Latent Task Adaptation with Large-Scale Hierarchies
Yangqing Jia, Trevor Darrell; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2080-2087
Recent years have witnessed the success of large-scale image classification systems that are able to identify objects among thousands of possible labels. However, it is yet unclear how general classifiers such as ones trained on ImageNet can be optimally adapted to specific tasks, each of which only covers a semantically related subset of all the objects in the world. It is inefficient and suboptimal to retrain classifiers whenever a new task is given, and is inapplicable when tasks are not given explicitly, but implicitly specified as a set of image queries. In this paper we propose a novel probabilistic model that jointly identifies the underlying task and performs prediction with a lineartime probabilistic inference algorithm, given a set of query images from a latent task. We present efficient ways to estimate parameters for the model, and an open-source toolbox to train classifiers distributedly at a large scale. Empirical results based on the ImageNet data showed significant performance increase over several baseline algorithms.
****

Image Co-segmentation via Consistent Functional Maps
Fan Wang, Qixing Huang, Leonidas J. Guibas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 849-856
Joint segmentation of image sets has great importance for object recognition, image classification, and image retrieval. In this paper, we aim to jointly segment a set of images starting from a small number of labeled images or none at all. To allow the images to share segmentation information with each other, we build a network that contains segmented as well as unsegmented images, and extract functional maps between connected image pairs based on image appearance features. These functional maps act as general property transporters between the images and, in particular, are used to transfer segmentations. We define and operate in a reduced functional space optimized so that the functional maps approximately satisfy cycle-consistency under composition in the network. A joint optimization framework is proposed to simultaneously generate all segmentation functions over the images so that they both align with local segmentation cues in each particular image, and agree with each other under network transportation. This formulation allows us to extract segmentations even with no training data, but can also exploit such data when available. The collective effect of the joint processing using functional maps leads to accurate information sharing among images and yields superior segmentation results, as shown on the iCoseg, MSRC, and PASCAL datasets.
****

Manipulation Pattern Discovery: A Nonparametric Bayesian Approach
Bingbing Ni, Pierre Moulin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1361-1368
We aim to unsupervisedly discover human's action (motion) patterns of manipulating various objects in scenarios such as assisted living. We are motivated by two key observations. First, large variation exists in motion patterns associated with various types of objects being manipulated, thus manually defining motion primitives is infeasible. Second, some motion patterns are shared among different objects being manipulated while others are object specific. We therefore propose a nonparametric Bayesian method that adopts a hierarchical Dirichlet process prior to learn representative manipulation (motion) patterns in an unsupervised manner. Taking easy-to-obtain object detection score maps and dense motion trajectories as inputs, the proposed probabilistic model can discover motion pattern groups associated with different types of objects being manipulated with a shared manipulation pattern dictionary. The size of the learned dictionary is automatically inferred. Comprehensive experiments on two assisted living benchmarks and a cooking motion dataset demonstrate superiority of our learned manipulation pattern dictionary in representing manipulation actions for recognition.
****

Large-Scale Image Annotation by Efficient and Robust Kernel Metric Learning

Zheyun Feng, Rong Jin, Anil Jain; Proceedings of the IEEE International Conferen ce on Computer Vision (ICCV), 2013, pp. 1609-1616
One of the key challenges in search-based image annotation models is to define a n appropriate similarity measure between images. Many kernel distance metric lea rning (KML) algorithms have been developed in order to capture the nonlinear rel ationships between visual features and semantics of the images. One fundamental limitation in applying KML to image annotation is that it requires converting im age annotations into binary constraints, leading to a significant information lo ss. In addition, most KML algorithms suffer from high computational cost due to the requirement that the learned matrix has to be positive semi-definitive (PSD) . In this paper, we propose a robust kernel metric learning (RKML) algorithm bas ed on the regression technique that is able to directly utilize image annotation s. The proposed method is also computationally more efficient because PSD proper ty is automatically ensured by regression. We provide the theoretical guarantee for the proposed algorithm, and verify its efficiency and effectiveness for imag e annotation by comparing it to state-of-the-art approaches for both distance me tric learning and image annotation.
****

Hybrid Deep Learning for Face Verification
Yi Sun, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE International Confer ence on Computer Vision (ICCV), 2013, pp. 1489-1496
This paper proposes a hybrid convolutional network (ConvNet)-Restricted Boltzman n Machine (RBM) model for face verification in wild conditions. A key contributi on of this work is to directly learn relational visual features, which indicate identity similarities, from raw pixels of face pairs with a hybrid deep network. The deep ConvNets in our model mimic the primary visual cortex to jointly extra ct local relational visual features from two face images compared with the learn ed filter pairs. These relational features are further processed through multipl e layers to extract high-level and global features. Multiple groups of ConvNets are constructed in order to achieve robustness and characterize face similaritie s from different aspects. The top-layer RBM performs inference from complementar y high-level features extracted from different ConvNet groups with a two-level a verage pooling hierarchy. The entire hybrid deep network is jointly fine-tuned t o optimize for the task of face verification. Our model achieves competitive fac e verification performance on the LFW dataset.
****

Latent Data Association: Bayesian Model Selection for Multi-target Tracking
Aleksandr V. Segal, Ian Reid; Proceedings of the IEEE International Conference o n Computer Vision (ICCV), 2013, pp. 2904-2911
We propose a novel parametrization of the data association problem for multi-tar get tracking. In our formulation, the number of targets is implicitly inferred t ogether with the data association, effectively solving data association and mode l selection as a single inference problem. The novel formulation allows us to in terpret data association and tracking as a single Switching Linear Dynamical Sys tem (SLDS). We compute an approximate posterior solution to this problem using a dynamic programming/message passing technique. This inference-based approach al lows us to incorporate richer probabilistic models into the tracking system. In particular, we incorporate inference over inliers/outliers and track termination times into the system. We evaluate our approach on publicly available datasets and demonstrate results competitive with, and in some cases exceeding the state of the art.
****

Recursive Estimation of the Stein Center of SPD Matrices and Its Applications
Hesamoddin Salehian, Guang Cheng, Baba C. Vemuri, Jeffrey Ho; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1793-1800
Symmetric positive-definite (SPD) matrices are ubiquitous in Computer Vision, Ma chine Learning and Medical Image Analysis. Finding the center/average of a popul ation of such matrices is a common theme in many algorithms such as clustering, segmentation, principal geodesic analysis, etc. The center of a population of su ch matrices can be defined using a variety of distance/divergence measures as th

e minimizer of the sum of squared distances/divergences from the unknown center to the members of the population. It is well known that the computation of the Karcher mean for the space of SPD matrices which is a negativelycurved Riemannian manifold is computationally expensive. Recently, the LogDet divergence-based center was shown to be a computationally attractive alternative. However,the LogDet-based mean of more than two matrices can not be computed in closed form, which makes it computationally less attractive for large populations. In this paper we present a novel recursive estimator for center based on the Stein distance â■■ which is the square root of the LogDet divergence â■■ that is significantly faster than the batch mode computation of this center. The key theoretical contribution is a closed-form solution for the weighted Stein center of two SPD matrices, which is used in the recursive computation of the Stein center for a population of SPD matrices. Additionally, we show experimental evidence of the convergence of our recursive Stein center estimator to the batch mode Stein center. We present applications of our recursive estimator to K-means clustering and image indexing depicting significant time gains over corresponding algorithms that use the batch mode computations. For the latter application, we develop novel hashing functions using the Stein distance and apply it to publicly available data sets, and experimental results have shown favorable comparisons to other competing methods.

****

Real-Time Solution to the Absolute Pose Problem with Unknown Radial Distortion and Focal Length

Zuzana Kukelova, Martin Bujnak, Tomas Pajdla; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2816-2823

The problem of determining the absolute position and orientation of a camera from a set of 2D-to-3D point correspondences is one of the most important problems in computer vision with a broad range of applications. In this paper we present a new solution to the absolute pose problem for camera with unknown radial distortion and unknown focal length from five 2D-to-3D point correspondences. Our new solver is numerically more stable, more accurate, and significantly faster than the existing state-of-the-art minimal four point absolute pose solvers for this problem. Moreover, our solver results in less solutions and can handle larger radial distortions. The new solver is straightforward and uses only simple concepts from linear algebra. Therefore it is simpler than the state-of-the-art Gr??bner basis solvers. We compare our new solver with the existing state-of-theart solvers and show its usefulness on synthetic and real datasets. 1

****

Sieving Regression Forest Votes for Facial Feature Detection in the Wild

Heng Yang, Ioannis Patras; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1936-1943

In this paper we propose a method for the localization of multiple facial features on challenging face images. In the regression forests (RF) framework, observations (patches) that are extracted at several image locations cast votes for the localization of several facial features. In order to filter out votes that are not relevant, we pass them through two types of sieves, that are organised in a cascade, and which enforce geometric constraints. The first sieve filters out votes that are not consistent with a hypothesis for the location of the face center. Several sieves of the second type, one associated with each individual facial point, filter out distant votes. We propose a method that adjusts onthe-fly the proximity threshold of each second type sieve by applying a classifier which, based on middle-level features extracted from voting maps for the facial feature in question, makes a sequence of decisions on whether the threshold should be reduced or not. We validate our proposed method on two challenging datasets with images collected from the Internet in which we obtain state of the art results without resorting to explicit facial shape models. We also show the benefits of our method for proximity threshold adjustment especially on 'difficult' face images.

****

Constant Time Weighted Median Filtering for Stereo Matching and Beyond

Ziyang Ma, Kaiming He, Yichen Wei, Jian Sun, Enhua Wu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 49-56
Despite the continuous advances in local stereo matching for years, most efforts are on developing robust cost computation and aggregation methods. Little attention has been seriously paid to the disparity refinement. In this work, we study weighted median filtering for disparity refinement. We discover that with this refinement, even the simple box filter aggregation achieves comparable accuracy with various sophisticated aggregation methods (with the same refinement). This is due to the nice weighted median filtering properties of removing outlier error while respecting edges/structures. This reveals that the previously overlooked refinement can be at least as crucial as aggregation. We also develop the first constant time algorithm for the previously time-consuming weighted median filter. This makes the simple combination "box aggregation + weighted median" an attractive solution in practice for both speed and accuracy. As a byproduct, the fast weighted median filtering unleashes its potential in other applications that were hampered by high complexities. We show its superiority in various applications such as depth upsampling, clip-art JPEG artifact removal, and image stylization.
****

Feature Weighting via Optimal Thresholding for Video Analysis
Zhongwen Xu, Yi Yang, Ivor Tsang, Nicu Sebe, Alexander G. Hauptmann; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3440-3447
Fusion of multiple features can boost the performance of large-scale visual classification and detection tasks like TRECVID Multimedia Event Detection (MED) competition [1]. In this paper, we propose a novel feature fusion approach, namely Feature Weighting via Optimal Thresholding (FWOT) to effectively fuse various features. FWOT learns the weights, thresholding and smoothing parameters in a joint framework to combine the decision values obtained from all the individual features and the early fusion. To the best of our knowledge, this is the first work to consider the weight and threshold factors of fusion problem simultaneously. Compared to state-of-the-art fusion algorithms, our approach achieves promising improvements on HMDB [8] action recognition dataset and CCV [5] video classification dataset. In addition, experiments on two TRECVID MED 2011 collections show that our approach outperforms the state-of-the-art fusion methods for complex event detection.
****

Restoring an Image Taken through a Window Covered with Dirt or Rain
David Eigen, Dilip Krishnan, Rob Fergus; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 633-640
Photographs taken through a window are often compromised by dirt or rain present on the window surface. Common cases of this include pictures taken from inside a vehicle, or outdoor security cameras mounted inside a protective enclosure. At capture time, defocus can be used to remove the artifacts, but this relies on achieving a shallow depth-of-field and placement of the camera close to the window. Instead, we present a post-capture image processing solution that can remove localized rain and dirt artifacts from a single image. We collect a dataset of clean/corrupted image pairs which are then used to train a specialized form of convolutional neural network. This learns how to map corrupted image patches to clean ones, implicitly capturing the characteristic appearance of dirt and water droplets in natural images. Our models demonstrate effective removal of dirt and rain in outdoor test conditions.
****

Tracking via Robust Multi-task Multi-view Joint Sparse Representation
Zhibin Hong, Xue Mei, Danil Prokhorov, Dacheng Tao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 649-656
Combining multiple observation views has proven beneficial for tracking. In this paper, we cast tracking as a novel multi-task multi-view sparse learning problem and exploit the cues from multiple views including various types of visual features, such as intensity, color, and edge, where each feature observation can be

sparsely represented by a linear combination of atoms from an adaptive feature dictionary. The proposed method is integrated in a particle filter framework where every view in each particle is regarded as an individual task. We jointly consider the underlying relationship between tasks across different views and different particles, and tackle it in a unified robust multi-task formulation. In addition, to capture the frequently emerging outlier tasks, we decompose the representation matrix to two collaborative components which enable a more robust and accurate approximation. We show that the proposed formulation can be efficiently solved using the Accelerated Proximal Gradient method with a small number of closed-form updates. The presented tracker is implemented using four types of features and is tested on numerous benchmark video sequences. Both the qualitative and quantitative results demonstrate the superior performance of the proposed approach compared to several stateof-the-art trackers.

****

A Simple Model for Intrinsic Image Decomposition with Depth Cues

Qifeng Chen, Vladlen Koltun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 241-248

We present a model for intrinsic decomposition of RGB-D images. Our approach analyzes a single RGB-D image and estimates albedo and shading fields that explain the input. To disambiguate the problem, our model estimates a number of components that jointly account for the reconstructed shading. By decomposing the shading field, we can build in assumptions about image formation that help distinguish reflectance variation from shading. These assumptions are expressed as simple nonlocal regularizers. We evaluate the model on real-world images and on a challenging synthetic dataset. The experimental results demonstrate that the presented approach outperforms prior models for intrinsic decomposition of RGB-D images.

****

Holistic Scene Understanding for 3D Object Detection with RGBD Cameras

Dahua Lin, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1417-1424

In this paper, we tackle the problem of indoor scene understanding using RGBD data. Towards this goal, we propose a holistic approach that exploits 2D segmentation, 3D geometry, as well as contextual relations between scenes and objects. Specifically, we extend the CPMC [3] framework to 3D in order to generate candidate cuboids, and develop a conditional random field to integrate information from different sources to classify the cuboids. With this formulation, scene classification and 3D object recognition are coupled and can be jointly solved through probabilistic inference. We test the effectiveness of our approach on the challenging NYU v2 dataset. The experimental results demonstrate that through effective evidence integration and holistic reasoning, our approach achieves substantial improvement over the state-of-the-art.

****

Pose-Free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model

Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, Dimitris N. Metaxas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1944-1951

This paper addresses the problem of facial landmark localization and tracking from a single camera. We present a two-stage cascaded deformable shape model to effectively and efficiently localize facial landmarks with large head pose variations. For face detection, we propose a group sparse learning method to automatically select the most salient facial landmarks. By introducing 3D face shape model, we use procrustes analysis to achieve pose-free facial landmark initialization. For deformation, the first step uses mean-shift local search with constrained local model to rapidly approach the global optimum. The second step uses component-wise active contours to discriminatively refine the subtle shape variation. Our framework can simultaneously handle face detection, pose-free landmark localization and tracking in real time. Extensive experiments are conducted on both laboratory environmental face databases and face-in-the-wild databases. All results demonstrate that our approach has certain advantages over state-of-theart meth

ods in handling pose variations 1 .
****
Online Robust Non-negative Dictionary Learning for Visual Tracking
Naiyan Wang, Jingdong Wang, Dit-Yan Yeung; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 657-664
This paper studies the visual tracking problem in video sequences and presents a novel robust sparse tracker under the particle filter framework. In particular, we propose an online robust non-negative dictionary learning algorithm for updating the object templates so that each learned template can capture a distinctive aspect of the tracked object. Another appealing property of this approach is that it can automatically detect and reject the occlusion and cluttered background in a principled way. In addition, we propose a new particle representation formulation using the Huber loss function. The advantage is that it can yield robust estimation without using trivial templates adopted by previous sparse trackers, leading to faster computation. We also reveal the equivalence between this new formulation and the previous one which uses trivial templates. The proposed tracker is empirically compared with state-of-the-art trackers on some challenging video sequences. Both quantitative and qualitative comparisons show that our proposed tracker is superior and more stable.
****
A Max-Margin Perspective on Sparse Representation-Based Classification
Zhaowen Wang, Jianchao Yang, Nasser Nasrabadi, Thomas Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1217-1224
Sparse Representation-based Classification (SRC) is a powerful tool in distinguishing signal categories which lie on different subspaces. Despite its wide application to visual recognition tasks, current understanding of SRC is solely based on a reconstructive perspective, which neither offers any guarantee on its classification performance nor provides any insight on how to design a discriminative dictionary for SRC. In this paper, we present a novel perspective towards SRC and interpret it as a margin classifier. The decision boundary and margin of SRC are analyzed in local regions where the support of sparse code is stable. Based on the derived margin, we propose a hinge loss function as the gauge for the classification performance of SRC. A stochastic gradient descent algorithm is implemented to maximize the margin of SRC and obtain more discriminative dictionaries. Experiments validate the effectiveness of the proposed approach in predicting classification performance and improving dictionary quality over reconstructive ones. Classification results competitive with other state-ofthe-art sparse coding methods are reported on several data sets.
****
Semantic Transform: Weakly Supervised Semantic Inference for Relating Visual Attributes
Sukrit Shankar, Joan Lasenby, Roberto Cipolla; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 361-368
Relative (comparative) attributes are promising for thematic ranking of visual entities, which also aids in recognition tasks [19, 23]. However, attribute rank learning often requires a substantial amount of relational supervision, which is highly tedious, and apparently impractical for realworld applications. In this paper, we introduce the Semantic Transform, which under minimal supervision, adaptively finds a semantic feature space along with a class ordering that is related in the best possible way. Such a semantic space is found for every attribute category. To relate the classes under weak supervision, the class ordering needs to be refined according to a cost function in an iterative procedure. This problem is ideally NP-hard, and we thus propose a constrained search tree formulation for the same. Driven by the adaptive semantic feature space representation, our model achieves the best results to date for all of the tasks of relative, absolute and zero-shot classification on two popular datasets.
****
Correlation Adaptive Subspace Segmentation by Trace Lasso
Canyi Lu, Jiashi Feng, Zhouchen Lin, Shuicheng Yan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1345-1352

This paper studies the subspace segmentation problem. Given a set of data points drawn from a union of subspaces, the goal is to partition them into their underlying subspaces they were drawn from. The spectral clustering method is used as the framework. It requires to find an affinity matrix which is close to block diagonal, with nonzero entries corresponding to the data point pairs from the same subspace. In this work, we argue that both sparsity and the grouping effect are important for subspace segmentation. A sparse affinity matrix tends to be block diagonal, with less connections between data points from different subspaces. The grouping effect ensures that the highly corrected data which are usually from the same subspace can be grouped together. Sparse Subspace Clustering (SSC), by using dt-minimization, encourages sparsity for data selection, but it lacks of the grouping effect. On the contrary, Low-Rank Representation (LRR), by rank minimization, and Least Squares Regression (LSR), by nk-regularization, exhibit strong grouping effect, but they are short in subset selection. Thus the obtained affinity matrix is usually very sparse by SSC, yet very dense by LRR and LSR. In this work, we propose the Correlation Adaptive Subspace Segmentation (CASS) method by using trace Lasso. CASS is a data correlation dependent method which simultaneously performs automatic data selection and groups correlated data together. It can be regarded as a method which adaptively balances SSC and LSR. Both theoretical and experimental results show the effectiveness of CASS.

****

DCSH - Matching Patches in RGBD Images

Yaron Eshet, Simon Korman, Eyal Ofek, Shai Avidan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 89-96

We extend patch based methods to work on patches in 3D space. We start with Coherency Sensitive Hashing [12] (CSH), which is an algorithm for matching patches between two RGB images, and extend it to work with RGBD images. This is done by warping all 3D patches to a common virtual plane in which CSH is performed. To avoid noise due to warping of patches of various normals and depths, we estimate a group of dominant planes and compute CSH on each plane separately, before merging the matching patches. The result is DCSH an algorithm that matches world (3D) patches in order to guide the search for image plane matches. An independent contribution is an extension of CSH, which we term Social-CSH. It allows a major speedup of the k nearest neighbor (kNN) version of CSH its runtime growing linearly, rather than quadratically, in k. Social-CSH is used as a subcomponent of DCSH when many NNs are required, as in the case of image denoising. We show the benefits of using depth information to image reconstruction and image denoising, demonstrated on several RGBD images.

****

Simultaneous Clustering and Tracklet Linking for Multi-face Tracking in Videos

Baoyuan Wu, Siwei Lyu, Bao-Gang Hu, Qiang Ji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2856-2863

We describe a novel method that simultaneously clusters and associates short sequences of detected faces (termed as face tracklets) in videos. The rationale of our method is that face tracklet clustering and linking are related problems that can benefit from the solutions of each other. Our method is based on a hidden Markov random field model that represents the joint dependencies of cluster labels and tracklet linking associations . We provide an efficient algorithm based on constrained clustering and optimal matching for the simultaneous inference of cluster labels and tracklet associations. We demonstrate significant improvements on the state-of-the-art results in face tracking and clustering performances on several video datasets.

****

Subpixel Scanning Invariant to Indirect Lighting Using Quadratic Code Length

Nicolas Martin, Vincent Couture, Sebastien Roy; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1441-1448

We present a scanning method that recovers dense subpixel camera-projector correspondence without requiring any photometric calibration nor preliminary knowledge of their relative geometry. Subpixel accuracy is achieved by considering several zero-crossings defined by the difference between pairs of unstructured patter

ns. We use gray-level band-pass white noise patterns that increase robustness to indirect lighting and scene discontinuities. Simulated and experimental results show that our method recovers scene geometry with high subpixel precision, and that it can handle many challenges of active reconstruction systems. We compare our results to state of the art methods such as micro phase shifting and modulated phase shifting.
****

PM-Huber: PatchMatch with Huber Regularization for Stereo Matching
Philipp Heise, Sebastian Klose, Brian Jensen, Alois Knoll; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2360-2367
Most stereo correspondence algorithms match support windows at integer-valued disparities and assume a constant disparity value within the support window. The recently proposed PatchMatch stereo algorithm [7] overcomes this limitation of previous algorithms by directly estimating planes. This work presents a method that integrates the PatchMatch stereo algorithm into a variational smoothing formulation using quadratic relaxation. The resulting algorithm allows the explicit regularization of the disparity and normal gradients using the estimated plane parameters. Evaluation of our method in the Middlebury benchmark shows that our method outperforms the traditional integer-valued disparity strategy as well as the original algorithm and its variants in sub-pixel accurate disparity estimation.
****

Relative Attributes for Large-Scale Abandoned Object Detection
Quanfu Fan, Prasad Gabbur, Sharath Pankanti; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2736-2743
Effective reduction of false alarms in large-scale video surveillance is rather challenging, especially for applications where abnormal events of interest rarely occur, such as abandoned object detection. We develop an approach to prioritize alerts by ranking them, and demonstrate its great effectiveness in reducing false positives while keeping good detection accuracy. Our approach benefits from a novel representation of abandoned object alerts by relative attributes, namely staticness, foregroundness and abandonment. The relative strengths of these attributes are quantified using a ranking function[19] learnt on suitably designed low-level spatial and temporal features.These attributes of varying strengths are not only powerful in distinguishing abandoned objects from false alarms such as people and light artifacts, but also computationally efficient for large-scale deployment. With these features, we apply a linear ranking algorithm to sort alerts according to their relevance to the end-user. We test the effectiveness of our approach on both public data sets and large ones collected from the real world.
****

Random Grids: Fast Approximate Nearest Neighbors and Range Searching for Image Search
Dror Aiger, Efi Kokiopoulou, Ehud Rivlin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3471-3478
We propose two solutions for both nearest neighbors and range search problems. For the nearest neighbors problem, we propose a c-approximate solution for the restricted version of the decision problem with bounded radius which is then reduced to the nearest neighbors by a known reduction. For range searching we propose a scheme that learns the parameters in a learning stage adopting them to the case of a set of points with low intrinsic dimension that are embedded in high dimensional space (common scenario for image point descriptors). We compare our algorithms to the best known methods for these problems, i.e. LSH, ANN and FLANN. We show analytically and experimentally that we can do better for moderate approximation factor. Our algorithms are trivial to parallelize. In the experiments conducted, running on couple of million images, our algorithms show meaningful speed-ups when compared with the above mentioned methods.
****

Image Guided Depth Upsampling Using Anisotropic Total Generalized Variation
David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Ruether, Horst Bischof; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 20

13, pp. 993-1000
In this work we present a novel method for the challenging problem of depth image upsampling. Modern depth cameras such as Kinect or Time of Flight cameras deliver dense, high quality depth measurements but are limited in their lateral resolution. To overcome this limitation we formulate a convex optimization problem using higher order regularization for depth image upsampling. In this optimization an anisotropic diffusion tensor, calculated from a high resolution intensity image, is used to guide the upsampling. We derive a numerical algorithm based on a primaldual formulation that is efficiently parallelized and runs at multiple frames per second. We show that this novel upsampling clearly outperforms state of the art approaches in terms of speed and accuracy on the widely used Middlebury 2007 datasets. Furthermore, we introduce novel datasets with highly accurate groundtruth, which, for the first time, enable to benchmark depth upsampling methods using real sensor data.
****

3D Scene Understanding by Voxel-CRF
Byung-Soo Kim, Pushmeet Kohli, Silvio Savarese; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1425-1432
Scene understanding is an important yet very challenging problem in computer vision. In the past few years, researchers have taken advantage of the recent diffusion of depth-RGB (RGB-D) cameras to help simplify the problem of inferring scene semantics. However, while the added 3D geometry is certainly useful to segment out objects with different depth values, it also adds complications in that the 3D geometry is often incorrect because of noisy depth measurements and the actual 3D extent of the objects is usually unknown because of occlusions. In this paper we propose a new method that allows us to jointly refine the 3D reconstruction of the scene (raw depth values) while accurately segmenting out the objects or scene elements from the 3D reconstruction. This is achieved by introducing a new model which we called Voxel-CRF. The Voxel-CRF model is based on the idea of constructing a conditional random field over a 3D volume of interest which captures the semantic and 3D geometric relationships among different elements (voxels) of the scene. Such model allows to jointly estimate (1) a dense voxel-based 3D reconstruction and (2) the semantic labels associated with each voxel even in presence of partial occlusions using an approximate yet efficient inference strategy. We evaluated our method on the challenging NYU Depth dataset (Version 1 and 2). Experimental results show that our method achieves competitive accuracy in inferring scene semantics and visually appealing results in improving the quality of the 3D reconstruction. We also demonstrate an interesting application of object removal and scene completion from RGB-D images.
****

No Matter Where You Are: Flexible Graph-Guided Multi-task Learning for Multi-view Head Pose Classification under Target Motion
Yan Yan, Elisa Ricci, Ramanathan Subramanian, Oswald Lanz, Nicu Sebe; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1177-1184
We propose a novel Multi-Task Learning framework (FEGA-MTL) for classifying the head pose of a person who moves freely in an environment monitored by multiple, large field-of-view surveillance cameras. As the target (person) moves, distortions in facial appearance owing to camera perspective and scale severely impede performance of traditional head pose classification methods. FEGA-MTL operates on a dense uniform spatial grid and learns appearance relationships across partitions as well as partition-specific appearance variations for a given head pose to build region-specific classifiers. Guided by two graphs which a-priori model appearance similarity among (i) grid partitions based on camera geometry and (ii) head pose classes, the learner efficiently clusters appearancewise related grid partitions to derive the optimal partitioning. For pose classification, upon determining the target's position using a person tracker, the appropriate regionspecific classifier is invoked. Experiments confirm that FEGA-MTL achieves state-of-the-art classification with few training data.
****

Dynamic Probabilistic Volumetric Models
Ali Osman Ulusoy, Octavian Biris, Joseph L. Mundy; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 505-512
This paper presents a probabilistic volumetric framework for image based modeling of general dynamic 3-d scenes. The framework is targeted towards high quality modeling of complex scenes evolving over thousands of frames. Extensive storage and computational resources are required in processing large scale space-time (4-d) data. Existing methods typically store separate 3-d models at each time step and do not address such limitations. A novel 4-d representation is proposed that adaptively subdivides in space and time to explain the appearance of 3-d dynamic surfaces. This representation is shown to achieve compression of 4-d data and provide efficient spatio-temporal processing. The advances of the proposed framework is demonstrated on standard datasets using free-viewpoint video and 3-d tracking applications.
****

Predicting an Object Location Using a Global Image Representation
Jose A. Rodriguez Serrano, Diane Larlus; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1729-1736
We tackle the detection of prominent objects in images as a retrieval task: given a global image descriptor, we find the most similar images in an annotated dataset, and transfer the object bounding boxes. We refer to this approach as data driven detection (DDD), that is an alternative to sliding windows. Previous works have used similar notions but with task-independent similarities and representations, i.e. they were not tailored to the end-goal of localization. This article proposes two contributions: (i) a metric learning algorithm and (ii) a representation of images as object probability maps, that are both optimized for detection. We show experimentally that these two contributions are crucial to DDD, do not require costly additional operations, and in some cases yield comparable or better results than state-of-the-art detectors despite conceptual simplicity and increased speed. As an application of prominent object detection, we improve fine-grained categorization by precropping images with the proposed approach.
****

Anchored Neighborhood Regression for Fast Example-Based Super-Resolution
Radu Timofte, Vincent De Smet, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1920-1927
Recently there have been significant advances in image upscaling or image super-resolution based on a dictionary of low and high resolution exemplars. The running time of the methods is often ignored despite the fact that it is a critical factor for real applications. This paper proposes fast super-resolution methods while making no compromise on quality. First, we support the use of sparse learned dictionaries in combination with neighbor embedding methods. In this case, the nearest neighbors are computed using the correlation with the dictionary atoms rather than the Euclidean distance. Moreover, we show that most of the current approaches reach top performance for the right parameters. Second, we show that using global collaborative coding has considerable speed advantages, reducing the super-resolution mapping to a precomputed projective matrix. Third, we propose the anchored neighborhood regression. That is to anchor the neighborhood embedding of a low resolution patch to the nearest atom in the dictionary and to precompute the corresponding embedding matrix. These proposals are contrasted with current state-ofthe-art methods on standard images. We obtain similar or improved quality and one or two orders of magnitude speed improvements.
****

Robust Object Tracking with Online Multi-lifespan Dictionary Learning
Junliang Xing, Jin Gao, Bing Li, Weiming Hu, Shuicheng Yan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 665-672
Recently, sparse representation has been introduced for robust object tracking. By representing the object sparsely, i.e., using only a few templates via 1 -norm minimization, these so-called 1 -trackers exhibit promising tracking results. In this work, we address the object template building and updating problem in these 1 -tracking approaches, which has not been fully studied. We propose to perf

orm template updating, in a new perspective, as an online incremental dictionary learning problem, which is efficiently solved through an online optimization procedure. To guarantee the robustness and adaptability of the tracking algorithm, we also propose to build a multi-lifespan dictionary model. By building target dictionaries of different lifespans, effective object observations can be obtained to deal with the well-known drifting problem in tracking and thus improve the tracking accuracy. We derive effective observation models both generatively and discriminatively based on the online multi-lifespan dictionary learning model and deploy them to the Bayesian sequential estimation framework to perform tracking. The proposed approach has been extensively evaluated on ten challenging video sequences. Experimental results demonstrate the effectiveness of the online learned templates, as well as the state-of-the-art tracking performance of the proposed approach.

****

Finding the Best from the Second Bests - Inhibiting Subjective Bias in Evaluation of Visual Tracking Algorithms

Yu Pang, Haibin Ling; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2784-2791

Evaluating visual tracking algorithms, or "trackers" for short, is of great importance in computer vision. However, it is hard to "fairly" compare trackers due to many parameters need to be tuned in the experimental configurations. On the other hand, when introducing a new tracker, a recent trend is to validate it by comparing it with several existing ones. Such an evaluation may have subjective biases towards the new tracker which typically performs the best. This is mainly due to the difficulty to optimally tune all its competitors and sometimes the selected testing sequences. By contrast, little subjective bias exists towards the "second best" ones 1 in the contest. This observation inspires us with a novel perspective towards inhibiting subjective bias in evaluating trackers by analyzing the results between the second bests. In particular, we first collect all tracking papers published in major computer vision venues in recent years. From these papers, after filtering out potential biases in various aspects, we create a dataset containing many records of comparison results between various visual trackers. Using these records, we derive performance rankings of the involved trackers by four different methods. The first two methods model the dataset as a graph and then derive the rankings over the graph, one by a rank aggregation algorithm and the other by a PageRank-like solution. The other two methods take the records as generated from sports contests and adopt widely used Elo's and Glicko's rating systems to derive the rankings. The experimental results are presented and may serve as a reference for related research.

****

Write a Classifier: Zero-Shot Learning Using Purely Textual Descriptions

Mohamed Elhoseiny, Babak Saleh, Ahmed Elgammal; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2584-2591

The main question we address in this paper is how to use purely textual description of categories with no training images to learn visual classifiers for these categories. We propose an approach for zero-shot learning of object categories where the description of unseen categories comes in the form of typical text such as an encyclopedia entry, without the need to explicitly defined attributes. We propose and investigate two baseline formulations, based on regression and domain adaptation. Then, we propose a new constrained optimization formulation that combines a regression function and a knowledge transfer function with additional constraints to predict the classifier parameters for new classes. We applied the proposed approach on two fine-grained categorization datasets, and the results indicate successful classifier prediction.

****

Detecting Dynamic Objects with Multi-view Background Subtraction

Raul Diaz, Sam Hallman, Charless C. Fowlkes; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 273-280

The confluence of robust algorithms for structure from motion along with high-coverage mapping and imaging of the world around us suggests that it will soon be

feasible to accurately estimate camera pose for a large class photographs taken in outdoor, urban environments. In this paper, we investigate how such information can be used to improve the detection of dynamic objects such as pedestrians and cars. First, we show that when rough camera location is known, we can utilize detectors that have been trained with a scene-specific background model in order to improve detection accuracy. Second, when precise camera pose is available, dense matching to a database of existing images using multi-view stereo provides a way to eliminate static backgrounds such as building facades, akin to background-subtraction often used in video analysis. We evaluate these ideas using a dataset of tourist photos with estimated camera pose. For template-based pedestrian detection, we achieve a 50 percent boost in average precision over baseline.
****

Face Recognition via Archetype Hull Ranking
Yuanjun Xiong, Wei Liu, Deli Zhao, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 585-592
The archetype hull model is playing an important role in large-scale data analytics and mining, but rarely applied to vision problems. In this paper, we migrate such a geometric model to address face recognition and verification together through proposing a unified archetype hull ranking framework. Upon a scalable graph characterized by a compact set of archetype exemplars whose convex hull encompasses most of the training images, the proposed framework explicitly captures the relevance between any query and the stored archetypes, yielding a rank vector over the archetype hull. The archetype hull ranking is then executed on every block of face images to generate a blockwise similarity measure that is achieved by comparing two different rank vectors with respect to the same archetype hull. After integrating blockwise similarity measurements with learned importance weights, we accomplish a sensible face similarity measure which can support robust and effective face recognition and verification. We evaluate the face similarity measure in terms of experiments performed on three benchmark face databases Multi-PIE, Pubfig83, and LFW, demonstrating its performance superior to the state-of-the-arts.
****

Compositional Models for Video Event Detection: A Multiple Kernel Learning Latent Variable Approach
Arash Vahdat, Kevin Cannons, Greg Mori, Sangmin Oh, Ilseo Kim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1185-1192
We present a compositional model for video event detection. A video is modeled using a collection of both global and segment-level features and kernel functions are employed for similarity comparisons. The locations of salient, discriminative video segments are treated as a latent variable, allowing the model to explicitly ignore portions of the video that are unimportant for classification. A novel, multiple kernel learning (MKL) latent support vector machine (SVM) is defined, that is used to combine and re-weight multiple feature types in a principled fashion while simultaneously operating within the latent variable framework. The compositional nature of the proposed model allows it to respond directly to the challenges of temporal clutter and intra-class variation, which are prevalent in unconstrained internet videos. Experimental results on the TRECVID Multimedia Event Detection 2011 (MED11) dataset demonstrate the efficacy of the method.
****

Nested Shape Descriptors
Jeffrey Byrne, Jianbo Shi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1201-1208
In this paper, we propose a new family of binary local feature descriptors called nested shape descriptors. These descriptors are constructed by pooling oriented gradients over a large geometric structure called the Hawaiian earring, which is constructed with a nested correlation structure that enables a new robust local distance function called the nesting distance. This distance function is unique to the nested descriptor and provides robustness to outliers from order statistics. In this paper, we define the nested shape descriptor family and introduce a specific member called the seed-of-life descriptor. We perform a trade study

to determine optimal descriptor parameters for the task of image matching. Finally, we evaluate performance compared to state-of-the-art local feature descriptors on the VGGAffine image matching benchmark, showing significant performance gains. Our descriptor is the first binary descriptor to outperform SIFT on this benchmark.
****

Coarse-to-Fine Semantic Video Segmentation Using Supervoxel Trees
Aastha Jain, Shuanak Chatterjee, Rene Vidal; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1865-1872
We propose an exact, general and efficient coarse-to-fine energy minimization strategy for semantic video segmentation. Our strategy is based on a hierarchical abstraction of the supervoxel graph that allows us to minimize an energy defined at the finest level of the hierarchy by minimizing a series of simpler energies defined over coarser graphs. The strategy is exact, i.e., it produces the same solution as minimizing over the finest graph. It is general, i.e., it can be used to minimize any energy function (e.g., unary, pairwise, and higher-order terms) with any existing energy minimization algorithm (e.g., graph cuts and belief propagation). It also gives significant speedups in inference for several datasets with varying degrees of spatio-temporal continuity. We also discuss the strengths and weaknesses of our strategy relative to existing hierarchical approaches, and the kinds of image and video data that provide the best speedups.
****

Local Signal Equalization for Correspondence Matching
Derek Bradley, Thabo Beeler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1881-1887
Correspondence matching is one of the most common problems in computer vision, and it is often solved using photo-consistency of local regions. These approaches typically assume that the frequency content in the local region is consistent in the image pair, such that matching is performed on similar signals. However, in many practical situations this is not the case, for example with low depth of field cameras a scene point may be out of focus in one view and in-focus in the other, causing a mismatch of frequency signals. Furthermore, this mismatch can vary spatially over the entire image. In this paper we propose a local signal equalization approach for correspondence matching. Using a measure of local image frequency, we equalize local signals using an efficient scale-space image representation such that their frequency contents are optimally suited for matching. Our approach allows better correspondence matching, which we demonstrate with a number of stereo reconstruction examples on synthetic and real datasets.
****

On One-Shot Similarity Kernels: Explicit Feature Maps and Properties
Stefanos Zafeiriou, Irene Kotsia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2392-2399
Kernels have been a common tool of machine learning and computer vision applications for modeling nonlinearities and/or the design of robust 1 similarity measures between objects. Arguably, the class of positive semidefinite (psd) kernels, widely known as Mercerâ■■s Kernels,constitutes one of the most well-studied cases. For every psd kernel there exists an associated feature map to an arbitrary dimensional Hilbert space H, the so-called feature space. The main reason behind psd kernelsâ■■ popularity is the fact that classification/regression techniques (such as Support Vector Machines (SVMs)) and component analysis algorithms (such as Kernel Principal Component Analysis (KPCA)) can be devised in H, without an explicit definition of the feature map, only by using the kernel (the so-called kernel trick). Recently, due to the development of very efficient solutions for large scale linear SVMs and for incremental linear component analysis, the research towards finding feature map approximations for classes of kernels has attracted significant interest. In this paper, we attempt the derivation of explicit feature maps of a recently proposed class of kernels, the so-called one-shot similarity kernels. We show that for this class of kernels either there exists an explicit representation in feature space or the kernel can be expressed in such a form that allows for exact incremental learning. We theoretically explore the pr

operties of these kernels and show how these kernels can be used for the develop ment of robust visual tracking, recognition and deformable fitting algorithms.
****

Combining the Right Features for Complex Event Recognition
Kevin Tang, Bangpeng Yao, Li Fei-Fei, Daphne Koller; Proceedings of the IEEE Int ernational Conference on Computer Vision (ICCV), 2013, pp. 2696-2703
In this paper, we tackle the problem of combining features extracted from video for complex event recognition. Feature combination is an especially relevant tas k in video data, as there are many features we can extract, ranging from image f eatures computed from individual frames to video features that take temporal inf ormation into account. To combine features effectively, we propose a method that is able to be selective of different subsets of features, as some features or f eature combinations may be uninformative for certain classes. We introduce a hie rarchical method for combining features based on the AND/OR graph structure, whe re nodes in the graph represent combinations of different sets of features. Our method automatically learns the structure of the AND/OR graph using score-based structure learning, and we introduce an inference procedure that is able to effi ciently compute structure scores. We present promising results and analysis on t he difficult and large-scale 2011 TRECVID Multimedia Event Detection dataset [17].
****

NEIL: Extracting Visual Knowledge from Web Data
Xinlei Chen, Abhinav Shrivastava, Abhinav Gupta; Proceedings of the IEEE Interna tional Conference on Computer Vision (ICCV), 2013, pp. 1409-1416
We propose NEIL (Never Ending Image Learner), a computer program that runs 24 ho urs per day and 7 days per week to automatically extract visual knowledge from I nternet data. NEIL uses a semi-supervised learning algorithm that jointly discov ers common sense relationships (e.g., "Corolla is a kind of/looks similar to Car ","Wheel is a part of Car") and labels instances of the given visual categories. It is an attempt to develop the world's largest visual structured knowledge bas e with minimum human labeling effort. As of 10 th October 2013, NEIL has been co ntinuously running for 2.5 months on 200 core cluster (more than 350K CPU hours) and has an ontology of 1152 object categories, 1034 scene categories and 87 att ributes. During this period, NEIL has discovered more than 1700 relationships an d has labeled more than 400K visual instances.
****

Joint Subspace Stabilization for Stereoscopic Video
Feng Liu, Yuzhen Niu, Hailin Jin; Proceedings of the IEEE International Conferen ce on Computer Vision (ICCV), 2013, pp. 73-80
Shaky stereoscopic video is not only unpleasant to watch but may also cause 3D f atigue. Stabilizing the left and right view of a stereoscopic video separately u sing a monocular stabilization method tends to both introduce undesirable vertic al disparities and damage horizontal disparities, which may destroy the stereosc opic viewing experience. In this paper, we present a joint subspace stabilizatio n method for stereoscopic video. We prove that the low-rank subspace constraint for monocular video [10] also holds for stereoscopic video. Particularly, the fe ature trajectories from the left and right video share the same subspace. Based on this proof, we develop a stereo subspace stabilization method that jointly co mputes a common subspace from the left and right video and uses it to stabilize the two videos simultaneously. Our method meets the stereoscopic constraints wit hout 3D reconstruction or explicit left-right correspondence. We test our method on a variety of stereoscopic videos with different scene content and camera mot ion. The experiments show that our method achieves high-quality stabilization fo r stereoscopic video in a robust and efficient way.
****

Learning CRFs for Image Parsing with Adaptive Subgradient Descent
Honghui Zhang, Jingdong Wang, Ping Tan, Jinglu Wang, Long Quan; Proceedings of t he IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3080-3087
We propose an adaptive subgradient descent method to efficiently learn the param eters of CRF models for image parsing. To balance the learning efficiency and pe

rformance of the learned CRF models, the parameter learning is iteratively carri ed out by solving a convex optimization problem in each iteration, which integra tes a proximal term to preserve the previously learned information and the large margin preference to distinguish bad labeling and the ground truth labeling. A solution of subgradient descent updating form is derived for the convex optimiza tion problem, with an adaptively determined updating step-size. Besides, to deal with partially labeled training data, we propose a new objective constraint mod eling both the labeled and unlabeled parts in the partially labeled training dat a for the parameter learning of CRF models. The superior learning efficiency of the proposed method is verified by the experiment results on two public datasets . We also demonstrate the powerfulness of our method for handling partially labe led training data.
****

Box in the Box: Joint 3D Layout and Object Reasoning from Single Images
Alexander G. Schwing, Sanja Fidler, Marc Pollefeys, Raquel Urtasun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 353-36 0
In this paper we propose an approach to jointly infer the room layout as well as the objects present in the scene. Towards this goal, we propose a branch and bo und algorithm which is guaranteed to retrieve the global optimum of the joint pr oblem. The main difficulty resides in taking into account occlusion in order to not over-count the evidence. We introduce a new decomposition method, which gene ralizes integral geometry to triangular shapes, and allows us to bound the diffe rent terms in constant time. We exploit both geometric cues and object detectors as image features and show large improvements in 2D and 3D object detection ove r state-of-the-art deformable part-based models.
****

A Global Linear Method for Camera Pose Registration
Nianjuan Jiang, Zhaopeng Cui, Ping Tan; Proceedings of the IEEE International Co nference on Computer Vision (ICCV), 2013, pp. 481-488
We present a linear method for global camera pose registration from pairwise rel ative poses encoded in essential matrices. Our method minimizes an approximate g eometric error to enforce the triangular relationship in camera triplets. This f ormulation does not suffer from the typical 'unbalanced scale' problem in linear methods relying on pairwise translation direction constraints, i.e. an algebrai c error; nor the system degeneracy from collinear motion. In the case of three c ameras, our method provides a good linear approximation of the trifocal tensor. It can be directly scaled up to register multiple cameras. The results obtained are accurate for point triangulation and can serve as a good initialization for final bundle adjustment. We evaluate the algorithm performance with different ty pes of data and demonstrate its effectiveness. Our system produces good accuracy , robustness, and outperforms some well-known systems on efficiency.
****

Heterogeneous Image Features Integration via Multi-modal Semi-supervised Learnin g Model
Xiao Cai, Feiping Nie, Weidong Cai, Heng Huang; Proceedings of the IEEE Internat ional Conference on Computer Vision (ICCV), 2013, pp. 1737-1744
Automatic image categorization has become increasingly important with the develo pment of Internet and the growth in the size of image databases. Although the im age categorization can be formulated as a typical multiclass classification prob lem, two major challenges have been raised by the real-world images. On one hand , though using more labeled training data may improve the prediction performance , obtaining the image labels is a time consuming as well as biased process. On t he other hand, more and more visual descriptors have been proposed to describe o bjects and scenes appearing in images and different features describe different aspects of the visual characteristics. Therefore, how to integrate heterogeneous visual features to do the semi-supervised learning is crucial for categorizing large-scale image data. In this paper, we propose a novel approach to integrate heterogeneous features by performing multi-modal semi-supervised classification on unlabeled as well as unsegmented images. Considering each type of feature as

one modality, taking advantage of the large amount of unlabeled data information , our new adaptive multimodal semi-supervised classification (AMMSS) algorithm l earns a commonly shared class indicator matrix and the weights for different mod alities (image features) simultaneously.
****

3DNN: Viewpoint Invariant 3D Geometry Matching for Scene Understanding
Scott Satkin, Martial Hebert; Proceedings of the IEEE International Conference o n Computer Vision (ICCV), 2013, pp. 1873-1880
We present a new algorithm 3DNN (3D NearestNeighbor), which is capable of matchi ng an image with 3D data, independently of the viewpoint from which the image wa s captured. By leveraging rich annotations associated with each image, our algor ithm can automatically produce precise and detailed 3D models of a scene from a single image. Moreover, we can transfer information across images to accurately label and segment objects in a scene. The true benefit of 3DNN compared to a tra ditional 2D nearest-neighbor approach is that by generalizing across viewpoints, we free ourselves from the need to have training examples captured from all pos sible viewpoints. Thus, we are able to achieve comparable results using orders o f magnitude less data, and recognize objects from never-beforeseen viewpoints. I n this work, we describe the 3DNN algorithm and rigorously evaluate its performa nce for the tasks of geometry estimation and object detection/segmentation. By d ecoupling the viewpoint and the geometry of an image, we develop a scene matchin g approach which is truly 100% viewpoint invariant, yielding state-of-the-art pe rformance on challenging data.
****

Correntropy Induced L2 Graph for Robust Subspace Clustering
Canyi Lu, Jinhui Tang, Min Lin, Liang Lin, Shuicheng Yan, Zhouchen Lin; Proceedi ngs of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 18 01-1808
In this paper, we study the robust subspace clustering problem, which aims to cl uster the given possibly noisy data points into their underlying subspaces. A la rge pool of previous subspace clustering methods focus on the graph construction by different regularization of the representation coefficient. We instead focus on the robustness of the model to non-Gaussian noises. We propose a new robust clustering method by using the correntropy induced metric, which is robust for h andling the non-Gaussian and impulsive noises. Also we further extend the method for handling the data with outlier rows/features. The multiplicative form of ha lf-quadratic optimization is used to optimize the nonconvex correntropy objectiv e function of the proposed models. Extensive experiments on face datasets well d emonstrate that the proposed methods are more robust to corruptions and occlusio ns.
****

Unsupervised Domain Adaptation by Domain Invariant Projection
Mahsa Baktashmotlagh, Mehrtash T. Harandi, Brian C. Lovell, Mathieu Salzmann; Pr oceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 769-776
Domain-invariant representations are key to addressing the domain shift problem where the training and test examples follow different distributions. Existing te chniques that have attempted to match the distributions of the source and target domains typically compare these distributions in the original feature space. Th is space, however, may not be directly suitable for such a comparison, since som e of the features may have been distorted by the domain shift, or may be domain specific. In this paper, we introduce a Domain Invariant Projection approach: An unsupervised domain adaptation method that overcomes this issue by extracting t he information that is invariant across the source and target domains. More spec ifically, we learn a projection of the data to a low-dimensional latent space wh ere the distance between the empirical distributions of the source and target ex amples is minimized. We demonstrate the effectiveness of our approach on the tas k of visual object recognition and show that it outperforms state-of-the-art met hods on a standard domain adaptation benchmark dataset.
****

Large-Scale Multi-resolution Surface Reconstruction from RGB-D Sequences
Frank Steinbrucker, Christian Kerl, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3264-3271
We propose a method to generate highly detailed, textured 3D models of large environments from RGB-D sequences. Our system runs in real-time on a standard desktop PC with a state-of-the-art graphics card. To reduce the memory consumption, we fuse the acquired depth maps and colors in a multi-scale octree representation of a signed distance function. To estimate the camera poses, we construct a pose graph and use dense image alignment to determine the relative pose between pairs of frames. We add edges between nodes when we detect loop-closures and optimize the pose graph to correct for long-term drift. Our implementation is highly parallelized on graphics hardware to achieve real-time performance. More specifically, we can reconstruct, store, and continuously update a colored 3D model of an entire corridor of nine rooms at high levels of detail in real-time on a single GPU with 2.5GB.
****
Detecting Curved Symmetric Parts Using a Deformable Disc Model
Tom Sie Ho Lee, Sanja Fidler, Sven Dickinson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1753-1760
Symmetry is a powerful shape regularity that's been exploited by perceptual grouping researchers in both human and computer vision to recover part structure from an image without a priori knowledge of scene content. Drawing on the concept of a medial axis, defined as the locus of centers of maximal inscribed discs that sweep out a symmetric part, we model part recovery as the search for a sequence of deformable maximal inscribed disc hypotheses generated from a multiscale superpixel segmentation, a framework proposed by [13]. However, we learn affinities between adjacent superpixels in a space that's invariant to bending and tapering along the symmetry axis, enabling us to capture a wider class of symmetric parts. Moreover, we introduce a global cost that perceptually integrates the hypothesis space by combining a pairwise and a higher-level smoothing term, which we minimize globally using dynamic programming. The new framework is demonstrated on two datasets, and is shown to significantly outperform the baseline [13].
****
Hierarchical Data-Driven Descent for Efficient Optimal Deformation Estimation
Yuandong Tian, Srinivasa G. Narasimhan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2288-2295
Real-world surfaces such as clothing, water and human body deform in complex ways. The image distortions observed are high-dimensional and non-linear, making it hard to estimate these deformations accurately. The recent datadriven descent approach [17] applies Nearest Neighbor estimators iteratively on a particular distribution of training samples to obtain a globally optimal and dense deformation field between a template and a distorted image. In this work, we develop a hierarchical structure for the Nearest Neighbor estimators, each of which can have only a local image support. We demonstrate in both theory and practice that this algorithm has several advantages over the nonhierarchical version: it guarantees global optimality with significantly fewer training samples, is several orders faster, provides a metric to decide whether a given image is "hard" (or "easy") requiring more (or less) samples, and can handle more complex scenes that include both global motion and local deformation. The proposed algorithm successfully tracks a broad range of non-rigid scenes including water, clothing, and medical images, and compares favorably against several other deformation estimation and tracking approaches that do not provide optimality guarantees.
****
Recognising Human-Object Interaction via Exemplar Based Modelling
Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, Shaogang Gong, Tao Xiang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3144-3151
Human action can be recognised from a single still image by modelling Human-object interaction (HOI), which infers the mutual spatial structure information between human and object as well as their appearance. Existing approaches rely heavi

ly on accurate detection of human and object, and estimation of human pose. They are thus sensitive to large variations of human poses, occlusion and unsatisfactory detection of small size objects. To overcome this limitation, a novel exemplar based approach is proposed in this work. Our approach learns a set of spatial pose-object interaction exemplars, which are density functions describing how a person is interacting with a manipulated object for different activities spatially in a probabilistic way. A representation based on our HOI exemplar thus has great potential for being robust to the errors in human/object detection and pose estimation. A new framework consists of a proposed exemplar based HOI descriptor and an activity specific matching model that learns the parameters is formulated for robust human activity recognition. Experiments on two benchmark activity datasets demonstrate that the proposed approach obtains state-ofthe-art performance.

****

How Do You Tell a Blackbird from a Crow?

Thomas Berg, Peter N. Belhumeur; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 9-16

How do you tell a blackbird from a crow? There has been great progress toward automatic methods for visual recognition, including fine-grained visual categorization in which the classes to be distinguished are very similar. In a task such as bird species recognition, automatic recognition systems can now exceed the performance of non-experts most people are challenged to name a couple dozen bird species, let alone identify them. This leads us to the question, "Can a recognition system show humans what to look for when identifying classes (in this case birds)?" In the context of fine-grained visual categorization, we show that we can automatically determine which classes are most visually similar, discover what visual features distinguish very similar classes, and illustrate the key features in a way meaningful to humans. Running these methods on a dataset of bird images, we can generate a visual field guide to birds which includes a tree of similarity that displays the similarity relations between all species, pages for each species showing the most similar other species, and pages for each pair of similar species illustrating their differences.

****

Video Synopsis by Heterogeneous Multi-source Correlation

Xiatian Zhu, Chen Change Loy, Shaogang Gong; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 81-88

Generating coherent synopsis for surveillance video stream remains a formidable challenge due to the ambiguity and uncertainty inherent to visual observations. In contrast to existing video synopsis approaches that rely on visual cues alone, we propose a novel multi-source synopsis framework capable of correlating visual data and independent non-visual auxiliary information to better describe and summarise subtle physical events in complex scenes. Specifically, our unsupervised framework is capable of seamlessly uncovering latent correlations among heterogeneous types of data sources, despite the non-trivial heteroscedasticity and dimensionality discrepancy problems. Additionally, the proposed model is robust to partial or missing non-visual information. We demonstrate the effectiveness of our framework on two crowded public surveillance datasets.

****

Semantic Segmentation without Annotating Segments

Wei Xia, Csaba Domokos, Jian Dong, Loong-Fah Cheong, Shuicheng Yan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2176-2183

Numerous existing object segmentation frameworks commonly utilize the object bounding box as a prior. In this paper, we address semantic segmentation assuming that object bounding boxes are provided by object detectors, but no training data with annotated segments are available. Based on a set of segment hypotheses, we introduce a simple voting scheme to estimate shape guidance for each bounding box. The derived shape guidance is used in the subsequent graph-cut-based figure-ground segmentation. The final segmentation result is obtained by merging the segmentation results in the bounding boxes. We conduct an extensive analysis of th

e effect of object bounding box accuracy. Comprehensive experiments on both the challenging PASCAL VOC object segmentation dataset and GrabCut50 image segmentation dataset show that the proposed approach achieves competitive results compared to previous detection or bounding box prior based methods, as well as other state-of-the-art semantic segmentation methods.
****

Action Recognition with Actons
Jun Zhu, Baoyuan Wang, Xiaokang Yang, Wenjun Zhang, Zhuowen Tu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3559-3566
With the improved accessibility to an exploding amount of video data and growing demands in a wide range of video analysis applications, video-based action recognition/classification becomes an increasingly important task in computer vision. In this paper, we propose a two-layer structure for action recognition to automatically exploit a mid-level "acton" representation. The weakly-supervised actons are learned via a new max-margin multi-channel multiple instance learning framework, which can capture multiple mid-level action concepts simultaneously. The learned actons (with no requirement for detailed manual annotations) observe the properties of being compact, informative, discriminative, and easy to scale. The experimental results demonstrate the effectiveness of applying the learned actons in our two-layer structure, and show the state-ofthe-art recognition performance on two challenging action datasets, i.e., Youtube and HMDB51.
****

Exemplar Cut
Jimei Yang, Yi-Hsuan Tsai, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 857-864
We present a hybrid parametric and nonparametric algorithm, exemplar cut, for generating class-specific object segmentation hypotheses. For the parametric part, we train a pylon model on a hierarchical region tree as the energy function for segmentation. For the nonparametric part, we match the input image with each exemplar by using regions to obtain a score which augments the energy function from the pylon model. Our method thus generates a set of highly plausible segmentation hypotheses by solving a series of exemplar augmented graph cuts. Experimental results on the Graz and PASCAL datasets show that the proposed algorithm achieves favorable segmentation performance against the state-of-the-art methods in terms of visual quality and accuracy.
****

Discovering Object Functionality
Bangpeng Yao, Jiayuan Ma, Li Fei-Fei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2512-2519
Object functionality refers to the quality of an object that allows humans to perform some specific actions. It has been shown in psychology that functionality (affordance) is at least as essential as appearance in object recognition by humans. In computer vision, most previous work on functionality either assumes exactly one functionality for each object, or requires detailed annotation of human poses and objects. In this paper, we propose a weakly supervised approach to discover all possible object functionalities. Each object functionality is represented by a specific type of human-object interaction. Our method takes any possible human-object interaction into consideration, and evaluates image similarity in 3D rather than 2D in order to cluster human-object interactions more coherently. Experimental results on a dataset of people interacting with musical instruments show the effectiveness of our approach.
****

Saliency Detection: A Boolean Map Approach
Jianming Zhang, Stan Sclaroff; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 153-160
A novel Boolean Map based Saliency (BMS) model is proposed. An image is characterized by a set of binary images, which are generated by randomly thresholding the image's color channels. Based on a Gestalt principle of figure-ground segregation, BMS computes saliency maps by analyzing the topological structure of Boolean maps. BMS is simple to implement and efficient to run. Despite its simplicity,

BMS consistently achieves state-of-the-art performance compared with ten leading methods on five eye tracking datasets. Furthermore, BMS is also shown to be advantageous in salient object detection.
****

Active MAP Inference in CRFs for Efficient Semantic Segmentation
Gemma Roig, Xavier Boix, Roderick De Nijs, Sebastian Ramos, Koljia Kuhnlenz, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2312-2319
Most MAP inference algorithms for CRFs optimize an energy function knowing all the potentials. In this paper, we focus on CRFs where the computational cost of instantiating the potentials is orders of magnitude higher than MAP inference. This is often the case in semantic image segmentation, where most potentials are instantiated by slow classifiers fed with costly features. We introduce Active MAP inference 1) to on-the-fly select a subset of potentials to be instantiated in the energy function, leaving the rest of the parameters of the potentials unknown, and 2) to estimate the MAP labeling from such incomplete energy function. Results for semantic segmentation benchmarks, namely PASCAL VOC 2010 [5] and MSRC-21 [19], show that Active MAP inference achieves similar levels of accuracy but with major efficiency gains.
****

PixelTrack: A Fast Adaptive Algorithm for Tracking Non-rigid Objects
Stefan Duffner, Christophe Garcia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2480-2487
In this paper, we present a novel algorithm for fast tracking of generic objects in videos. The algorithm uses two components: a detector that makes use of the generalised Hough transform with pixel-based descriptors, and a probabilistic segmentation method based on global models for foreground and background. These components are used for tracking in a combined way, and they adapt each other in a co-training manner. Through effective model adaptation and segmentation, the algorithm is able to track objects that undergo rigid and non-rigid deformations and considerable shape and appearance variations. The proposed tracking method has been thoroughly evaluated on challenging standard videos, and outperforms state-of-theart tracking methods designed for the same task. Finally, the proposed models allow for an extremely efficient implementation, and thus tracking is very fast.
****

Class-Specific Simplex-Latent Dirichlet Allocation for Image Classification
Mandar Dixit, Nikhil Rasiwasia, Nuno Vasconcelos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2672-2679
An extension of the latent Dirichlet allocation (LDA), denoted class-specific-simplex LDA (css-LDA), is proposed for image classification. An analysis of the supervised LDA models currently used for this task shows that the impact of class information on the topics discovered by these models is very weak in general. This implies that the discovered topics are driven by general image regularities, rather than the semantic regularities of interest for classification. To address this, we introduce a model that induces supervision in topic discovery, while retaining the original flexibility of LDA to account for unanticipated structures of interest. The proposed css-LDA is an LDA model with class supervision at the level of image features. In css-LDA topics are discovered per class, i.e. a single set of topics shared across classes is replaced by multiple class-specific topic sets. This model can be used for generative classification using the Bayes decision rule or even extended to discriminative classification with support vector machines (SVMs). A css-LDA model can endow an image with a vector of class and topic specific count statistics that are similar to the Bag-of-words (BoW) histogram. SVM-based discriminants can be learned for classes in the space of these histograms. The effectiveness of css-LDA model in both generative and discriminative classification frameworks is demonstrated through an extensive experimental evaluation, involving multiple benchmark datasets, where it is shown to outperform all existing LDA based image classification approaches.
****

BOLD Features to Detect Texture-less Objects
Federico Tombari, Alessandro Franchi, Luigi Di Stefano; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1265-1272
Object detection in images withstanding significant clutter and occlusion is still a challenging task whenever the object surface is characterized by poor informative content. We propose to tackle this problem by a compact and distinctive representation of groups of neighboring line segments aggregated over limited spatial supports and invariant to rotation, translation and scale changes. Peculiarly, our proposal allows for leveraging on the inherent strengths of descriptor-based approaches, i.e. robustness to occlusion and clutter and scalability with respect to the size of the model library, also when dealing with scarcely textured objects.
****

Bird Part Localization Using Exemplar-Based Models with Enforced Pose and Subcategory Consistency
Jiongxin Liu, Peter N. Belhumeur; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2520-2527
In this paper, we propose a novel approach for bird part localization, targeting fine-grained categories with wide variations in appearance due to different poses (including aspect and orientation) and subcategories. As it is challenging to represent such variations across a large set of diverse samples with tractable parametric models, we turn to individual exemplars. Specifically, we extend the exemplarbased models in [4] by enforcing pose and subcategory consistency at the parts. During training, we build posespecific detectors scoring part poses across subcategories, and subcategory-specific detectors scoring part appearance across poses. At the testing stage, likely exemplars are matched to the image, suggesting part locations whose pose and subcategory consistency are well-supported by the image cues. From these hypotheses, part configuration can be predicted with very high accuracy. Experimental results demonstrate significant performance gains from our method on an extensive dataset: CUB-200-2011 [30], for both localization and classification tasks.
****

Multiple Non-rigid Surface Detection and Registration
Yi Wu, Yoshihisa Ijiri, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1992-1999
Detecting and registering nonrigid surfaces are two important research problems for computer vision. Much work has been done with the assumption that there exists only one instance in the image. In this work, we propose an algorithm that detects and registers multiple nonrigid instances of given objects in a cluttered image. Specifically, after we use low level feature points to obtain the initial matches between templates and the input image, a novel high-order affinity graph is constructed to model the consistency of local topology. A hierarchical clustering approach is then used to locate the nonrigid surfaces. To remove the outliers in the cluster, we propose a deterministic annealing approach based on the Thin Plate Spline (TPS) model. The proposed method achieves high accuracy even when the number of outliers is nineteen times larger than the inliers. As the matches may appear sparsely in each instance, we propose a TPS based match growing approach to propagate the matches. Finally, an approach that fuses feature and appearance information is proposed to register each nonrigid surface. Extensive experiments and evaluations demonstrate that the proposed algorithm achieves promising results in detecting and registering multiple non-rigid surfaces in a cluttered scene.
****

Drosophila Embryo Stage Annotation Using Label Propagation
Tomas Kazmar, Evgeny Z. Kvon, Alexander Stark, Christoph H. Lampert; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1089-1096
In this work we propose a system for automatic classification of Drosophila embryos into developmental stages. While the system is designed to solve an actual problem in biological research, we believe that the principle underlying it is in

teresting not only for biologists, but also for researchers in computer vision. The main idea is to combine two orthogonal sources of information: one is a classifier trained on strongly invariant features, which makes it applicable to images of very different conditions, but also leads to rather noisy predictions. The other is a label propagation step based on a more powerful similarity measure that however is only consistent within specific subsets of the data at a time. In our biological setup, the information sources are the shape and the staining patterns of embryo images. We show experimentally that while neither of the methods can be used by itself to achieve satisfactory results, their combination achieves prediction quality comparable to human performance.
****

Parsing IKEA Objects: Fine Pose Estimation
Joseph J. Lim, Hamed Pirsiavash, Antonio Torralba; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2992-2999
We address the problem of localizing and estimating the fine-pose of objects in the image with exact 3D models. Our main focus is to unify contributions from the 1970s with recent advances in object detection: use local keypoint detectors to find candidate poses and score global alignment of each candidate pose to the image. Moreover, we also provide a new dataset containing fine-aligned objects with their exactly matched 3D models, and a set of models for widely used objects. We also evaluate our algorithm both on object detection and fine pose estimation, and show that our method outperforms state-of-the art algorithms.
****

Corrected-Moment Illuminant Estimation
Graham D. Finlayson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1904-1911
Image colors are biased by the color of the prevaling illumination. As such the color at pixel cannot always be used directly in solving vision tasks from recognition, to tracking to general scene understanding. Illuminant estimation algorithms attempt to infer the color of the light incident in a scene and then a color cast removal step discounts the color bias due to illumination. However, despite sustained research since almost the inception of computer vision, progress has been modest. The best algorithms - now often built on top of existing feature extraction and machine learning - are only about twice as good as the simplest approaches. This paper, in effect, will show how simple moment based algorithms - such as Gray-World - can, with the addition of a simple correction step, deliver much improved illuminant estimation performance. The corrected Gray-World algorithm maps the mean image color using a fixed (per camera) 3x3 matrix transform. More generally, our moment approach employs 1st, 2nd and higher order moments - of colors or features such as color derivatives - and these again are linearly corrected to give an illuminant estimate. The question of how to correct the moments is an important one yet we will show a simple alternating least-squares training procedure suffices. Remarkably, across the major datasets - evaluated using a 3-fold cross validation procedure - our simple corrected moment approach always delivers the best results (and the performance increment is often large compared with the prior art). Significantly, outlier performance was found to be much improved.
****

Group Sparsity and Geometry Constrained Dictionary Learning for Action Recognition from Depth Maps
Jiajia Luo, Wei Wang, Hairong Qi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1809-1816
Human action recognition based on the depth information provided by commodity depth sensors is an important yet challenging task. The noisy depth maps, different lengths of action sequences, and free styles in performing actions, may cause large intra-class variations. In this paper, a new framework based on sparse coding and temporal pyramid matching (TPM) is proposed for depthbased human action recognition. Especially, a discriminative class-specific dictionary learning algorithm is proposed for sparse coding. By adding the group sparsity and geometry constraints, features can be well reconstructed by the sub-dictionary belonging

to the same class, and the geometry relationships among features are also kept i
n the calculated coefficients. The proposed approach is evaluated on two benchma
rk datasets captured by depth cameras. Experimental results show that the propos
ed algorithm repeatedly achieves superior performance to the state of the art al
gorithms. Moreover, the proposed dictionary learning method also outperforms cla
ssic dictionary learning approaches.
****

Online Video SEEDS for Temporal Window Objectness
Michael Van Den Bergh, Gemma Roig, Xavier Boix, Santiago Manen, Luc Van Gool; Pr
oceedings of the IEEE International Conference on Computer Vision (ICCV), 2013,
pp. 377-384
Superpixel and objectness algorithms are broadly used as a pre-processing step t
o generate support regions and to speed-up further computations. Recently, many
algorithms have been extended to video in order to exploit the temporal consiste
ncy between frames. However, most methods are computationally too expensive for
real-time applications. We introduce an online, real-time video superpixel algor
ithm based on the recently proposed SEEDS superpixels. A new capability is incor
porated which delivers multiple diverse samples (hypotheses) of superpixels in t
he same image or video sequence. The multiple samples are shown to provide a str
ong cue to efficiently measure the objectness of image windows, and we introduce
 the novel concept of objectness in temporal windows. Experiments show that the
video superpixels achieve comparable performance to state-of-the-art offline met
hods while running at 30 fps on a single 2.8 GHz i7 CPU. State-of-the-art perfor
mance on objectness is also demonstrated, yet orders of magnitude faster and ext
ended to temporal windows in video.
****

Fast Subspace Search via Grassmannian Based Hashing
Xu Wang, Stefan Atev, John Wright, Gilad Lerman; Proceedings of the IEEE Interna
tional Conference on Computer Vision (ICCV), 2013, pp. 2776-2783
The problem of efficiently deciding which of a database of models is most simila
r to a given input query arises throughout modern computer vision. Motivated by
applications in recognition, image retrieval and optimization, there has been si
gnificant recent interest in the variant of this problem in which the database m
odels are linear subspaces and the input is either a point or a subspace. Curren
t approaches to this problem have poor scaling in high dimensions, and may not g
uarantee sublinear query complexity. We present a new approach to approximate ne
arest subspace search, based on a simple, new locality sensitive hash for subspa
ces. Our approach allows point-tosubspace query for a database of subspaces of a
rbitrary dimension d, in a time that depends sublinearly on the number of subspa
ces in the database. The query complexity of our algorithm is linear in the ambi
ent dimension D, allowing it to be directly applied to high-dimensional imagery
data. Numerical experiments on model problems in image repatching and automatic
face recognition confirm the advantages of our algorithm in terms of both speed
and accuracy.
****

Data-Driven 3D Primitives for Single Image Understanding
David F. Fouhey, Abhinav Gupta, Martial Hebert; Proceedings of the IEEE Internat
ional Conference on Computer Vision (ICCV), 2013, pp. 3392-3399
What primitives should we use to infer the rich 3D world behind an image? We arg
ue that these primitives should be both visually discriminative and geometricall
y informative and we present a technique for discovering such primitives. We dem
onstrate the utility of our primitives by using them to infer 3D surface normals
 given a single image. Our technique substantially outperforms the state-of-the-
art and shows improved cross-dataset performance.
****

Partial Enumeration and Curvature Regularization
Carl Olsson, Johannes Ulen, Yuri Boykov, Vladimir Kolmogorov; Proceedings of the
 IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2936-2943
Energies with high-order non-submodular interactions have been shown to be very
useful in vision due to their high modeling power. Optimization of such energies

, however, is generally NP-hard. A naive approach that works for small problem i nstances is exhaustive search, that is, enumeration of all possible labelings of the underlying graph. We propose a general minimization approach for large grap hs based on enumeration of labelings of certain small patches. This partial enum eration technique reduces complex highorder energy formulations to pairwise Cons traint Satisfaction Problems with unary costs (uCSP), which can be efficiently s olved using standard methods like TRW-S. Our approach outperforms a number of ex isting state-of-the-art algorithms on well known difficult problems (e.g. curvat ure regularization, stereo, deconvolution); it gives near global minimum and bet ter speed. Our main application of interest is curvature regularization. In the context of segmentation, our partial enumeration technique allows to evaluate cu rvature directly on small patches using a novel integral geometry approach. 1
****

Fast Face Detector Training Using Tailored Views

Kristina Scherbaum, James Petterson, Rogerio S. Feris, Volker Blanz, Hans-Peter Seidel; Proceedings of the IEEE International Conference on Computer Vision (ICC V), 2013, pp. 2848-2855

Face detection is an important task in computer vision and often serves as the f irst step for a variety of applications. State-of-the-art approaches use efficie nt learning algorithms and train on large amounts of manually labeled imagery. A cquiring appropriate training images, however, is very time-consuming and does n ot guarantee that the collected training data is representative in terms of data variability. Moreover, available data sets are often acquired under controlled settings, restricting, for example, scene illumination or 3D head pose to a narr ow range. This paper takes a look into the automated generation of adaptive trai ning samples from a 3D morphable face model. Using statistical insights, the tai lored training data guarantees full data variability and is enriched by arbitrar y facial attributes such as age or body weight. Moreover, it can automatically a dapt to environmental constraints, such as illumination or viewing angle of reco rded video footage from surveillance cameras. We use the tailored imagery to tra in a new many-core implementation of Viola Jones' AdaBoost object detection fram ework. The new implementation is not only faster but also enables the use of mul tiple feature channels such as color features at training time. In our experimen ts we trained seven view-dependent face detectors and evaluate these on the Face Detection Data Set and Benchmark (FDDB). Our experiments show that the use of t ailored training imagery outperforms state-of-the-art approaches on this challen ging dataset.
****

Image Retrieval Using Textual Cues

Anand Mishra, Karteek Alahari, C.V. Jawahar; Proceedings of the IEEE Internation al Conference on Computer Vision (ICCV), 2013, pp. 3040-3047

We present an approach for the text-to-image retrieval problem based on textual content present in images. Given the recent developments in understanding text i n images, an appealing approach to address this problem is to localize and recog nize the text, and then query the database, as in a text retrieval problem. We s how that such an approach, despite being based on state-of-the-art methods, is i nsufficient, and propose a method, where we do not rely on an exact localization and recognition pipeline. We take a query-driven search approach, where we find approximate locations of characters in the text query, and then impose spatial constraints to generate a ranked list of images in the database. The retrieval p erformance is evaluated on public scene text datasets as well as three large dat asets, namely IIIT scene text retrieval, Sports-10K and TV series-1M, we introdu ce.
****

Fluttering Pattern Generation Using Modified Legendre Sequence for Coded Exposur e Imaging

Hae-Gon Jeon, Joon-Young Lee, Yudeog Han, Seon Joo Kim, In So Kweon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1001-1008

Finding a good binary sequence is critical in determining the performance of the

coded exposure imaging, but previous methods mostly rely on a random search for finding the binary codes, which could easily fail to find good long sequences due to the exponentially growing search space. In this paper, we present a new computationally efficient algorithm for generating the binary sequence, which is especially well suited for longer sequences. We show that the concept of the low autocorrelation binary sequence that has been well exploited in the information theory community can be applied for generating the fluttering patterns of the shutter, propose a new measure of a good binary sequence, and present a new algorithm by modifying the Legendre sequence for the coded exposure imaging. Experiments using both synthetic and real data show that our new algorithm consistently generates better binary sequences for the coded exposure problem, yielding better deblurring and resolution enhancement results compared to the previous methods for generating the binary codes.

****

Prime Object Proposals with Randomized Prim's Algorithm

Santiago Manen, Matthieu Guillaumin, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2536-2543

Generic object detection is the challenging task of proposing windows that localize all the objects in an image, regardless of their classes. Such detectors have recently been shown to benefit many applications such as speedingup class-specific object detection, weakly supervised learning of object detectors and object discovery. In this paper, we introduce a novel and very efficient method for generic object detection based on a randomized version of Prim's algorithm. Using the connectivity graph of an image's superpixels, with weights modelling the probability that neighbouring superpixels belong to the same object, the algorithm generates random partial spanning trees with large expected sum of edge weights. Object localizations are proposed as bounding-boxes of those partial trees. Our method has several benefits compared to the stateof-the-art. Thanks to the efficiency of Prim's algorithm, it samples proposals very quickly: 1000 proposals are obtained in about 0.7s. With proposals bound to superpixel boundaries yet diversified by randomization, it yields very high detection rates and windows that tightly fit objects. In extensive experiments on the challenging PASCAL VOC 2007 and 2012 and SUN2012 benchmark datasets, we show that our method improves over state-of-the-art competitors for a wide range of evaluation scenarios.

****

Optimization Problems for Fast AAM Fitting in-the-Wild

Georgios Tzimiropoulos, Maja Pantic; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 593-600

We describe a very simple framework for deriving the most-well known optimization problems in Active Appearance Models (AAMs), and most importantly for providing efficient solutions. Our formulation results in two optimization problems for fast and exact AAM fitting, and one new algorithm which has the important advantage of being applicable to 3D. We show that the dominant cost for both forward and inverse algorithms is a few times mN which is the cost of projecting an image onto the appearance subspace. This makes both algorithms not only computationally realizable but also very attractive speed-wise for most current systems. Because exact AAM fitting is no longer computationally prohibitive, we trained AAMs in-the-wild with the goal of investigating whether AAMs benefit from such a training process. Our results show that although we did not use sophisticated shape priors, robust features or robust norms for improving performance, AAMs perform notably well and in some cases comparably with current state-ofthe-art methods. We provide Matlab source code for training, fitting and reproducing the results presented in this paper at http://ibug.doc.ic.ac.uk/resources.

****

Semi-supervised Robust Dictionary Learning via Efficient l-Norms Minimization

Hua Wang, Feiping Nie, Weidong Cai, Heng Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1145-1152

Representing the raw input of a data set by a set of relevant codes is crucial to many computer vision applications. Due to the intrinsic sparse property of real-world data, dictionary learning, in which the linear decomposition of a data p

oint uses a set of learned dictionary bases, i.e., codes, has demonstrated state-of-the-art performance. However, traditional dictionary learning methods suffer from three weaknesses: sensitivity to noisy and outlier samples, difficulty to determine the optimal dictionary size, and incapability to incorporate supervision information. In this paper, we address these weaknesses by learning a Semi-Supervised Robust Dictionary (SSR-D). Specifically, we use the l 2,0 + norm as the loss function to improve the robustness against outliers, and develop a new structured sparse regularization to incorporate the supervision information in dictionary learning, without incurring additional parameters. Moreover, the optimal dictionary size is automatically learned from the input data. Minimizing the derived objective function is challenging because it involves many non-smooth l 2,0 + -norm terms. We present an efficient algorithm to solve the problem with a rigorous proof of the convergence of the algorithm. Extensive experiments are presented to show the superior performance of the proposed method.

****

Cosegmentation and Cosketch by Unsupervised Learning

Jifeng Dai, Ying Nian Wu, Jie Zhou, Song-Chun Zhu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1305-1312

Cosegmentation refers to the problem of segmenting multiple images simultaneously by exploiting the similarities between the foreground and background regions in these images. The key issue in cosegmentation is to align common objects between these images. To address this issue, we propose an unsupervised learning framework for cosegmentation, by coupling cosegmentation with what we call "cosketch". The goal of cosketch is to automatically discover a codebook of deformable shape templates shared by the input images. These shape templates capture distinct image patterns and each template is matched to similar image patches in different images. Thus the cosketch of the images helps to align foreground objects, thereby providing crucial information for cosegmentation. We present a statistical model whose energy function couples cosketch and cosegmentation. We then present an unsupervised learning algorithm that performs cosketch and cosegmentation by energy minimization. Experiments show that our method outperforms state of the art methods for cosegmentation on the challenging MSRC and iCoseg datasets. We also illustrate our method on a new dataset called Coseg-Rep where cosegmentation can be performed within a single image with repetitive patterns.

****

Joint Learning of Discriminative Prototypes and Large Margin Nearest Neighbor Classifiers

Martin Kostinger, Paul Wohlhart, Peter M. Roth, Horst Bischof; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3112-3119

In this paper, we raise important issues concerning the evaluation complexity of existing Mahalanobis metric learning methods. The complexity scales linearly with the size of the dataset. This is especially cumbersome on large scale or for real-time applications with limited time budget. To alleviate this problem we propose to represent the dataset by a fixed number of discriminative prototypes. In particular, we introduce a new method that jointly chooses the positioning of prototypes and also optimizes the Mahalanobis distance metric with respect to these. We show that choosing the positioning of the prototypes and learning the metric in parallel leads to a drastically reduced evaluation effort while maintaining the discriminative essence of the original dataset. Moreover, for most problems our method performing k-nearest prototype (k-NP) classification on the condensed dataset leads to even better generalization compared to k-NN classification using all data. Results on a variety of challenging benchmarks demonstrate the power of our method. These include standard machine learning datasets as well as the challenging Public Figures Face Database. On the competitive machine learning benchmarks we are comparable to the state-of-the-art while being more efficient. On the face benchmark we clearly outperform the state-of-the-art in Mahalanobis metric learning with drastically reduced evaluation effort.

****

Joint Optimization for Consistent Multiple Graph Matching

Junchi Yan, Yu Tian, Hongyuan Zha, Xiaokang Yang, Ya Zhang, Stephen M. Chu; Proc

eedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1649-1656
The problem of graph matching in general is NP-hard and approaches have been proposed for its suboptimal solution, most focusing on finding the one-to-one node mapping between two graphs. A more general and challenging problem arises when one aims to find consistent mappings across a number of graphs more than two. Conventional graph pair matching methods often result in mapping inconsistency since the mapping between two graphs can either be determined by pair mapping or by an additional anchor graph. To address this issue, a novel formulation is derived which is maximized via alternating optimization. Our method enjoys several advantages: 1) the mappings are jointly optimized rather than sequentially performed by applying pair matching, allowing the global affinity information across graphs can be propagated and explored; 2) the number of concerned variables to optimize is in linear with the number of graphs, being superior to local pair matching resulting in O(n 2) variables; 3) the mapping consistency constraints are analytically satisfied during optimization; and 4) off-the-shelf graph pair matching solvers can be reused under the proposed framework in an â■■out-of-theboxâ■■ fashion. Competitive results on both the synthesized data and the real data are reported, by varying the level of deformation, outliers and edge densities.
****

Scene Collaging: Analysis and Synthesis of Natural Images with Semantic Layers
Phillip Isola, Ce Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3048-3055
To quickly synthesize complex scenes, digital artists often collage together visual elements from multiple sources: for example, mountains from New Zealand behind a Scottish castle with wisps of Saharan sand in front. In this paper, we propose to use a similar process in order to parse a scene. We model a scene as a collage of warped, layered objects sampled from labeled, reference images. Each object is related to the rest by a set of support constraints. Scene parsing is achieved through analysis-by-synthesis. Starting with a dataset of labeled exemplar scenes, we retrieve a dictionary of candidate object segments thatttriginal in lch imcchie e query ryEtem m-ige age. We then combine elements of this set into a "scene collage" that explains the query image. Beyond just assigning object labels to pixels, scene collaging produces a lot more information such as the number of each type of object in the scene, how they support one another, the ordinal depth of each object, and, to some degree, occluded content. We exploit this representation for several applications: image editing, random scene synthesis, and image-to-anaglyph.
****

Quadruplet-Wise Image Similarity Learning
Marc T. Law, Nicolas Thome, Matthieu Cord; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 249-256
This paper introduces a novel similarity learning framework. Working with inequality constraints involving quadruplets of images, our approach aims at efficiently modeling similarity from rich or complex semantic label relationships. From these quadruplet-wise constraints, we propose a similarity learning framework relying on a convex optimization scheme. We then study how our metric learning scheme can exploit specific class relationships, such as class ranking (relative attributes), and class taxonomy. We show that classification using the learned metrics gets improved performance over state-of-the-art methods on several datasets. We also evaluate our approach in a new application to learn similarities between webpage screenshots in a fully unsupervised way.
****

Facial Action Unit Event Detection by Cascade of Tasks
Xiaoyu Ding, Wen-Sheng Chu, Fernando De La Torre, Jeffery F. Cohn, Qiao Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2400-2407
Automatic facial Action Unit (AU) detection from video is a long-standing problem in facial expression analysis. AU detection is typically posed as a classification problem between frames or segments of positive examples and negative ones,

where existing work emphasizes the use of different features or classifiers. In this paper, we propose a method called Cascade of Tasks (CoT) that combines the use of different tasks (i.e., frame, segment and transition) for AU event detection. We train CoT in a sequential manner embracing diversity, which ensures robustness and generalization to unseen data. In addition to conventional framebased metrics that evaluate frames independently, we propose a new event-based metric to evaluate detection performance at event-level. We show how the CoT method consistently outperforms state-of-the-art approaches in both frame-based and event-based metrics, across three public datasets that differ in complexity: CK+, FERA and RUFACS.
****

Cascaded Shape Space Pruning for Robust Facial Landmark Detection
Xiaowei Zhao, Shiguang Shan, Xiujuan Chai, Xilin Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1033-1040
In this paper, we propose a novel cascaded face shape space pruning algorithm for robust facial landmark detection. Through progressively excluding the incorrect candidate shapes, our algorithm can accurately and efficiently achieve the globally optimal shape configuration. Specifically, individual landmark detectors are firstly applied to eliminate wrong candidates for each landmark. Then, the candidate shape space is further pruned by jointly removing incorrect shape configurations. To achieve this purpose, a discriminative structure classifier is designed to assess the candidate shape configurations. Based on the learned discriminative structure classifier, an efficient shape space pruning strategy is proposed to quickly reject most incorrect candidate shapes while preserve the true shape. The proposed algorithm is carefully evaluated on a large set of real world face images. In addition, comparison results on the publicly available BioID and LFW face databases demonstrate that our algorithm outperforms some state-of-the-art algorithms.
****

Efficient Higher-Order Clustering on the Grassmann Manifold
Suraj Jain, Venu Madhav Govindu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3511-3518
The higher-order clustering problem arises when data is drawn from multiple subspaces or when observations fit a higher-order parametric model. Most solutions to this problem either decompose higher-order similarity measures for use in spectral clustering or explicitly use low-rank matrix representations. In this paper we present our approach of Sparse Grassmann Clustering (SGC) that combines attributes of both categories. While we decompose the higherorder similarity tensor, we cluster data by directly finding a low dimensional representation without explicitly building a similarity matrix. By exploiting recent advances in online estimation on the Grassmann manifold (GROUSE) we develop an efficient and accurate algorithm that works with individual columns of similarities or partial observations thereof. Since it avoids the storage and decomposition of large similarity matrices, our method is efficient, scalable and has low memory requirements even for large-scale data. We demonstrate the performance of our SGC method on a variety of segmentation problems including planar segmentation of Kinect depth maps and motion segmentation of the Hopkins 155 dataset for which we achieve performance comparable to the state-of-the-art.
****

A Scalable Unsupervised Feature Merging Approach to Efficient Dimensionality Reduction of High-Dimensional Visual Data
Lingqiao Liu, Lei Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3008-3015
To achieve a good trade-off between recognition accuracy and computational efficiency, it is often needed to reduce high-dimensional visual data to medium-dimensional ones. For this task, even applying a simple full-matrixbased linear projection causes significant computation and memory use. When the number of visual data is large, how to efficiently learn such a projection could even become a problem. The recent feature merging approach offers an efficient way to reduce the dimensionality, which only requires a single scan of features to perform reducti

on. However, existing merging algorithms do not scale well with highdimensional data, especially in the unsupervised case. To address this problem, we formulate unsupervised feature merging as a PCA problem imposed with a special structure constraint. By exploiting its connection with kmeans, we transform this constrained PCA problem into a feature clustering problem. Moreover, we employ the hashing technique to improve its scalability. These produce a scalable feature merging algorithm for our dimensionality reduction task. In addition, we develop an extension of this method by leveraging the neighborhood structure in the data to further improve dimensionality reduction performance. In further, we explore the incorporation of bipolar merging a variant of merging function which allows the subtraction operation into our algorithms. Through three applications in visual recognition, we demonstrate that our methods can not only achieve good dimensionality reduction performance with little computational cost but also help to create more powerful representation at both image level and local feature level.

****

Deformable Part Descriptors for Fine-Grained Recognition and Attribute Prediction

Ning Zhang, Ryan Farrell, Forrest Iandola, Trevor Darrell; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 729-736

Recognizing objects in fine-grained domains can be extremely challenging due to the subtle differences between subcategories. Discriminative markings are often highly localized, leading traditional object recognition approaches to struggle with the large pose variation often present in these domains. Pose-normalization seeks to align training exemplars, either piecewise by part or globally for the whole object, effectively factoring out differences in pose and in viewing angle. Prior approaches relied on computationally-expensive filter ensembles for part localization and required extensive supervision. This paper proposes two pose-normalized descriptors based on computationally-efficient deformable part models. The first leverages the semantics inherent in strongly-supervised DPM parts. The second exploits weak semantic annotations to learn cross-component correspondences, computing pose-normalized descriptors from the latent parts of a weakly-supervised DPM. These representations enable pooling across pose and viewpoint, in turn facilitating tasks such as fine-grained recognition and attribute prediction. Experiments conducted on the Caltech-UCSD Birds 200 dataset and Berkeley Human Attribute dataset demonstrate significant improvements over state-of-art algorithms.

****

Compensating for Motion during Direct-Global Separation

Supreeth Achar, Stephen T. Nuske, Srinivasa G. Narasimhan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1481-1488

Separating the direct and global components of radiance can aid shape recovery algorithms and can provide useful information about materials in a scene. Practical methods for finding the direct and global components use multiple images captured under varying illumination patterns and require the scene, light source and camera to remain stationary during the image acquisition process. In this paper, we develop a motion compensation method that relaxes this condition and allows direct-global separation to be performed on video sequences of dynamic scenes captured by moving projector-camera systems. Key to our method is being able to register frames in a video sequence to each other in the presence of time varying, high frequency active illumination patterns. We compare our motion compensated method to alternatives such as single shot separation and frame interleaving as well as ground truth. We present results on challenging video sequences that include various types of motions and deformations in scenes that contain complex materials like fabric, skin, leaves and wax.

****

Shufflets: Shared Mid-level Parts for Fast Object Detection

Iasonas Kokkinos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1393-1400

We present a method to identify and exploit structures that are shared across different object categories, by using sparse coding to learn a shared basis for th

e 'part' and 'root' templates of Deformable Part Models (DPMs). Our first contri bution consists in using Shift-Invariant Sparse Coding (SISC) to learn mid-level elements that can translate during coding. This results in systematically bette r approximations than those attained using standard sparse coding. To emphasize that the learned mid-level structures are shiftable we call them shufflets. Our second contribution consists in using the resulting score to construct probabili stic upper bounds to the exact template scores, instead of taking them 'at face value' as is common in current works. We integrate shufflets in DualTree Branch-and-Bound and cascade-DPMs and demonstrate that we can achieve a substantial acc eleration, with practically no loss in performance.
****

GrabCut in One Cut
Meng Tang, Lena Gorelick, Olga Veksler, Yuri Boykov; Proceedings of the IEEE Int ernational Conference on Computer Vision (ICCV), 2013, pp. 1769-1776
Among image segmentation algorithms there are two major groups: (a) methods assu ming known appearance models and (b) methods estimating appearance models jointl y with segmentation. Typically, the first group optimizes appearance log-likelih oods in combination with some spacial regularization. This problem is relatively simple and many methods guarantee globally optimal results. The second group tr eats model parameters as additional variables transforming simple segmentation e nergies into highorder NP-hard functionals (Zhu-Yuille, Chan-Vese, GrabCut, etc) . It is known that such methods indirectly minimize the appearance overlap betwe en the segments. We propose a new energy term explicitly measuring L 1 distance between the object and background appearance models that can be globally maximiz ed in one graph cut. We show that in many applications our simple term makes NP-hard segmentation functionals unnecessary. Our one cut algorithm effectively rep laces approximate iterative optimization techniques based on block coordinate de scent.
****

Coupling Alignments with Recognition for Still-to-Video Face Recognition
Zhiwu Huang, Xiaowei Zhao, Shiguang Shan, Ruiping Wang, Xilin Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3296-3 303
The Still-to-Video (S2V) face recognition systems typically need to match faces in low-quality videos captured under unconstrained conditions against high quali ty still face images, which is very challenging because of noise, image blur, lo w face resolutions, varying head pose, complex lighting, and alignment difficult y. To address the problem, one solution is to select the frames of 'best quality ' from videos (hereinafter called quality alignment in this paper). Meanwhile, t he faces in the selected frames should also be geometrically aligned to the stil l faces offline well-aligned in the gallery. In this paper, we discover that the interactions among the three tasks-quality alignment, geometric alignment and f ace recognition-can benefit from each other, thus should be performed jointly. W ith this in mind, we propose a Coupling Alignments with Recognition (CAR) method to tightly couple these tasks via low-rank regularized sparse representation in a unified framework. Our method makes the three tasks promote mutually by a joi nt optimization in an Augmented Lagrange Multiplier routine. Extensive experimen ts on two challenging S2V datasets demonstrate that our method outperforms the s tate-of-the-art methods impressively.
****

Stacked Predictive Sparse Coding for Classification of Distinct Regions in Tumor Histopathology
Hang Chang, Yin Zhou, Paul Spellman, Bahram Parvin; Proceedings of the IEEE Inte rnational Conference on Computer Vision (ICCV), 2013, pp. 169-176
Image-based classification of histology sections, in terms of distinct component s (e.g., tumor, stroma, normal), provides a series of indices for tumor composit ion. Furthermore, aggregation of these indices, from each whole slide image (WSI) in a large cohort, can provide predictive models of the clinical outcome. Howe ver, performance of the existing techniques is hindered as a result of large tec hnical variations and biological heterogeneities that are always present in a la

rge cohort. We propose a system that automatically learns a series of basis func tions for representing the underlying spatial distribution using stacked predict ive sparse decomposition (PSD). The learned representation is then fed into the spatial pyramid matching framework (SPM) with a linear SVM classifier. The syste m has been evaluated for classification of (a) distinct histological components for two cohorts of tumor types, and (b) colony organization of normal and malign ant cell lines in 3D cell culture models. Throughput has been increased through the utility of graphical processing unit (GPU), and evaluation indicates a super ior performance results, compared with previous research.
****

Query-Adaptive Asymmetrical Dissimilarities for Visual Object Retrieval
Cai-Zhi Zhu, Herve Jegou, Shin Ichi Satoh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1705-1712
Visual object retrieval aims at retrieving, from a collection of images, all tho se in which a given query object appears. It is inherently asymmetric: the query object is mostly included in the database image, while the converse is not nece ssarily true. However, existing approaches mostly compare the images with symmet rical measures, without considering the different roles of query and database. T his paper first measure the extent of asymmetry on large-scale public datasets r eflecting this task. Considering the standard bag-of-words representation, we th en propose new asymmetrical dissimilarities accounting for the different inlier ratios associated with query and database images. These asymmetrical measures de pend on the query, yet they are compatible with an inverted file structure, with out noticeably impacting search efficiency. Our experiments show the benefit of our approach, and show that the visual object retrieval task is better treated a symmetrically, in the spirit of state-of-the-art text retrieval.
****

Direct Optimization of Frame-to-Frame Rotation
Laurent Kneip, Simon Lynen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2352-2359
This work makes use of a novel, recently proposed epipolar constraint for comput ing the relative pose between two calibrated images. By enforcing the coplanarit y of epipolar plane normal vectors, it constrains the three degrees of freedom o f the relative rotation between two camera views directly--independently of the translation. The present paper shows how the approach can be extended to n point s, and translated into an efficient eigenvalue minimization over the three rotat ional degrees of freedom. Each iteration in the non-linear optimization has cons tant execution time, independently of the number of features. Two global optimiz ation approaches are proposed. The first one consists of an efficient Levenberg- Marquardt scheme with randomized initial value, which already leads to stable an d accurate results. The second scheme consists of a globally optimal branch-and- bound algorithm based on a bound on the eigenvalue variation derived from symmet ric eigenvalue-perturbation theory. Analysis of the cost function reveals insigh ts into the nature of a specific relative pose problem, and outlines the complex ity under different conditions. The algorithm shows state-of-the-art performance w.r.t. essential-matrix based solutions, and a frameto-frame application to a v ideo sequence immediately leads to an alternative, real-time visual odometry sol ution. Note: All algorithms in this paper are made available in the OpenGV libra ry. Please visit http://laurentkneip.github.io/opengv
****

Unsupervised Intrinsic Calibration from a Single Frame Using a "Plumb-Line" Appr oach
R. Melo, M. Antunes, J.P. Barreto, G. Falcao, N. Goncalves; Proceedings of the I EEE International Conference on Computer Vision (ICCV), 2013, pp. 537-544
Estimating the amount and center of distortion from lines in the scene has been addressed in the literature by the socalled "plumb-line" approach. In this paper we propose a new geometric method to estimate not only the distortion parameter s but the entire camera calibration (up to an "angular" scale factor) using a mi nimum of 3 lines. We propose a new framework for the unsupervised simultaneous d etection of natural image of lines and camera parameters estimation, enabling a

robust calibration from a single image. Comparative experiments with existing au
tomatic approaches for the distortion estimation and with ground truth data are
presented.
****

Weakly Supervised Learning of Image Partitioning Using Decision Trees with Struc
tured Split Criteria
Christoph Straehle, Ullrich Koethe, Fred A. Hamprecht; Proceedings of the IEEE I
nternational Conference on Computer Vision (ICCV), 2013, pp. 1849-1856
We propose a scheme that allows to partition an image into a previously unknown
number of segments, using only minimal supervision in terms of a few must-link a
nd cannotlink annotations. We make no use of regional data terms, learning inste
ad what constitutes a likely boundary between segments. Since boundaries are onl
y implicitly specified through cannot-link constraints, this is a hard and nonco
nvex latent variable problem. We address this problem in a greedy fashion using
a randomized decision tree on features associated with interpixel edges. We use
a structured purity criterion during tree construction and also show how a backt
racking strategy can be used to prevent the greedy search from ending up in poor
 local optima. The proposed strategy is compared with prior art on natural image
s.
****

Discriminant Tracking Using Tensor Representation with Semi-supervised Improveme
nt
Jin Gao, Junliang Xing, Weiming Hu, Steve Maybank; Proceedings of the IEEE Inter
national Conference on Computer Vision (ICCV), 2013, pp. 1569-1576
Visual tracking has witnessed growing methods in object representation, which is
 crucial to robust tracking. The dominant mechanism in object representation is
using image features encoded in a vector as observations to perform tracking, wi
thout considering that an image is intrinsically a matrix, or a 2 nd -order tens
or. Thus approaches following this mechanism inevitably lose a lot of useful inf
ormation, and therefore cannot fully exploit the spatial correlations within the
 2D image ensembles. In this paper, we address an image as a 2 nd -order tensor
in its original form, and find a discriminative linear embedding space approxima
tion to the original nonlinear submanifold embedded in the tensor space based on
 the graph embedding framework. We specially design two graphs for characterizin
g the intrinsic local geometrical structure of the tensor space, so as to retain
 more discriminant information when reducing the dimension along certain tensor
dimensions. However, spatial correlations within a tensor are not limited to the
 elements along these dimensions. This means that some part of the discriminant
information may not be encoded in the embedding space. We introduce a novel tech
nique called semi-supervised improvement to iteratively adjust the embedding spa
ce to compensate for the loss of discriminant information, hence improving the p
erformance of our tracker. Experimental results on challenging videos demonstrat
e the effectiveness and robustness of the proposed tracker.
****

Adapting Classification Cascades to New Domains
Vidit Jain, Sachin Sudhakar Farfade; Proceedings of the IEEE International Confe
rence on Computer Vision (ICCV), 2013, pp. 105-112
Classification cascades have been very effective for object detection. Such a ca
scade fails to perform well in data domains with variations in appearances that
may not be captured in the training examples. This limited generalization severe
ly restricts the domains for which they can be used effectively. A common approa
ch to address this limitation is to train a new cascade of classifiers from scra
tch for each of the new domains. Building separate detectors for each of the dif
ferent domains requires huge annotation and computational effort, making it not
scalable to a large number of data domains. Here we present an algorithm for qui
ckly adapting a pre-trained cascade of classifiers using a small number of label
ed positive instances from a different yet similar data domain. In our experimen
ts with images of human babies and human-like characters from movies, we demonst
rate that the adapted cascade significantly outperforms both of the original cas
cade and the one trained from scratch using the given training examples.

****
Collaborative Active Learning of a Kernel Machine Ensemble for Recognition
Gang Hua, Chengjiang Long, Ming Yang, Yan Gao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1209-1216
Active learning is an effective way of engaging users to interactively train models for visual recognition. The vast majority of previous works, if not all of them, focused on active learning with a single human oracle. The problem of active learning with multiple oracles in a collaborative setting has not been well explored. Moreover, most of the previous works assume that the labels provided by the human oracles are noise free, which may often be violated in reality. We present a collaborative computational model for active learning with multiple human oracles. It leads to not only an ensemble kernel machine that is robust to label noises, but also a principled label quality measure to online detect irresponsible labelers. Instead of running independent active learning processes for each individual human oracle, our model captures the inherent correlations among the labelers through shared data among them. Our simulation experiments and experiments with real crowd-sourced noisy labels demonstrated the efficacy of our model.
****
Accurate and Robust 3D Facial Capture Using a Single RGBD Camera
Yen-Lin Chen, Hsiang-Tao Wu, Fuhao Shi, Xin Tong, Jinxiang Chai; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3615-3622
This paper presents an automatic and robust approach that accurately captures high-quality 3D facial performances using a single RGBD camera. The key of our approach is to combine the power of automatic facial feature detection and image-based 3D nonrigid registration techniques for 3D facial reconstruction. In particular, we develop a robust and accurate image-based nonrigid registration algorithm that incrementally deforms a 3D template mesh model to best match observed depth image data and important facial features detected from single RGBD images. The whole process is fully automatic and robust because it is based on single frame facial registration framework. The system is flexible because it does not require any strong 3D facial priors such as blendshape models. We demonstrate the power of our approach by capturing a wide range of 3D facial expressions using a single RGBD camera and achieve state-of-the-art accuracy by comparing against alternative methods.
****
Domain Adaptive Classification
Fatemeh Mirrashed, Mohammad Rastegari; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2608-2615
We propose an unsupervised domain adaptation method that exploits intrinsic compact structures of categories across different domains using binary attributes. Our method directly optimizes for classification in the target domain. The key insight is finding attributes that are discriminative across categories and predictable across domains. We achieve a performance that significantly exceeds the state-of-the-art results on standard benchmarks. In fact, in many cases, our method reaches the same-domain performance, the upper bound, in unsupervised domain adaptation scenarios.
****
GOSUS: Grassmannian Online Subspace Updates with Structured-Sparsity
Jia Xu, Vamsi K. Ithapu, Lopamudra Mukherjee, James M. Rehg, Vikas Singh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3376-3383
We study the problem of online subspace learning in the context of sequential observations involving structured perturbations. In online subspace learning, the observations are an unknown mixture of two components presented to the model sequentially -the main effect which pertains to the subspace and a residual/error term. If no additional requirement is imposed on the residual, it often corresponds to noise terms in the signal which were unaccounted for by the main effect. To remedy this, one may impose 'structural' contiguity, which has the intended effect of leveraging the secondary terms as a covariate that helps the estimation

of the subspace itself, instead of merely serving as a noise residual. We show t
hat the corresponding online estimation procedure can be written as an approxima
te optimization process on a Grassmannian. We propose an efficient numerical sol
ution, GOSUS, Grassmannian Online ficintnnumeriallsowith n,GGOSSUUS,GGrasssmaafo
r this problem. GOSUS is expressive enough in modeling both homogeneous perturba
tions of the subspace and structural contiguities of outliers, and after certain
 manipulations, solvable via an alternating direction method of multipliers (ADM
M). We evaluate the empirical performance of this algorithm on two problems of i
nterest: online background subtraction and online multiple face tracking, and de
monstrate that it achieves competitive performance with the state-of-the-art in
near real time.
****

Analysis of Scores, Datasets, and Models in Visual Saliency Prediction
Ali Borji, Hamed R. Tavakoli, Dicky N. Sihite, Laurent Itti; Proceedings of the
IEEE International Conference on Computer Vision (ICCV), 2013, pp. 921-928
Significant recent progress has been made in developing high-quality saliency mo
dels. However, less effort has been undertaken on fair assessment of these model
s, over large standardized datasets and correctly addressing confounding factors
. In this study, we pursue a critical and quantitative look at challenges (e.g.,
 center-bias, map smoothing) in saliency modeling and the way they affect model
accuracy. We quantitatively compare 32 state-of-the-art models (using the shuffl
ed AUC score to discount center-bias) on 4 benchmark eye movement datasets, for
prediction of human fixation locations and scanpath sequence. We also account fo
r the role of map smoothing. We find that, although model rankings vary, some (e
.g., AWS, LG, AIM, and HouNIPS) consistently outperform other models over all da
tasets. Some models work well for prediction of both fixation locations and scan
path sequence (e.g., Judd, GBVS). Our results show low prediction accuracy for m
odels over emotional stimuli from the NUSEF dataset. Our last benchmark, for the
 first time, gauges the ability of models to decode the stimulus category from s
tatistics of fixations, saccades, and model saliency values at fixated locations
. In this test, ITTI and AIM models win over other models. Our benchmark provide
s a comprehensive high-level picture of the strengths and weaknesses of many pop
ular models, and suggests future research directions in saliency modeling.
****

A Color Constancy Model with Double-Opponency Mechanisms
Shaobing Gao, Kaifu Yang, Chaoyi Li, Yongjie Li; Proceedings of the IEEE Interna
tional Conference on Computer Vision (ICCV), 2013, pp. 929-936
The double-opponent color-sensitive cells in the primary visual cortex (V1) of t
he human visual system (HVS) have long been recognized as the physiological basi
s of color constancy. We introduce a new color constancy model by imitating the
functional properties of the HVS from the retina to the double-opponent cells in
 V1. The idea behind the model originates from the observation that the color di
stribution of the responses of double-opponent cells to the input color-biased i
mages coincides well with the light source direction. Then the true illuminant c
olor of a scene is easily estimated by searching for the maxima of the separate
RGB channels of the responses of double-opponent cells in the RGB space. Our sys
tematical experimental evaluations on two commonly used image datasets show that
 the proposed model can produce competitive results in comparison to the complex
 state-of-the-art approaches, but with a simple implementation and without the n
eed for training.
****

Latent Multitask Learning for View-Invariant Action Recognition
Behrooz Mahasseni, Sinisa Todorovic; Proceedings of the IEEE International Confe
rence on Computer Vision (ICCV), 2013, pp. 3128-3135
This paper presents an approach to view-invariant action recognition, where huma
n poses and motions exhibit large variations across different camera viewpoints.
 When each viewpoint of a given set of action classes is specified as a learning
 task then multitask learning appears suitable for achieving view invariance in
recognition. We extend the standard multitask learning to allow identifying: (1)
 latent groupings of action views (i.e., tasks), and (2) discriminative action p

arts, along with joint learning of all tasks. This is because it seems reasonable to expect that certain distinct views are more correlated than some others, and thus identifying correlated views could improve recognition. Also, part-based modeling is expected to improve robustness against self-occlusion when actors are imaged from different views. Results on the benchmark datasets show that we outperform standard multitask learning by 21.9%, and the state-of-the-art alternatives by 4.5-6%.
****

Translating Video Content to Natural Language Descriptions
Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, Bernt Schiele; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 433-440
Humans use rich natural language to describe and communicate visual perceptions. In order to provide natural language descriptions for visual content, this paper combines two important ingredients. First, we generate a rich semantic representation of the visual content including e.g. object and activity labels. To predict the semantic representation we learn a CRF to model the relationships between different components of the visual input. And second, we propose to formulate the generation of natural language as a machine translation problem using the semantic representation as source language and the generated sentences as target language. For this we exploit the power of a parallel corpus of videos and textual descriptions and adapt statistical machine translation to translate between our two languages. We evaluate our video descriptions on the TACoS dataset [23], which contains video snippets aligned with sentence descriptions. Using automatic evaluation and human judgments we show significant improvements over several baseline approaches, motivated by prior work. Our translation approach also shows improvements over related work on an image description task.
****

Robust Dictionary Learning by Error Source Decomposition
Zhuoyuan Chen, Ying Wu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2216-2223
Sparsity models have recently shown great promise in many vision tasks. Using a learned dictionary in sparsity models can in general outperform predefined bases in clean data. In practice, both training and testing data may be corrupted and contain noises and outliers. Although recent studies attempted to cope with corrupted data and achieved encouraging results in testing phase, how to handle corruption in training phase still remains a very difficult problem. In contrast to most existing methods that learn the dictionary from clean data, this paper is targeted at handling corruptions and outliers in training data for dictionary learning. We propose a general method to decompose the reconstructive residual into two components: a non-sparse component for small universal noises and a sparse component for large outliers, respectively. In addition, further analysis reveals the connection between our approach and the "partial" dictionary learning approach, updating only part of the prototypes (or informative codewords) with remaining (or noisy codewords) fixed. Experiments on synthetic data as well as real applications have shown satisfactory performance of this new robust dictionary learning approach.
****

Accurate Blur Models vs. Image Priors in Single Image Super-resolution
Netalee Efrat, Daniel Glasner, Alexander Apartsin, Boaz Nadler, Anat Levin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2832-2839
Over the past decade, single image Super-Resolution (SR) research has focused on developing sophisticated image priors, leading to significant advances. Estimating and incorporating the blur model, that relates the high-res and low-res images, has received much less attention, however. In particular, the reconstruction constraint, namely that the blurred and downsampled high-res output should approximately equal the low-res input image, has been either ignored or applied with default fixed blur models. In this work, we examine the relative importance of the image prior and the reconstruction constraint. First, we show that an accura

te reconstruction constraint combined with a simple gradient regularization achi eves SR results almost as good as those of state-of-the-art algorithms with soph isticated image priors. Second, we study both empirically and theoretically the sensitivity of SR algorithms to the blur model assumed in the reconstruction con straint. We find that an accurate blur model is more important than a sophistica ted image prior. Finally, using real camera data, we demonstrate that the defaul t blur models of various SR algorithms may differ from the camera blur, typicall y leading to oversmoothed results. Our findings highlight the importance of accu rately estimating camera blur in reconstructing raw lowres images acquired by an actual camera.

****

Monte Carlo Tree Search for Scheduling Activity Recognition
Mohamed R. Amer, Sinisa Todorovic, Alan Fern, Song-Chun Zhu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1353-1360
This paper presents an efficient approach to video parsing. Our videos show a nu mber of co-occurring individual and group activities. To address challenges of t he domain, we use an expressive spatiotemporal AND-OR graph (ST-AOG) that jointl y models activity parts, their spatiotemporal relations, and context, as well as enables multitarget tracking. The standard ST-AOG inference is prohibitively ex pensive in our setting, since it would require running a multitude of detectors, and tracking their detections in a long video footage. This problem is addresse d by formulating a cost-sensitive inference of ST-AOG as Monte Carlo Tree Search (MCTS). For querying an activity in the video, MCTS optimally schedules a seque nce of detectors and trackers to be run, and where they should be applied in the space-time volume. Evaluation on the benchmark datasets demonstrates that MCTS enables two-magnitude speed-ups without compromising accuracy relative to the st andard cost-insensitive inference.

****

Multi-stage Contextual Deep Learning for Pedestrian Detection
Xingyu Zeng, Wanli Ouyang, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 121-128
Cascaded classifiers 1 have been widely used in pedestrian detection and achieve d great success. These classifiers are trained sequentially without joint optimi zation. In this paper, we propose a new deep model that can jointly train multi-stage classifiers through several stages of backpropagation. It keeps the score map output by a classifier within a local region and uses it as contextual infor mation to support the decision at the next stage. Through a specific design of t he training strategy, this deep architecture is able to simulate the cascaded cl assifiers by mining hard samples to train the network stage-by-stage. Each class ifier handles samples at a different difficulty level. Unsupervised pre-training and specifically designed stage-wise supervised training are used to regularize the optimization problem. Both theoretical analysis and experimental results sh ow that the training strategy helps to avoid overfitting. Experimental results o n three datasets (Caltech, ETH and TUD-Brussels) show that our approach outperfo rms the state-of-the-art approaches.

****

Unbiased Metric Learning: On the Utilization of Multiple Datasets and Web Images for Softening Bias
Chen Fang, Ye Xu, Daniel N. Rockmore; Proceedings of the IEEE International Conf erence on Computer Vision (ICCV), 2013, pp. 1657-1664
Many standard computer vision datasets exhibit biases due to a variety of source s including illumination condition, imaging system, and preference of dataset co llectors. Biases like these can have downstream effects in the use of vision dat asets in the construction of generalizable techniques, especially for the goal o f the creation of a classification system capable of generalizing to unseen and novel datasets. In this work we propose Unbiased Metric Learning (UML), a metric learning approach, to achieve this goal. UML operates in the following two step s: (1) By varying hyperparameters, it learns a set of less biased candidate dist ance metrics on training examples from multiple biased datasets. The key idea is to learn a neighborhood for each example, which consists of not only examples o

f the same category from the same dataset, but those from other datasets. The le
arning framework is based on structural SVM. (2) We do model validation on a set
 of weakly-labeled web images retrieved by issuing class labels as keywords to s
earch engine. The metric with best validation performance is selected. Although
the web images sometimes have noisy labels, they often tend to be less biased, w
hich makes them suitable for the validation set in our task. Cross-dataset image
 classification experiments are carried out. Results show significant performanc
e improvement on four well-known computer vision datasets.
****

Pyramid Coding for Functional Scene Element Recognition in Video Scenes
Eran Swears, Anthony Hoogs, Kim Boyer; Proceedings of the IEEE International Con
ference on Computer Vision (ICCV), 2013, pp. 345-352
Recognizing functional scene elemeeents in video scenes based on the behaviors o
f moving o bjects that interact with them is an emerging problem of interest. Ex
isting approaches have a limited ability to chhharacterize elements such as cros
s-walks, intersections, anddd buildings that have low activity, are multi-modal,
 or haveee indirect evidence. Our approach recognizes the low activvvity and mul
ti-model elements (crosswalks/intersections) by introducing a hierarchy of descr
iptive clusters to ffform a pyramid of codebooks that is sparse in the numbbber
of clusters and dense in content. The incorporation ooof local behavioral contex
t such as person-enter-building aaand vehicle-parking nearby enables the detecti
on of elemennnts that do not have direct motion-based evidence, e.g. buuuildings
. These two contributions significantly improveee scene element recognition when
 compared against thhhree state-of-the-art approaches. Results are shown on tyyy
pical ground level surveillance video and for the first time on the more complex
 Wide Area Motion Imagery.
****

Fast Object Segmentation in Unconstrained Video
Anestis Papazoglou, Vittorio Ferrari; Proceedings of the IEEE International Conf
erence on Computer Vision (ICCV), 2013, pp. 1777-1784
We present a technique for separating foreground objects from the background in
a video. Our method is fast, fully automatic, and makes minimal assumptions abou
t the video. This enables handling essentially unconstrained settings, including
 rapidly moving background, arbitrary object motion and appearance, and non-rigi
d deformations and articulations. In experiments on two datasets containing over
 1400 video shots, our method outperforms a state-of-theart background subtracti
on technique [4] as well as methods based on clustering point tracks [6, 18, 19]
. Moreover, it performs comparably to recent video object segmentation methods b
ased on object proposals [14, 16, 27], while being orders of magnitude faster.
****

Offline Mobile Instance Retrieval with a Small Memory Footprint
Jayaguru Panda, Michael S. Brown, C.V. Jawahar; Proceedings of the IEEE Internat
ional Conference on Computer Vision (ICCV), 2013, pp. 1257-1264
Existing mobile image instance retrieval applications assume a network-based usa
ge where image features are sent to a server to query an online visual database.
 In this scenario, there are no restrictions on the size of the visual database.
 This paper, however, examines how to perform this same task offline, where the
entire visual index must reside on the mobile device itself within a small memor
y footprint. Such solutions have applications on location recognition and produc
t recognition. Mobile instance retrieval requires a significant reduction in the
 visual index size. To achieve this, we describe a set of strategies that can re
duce the visual index up to 60-80 x compared to a standard instance retrieval im
plementation found on desktops or servers. While our proposed reduction steps af
fect the overall mean Average Precision (mAP), they are able to maintain a good
Precision for the top K results (P K). We argue that for such offline applicati
on, maintaining a good P K is sufficient. The effectiveness of this approach is
demonstrated on several standard databases. A working application designed for a
 remote historical site is also presented. This application is able to reduce an
 50,000 image index structure to 25 MBs while providing a precision of 97% for P
 10 and 100% for P 1 .

****
Contextual Hypergraph Modeling for Salient Object Detection
Xi Li, Yao Li, Chunhua Shen, Anthony Dick, Anton Van Den Hengel; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3328-3335
Salient object detection aims to locate objects that capture human attention wit hin images. Previous approaches often pose this as a problem of image contrast a nalysis. In this work, we model an image as a hypergraph that utilizes a set of hyperedges to capture the contextual properties of image pixels or regions. As a result, the problem of salient object detection becomes one of finding salient vertices and hyperedges in the hypergraph. The main advantage of hypergraph mode ling is that it takes into account each pixel's (or region's) affinity with its neighborhood as well as its separation from image background. Furthermore, we pr opose an alternative approach based on centerversus-surround contextual contrast analysis, which performs salient object detection by optimizing a cost-sensitiv e support vector machine (SVM) objective function. Experimental results on four challenging datasets demonstrate the effectiveness of the proposed approaches ag ainst the stateof-the-art approaches to salient object detection.
****
Automatic Kronecker Product Model Based Detection of Repeated Patterns in 2D Urb an Images
Juan Liu, Emmanouil Psarakis, Ioannis Stamos; Proceedings of the IEEE Internatio nal Conference on Computer Vision (ICCV), 2013, pp. 401-408
Repeated patterns (such as windows, tiles, balconies and doors) are prominent an d significant features in urban scenes. Therefore, detection of these repeated p atterns becomes very important for city scene analysis. This paper attacks the p roblem of repeated patterns detection in a precise, efficient and automatic way, by combining traditional feature extraction followed by a Kronecker product low rank modeling approach. Our method is tailored for 2D images of building fac,ade s. We have developed algorithms for automatic selection of a representative text ure within fac,ade images using vanishing points and Harris corners. After recti fying the input images, we describe novel algorithms that extract repeated patte rns by using Kronecker product based modeling that is based on a solid theoretic al foundation. Our approach is unique and has not ever been used for fac,ade ana lysis. We have tested our algorithms in a large set of images.
****
From Where and How to What We See
S. Karthikeyan, Vignesh Jagadeesh, Renuka Shenoy, Miguel Ecksteinz, B.S. Manjuna th; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 625-632
Eye movement studies have confirmed that overt attention is highly biased toward s faces and text regions in images. In this paper we explore a novel problem of predicting face and text regions in images using eye tracking data from multiple subjects. The problem is challenging as we aim to predict the semantics (face/t ext/background) only from eye tracking data without utilizing any image informat ion. The proposed algorithm spatially clusters eye tracking data obtained in an image into different coherent groups and subsequently models the likelihood of t he clusters containing faces and text using a fully connected Markov Random Fiel d (MRF). Given the eye tracking data from a test image, it predicts potential fa ce/head (humans, dogs and cats) and text locations reliably. Furthermore, the ap proach can be used to select regions of interest for further analysis by object detectors for faces and text. The hybrid eye position/object detector approach a chieves better detection performance and reduced computation time compared to us ing only the object detection algorithm. We also present a new eye tracking data set on 300 images selected from ICDAR, Street-view, Flickr and Oxford-IIIT Pet D ataset from 15 subjects.
****
Saliency Detection via Absorbing Markov Chain
Bowen Jiang, Lihe Zhang, Huchuan Lu, Chuan Yang, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1665-167 2

In this paper, we formulate saliency detection via absorbing Markov chain on an image graph model. We jointly consider the appearance divergence and spatial distribution of salient objects and the background. The virtual boundary nodes are chosen as the absorbing nodes in a Markov chain and the absorbed time from each transient node to boundary absorbing nodes is computed. The absorbed time of transient node measures its global similarity with all absorbing nodes, and thus salient objects can be consistently separated from the background when the absorbed time is used as a metric. Since the time from transient node to absorbing nodes relies on the weights on the path and their spatial distance, the background region on the center of image may be salient. We further exploit the equilibrium distribution in an ergodic Markov chain to reduce the absorbed time in the long-range smooth background regions. Extensive experiments on four benchmark datasets demonstrate robustness and efficiency of the proposed method against the state-of-the-art methods.

****

Semantic-Aware Co-indexing for Image Retrieval

Shiliang Zhang, Ming Yang, Xiaoyu Wang, Yuanqing Lin, Qi Tian; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1673-1680

Inverted indexes in image retrieval not only allow fast access to database images but also summarize all knowledge about the database, so that their discriminative capacity largely determines the retrieval performance. In this paper, for vocabulary tree based image retrieval, we propose a semantic-aware co-indexing algorithm to jointly embed two strong cues into the inverted indexes: 1) local invariant features that are robust to delineate low-level image contents, and 2) semantic attributes from large-scale object recognition that may reveal image semantic meanings. For an initial set of inverted indexes of local features, we utilize 1000 semantic attributes to filter out isolated images and insert semantically similar images to the initial set. Encoding these two distinct cues together effectively enhances the discriminative capability of inverted indexes. Such co-indexing operations are totally off-line and introduce small computation overhead to online query cause only local features but no semantic attributes are used for query. Experiments and comparisons with recent retrieval methods on 3 datasets, i.e., UKbench, Holidays, Oxford5K, and 1.3 million images from Flickr as distractors, manifest the competitive performance of our method 1 .

****

Stable Hyper-pooling and Query Expansion for Event Detection

Matthijs Douze, Jerome Revaud, Cordelia Schmid, Herve Jegou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1825-1832

This paper makes two complementary contributions to event retrieval in large collections of videos. First, we propose hyper-pooling strategies that encode the frame descriptors into a representation of the video sequence in a stable manner. Our best choices compare favorably with regular pooling techniques based on k-means quantization. Second, we introduce a technique to improve the ranking. It can be interpreted either as a query expansion method or as a similarity adaptation based on the local context of the query video descriptor. Experiments on public benchmarks show that our methods are complementary and improve event retrieval results, without sacrificing efficiency.

****

Predicting Sufficient Annotation Strength for Interactive Foreground Segmentation

Suyog Dutt Jain, Kristen Grauman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1313-1320

The mode of manual annotation used in an interactive segmentation algorithm affects both its accuracy and easeof-use. For example, bounding boxes are fast to supply, yet may be too coarse to get good results on difficult images; freehand outlines are slower to supply and more specific, yet they may be overkill for simple images. Whereas existing methods assume a fixed form of input no matter the image, we propose to predict the tradeoff between accuracy and effort. Our approach learns whether a graph cuts segmentation will succeed if initialized with a given annotation mode, based on the image's visual separability and foreground un

certainty. Using these predictions, we optimize the mode of input requested on n ew images a user wants segmented. Whether given a single image that should be se gmented as quickly as possible, or a batch of images that must be segmented with in a specified time budget, we show how to select the easiest modality that will be sufficiently strong to yield high quality segmentations. Extensive results w ith real users and three datasets demonstrate the impact.
****

What is the Most Efficient Way to Select Nearest Neighbor Candidates for Fast Ap proximate Nearest Neighbor Search?
Masakazu Iwamura, Tomokazu Sato, Koichi Kise; Proceedings of the IEEE Internatio nal Conference on Computer Vision (ICCV), 2013, pp. 3535-3542
Approximate nearest neighbor search (ANNS) is a basic and important technique us ed in many tasks such as object recognition. It involves two processes: selectin g nearest neighbor candidates and performing a brute-force search of these candi dates. Only the former though has scope for improvement. In most existing method s, it approximates the space by quantization. It then calculates all the distanc es between the query and all the quantized values (e.g., clusters or bit sequenc es), and selects a fixed number of candidates close to the query. The performanc e of the method is evaluated based on accuracy as a function of the number of ca ndidates. This evaluation seems rational but poses a serious problem; it ignores the computational cost of the process of selection. In this paper, we propose a new ANNS method that takes into account costs in the selection process. Whereas existing methods employ computationally expensive techniques such as comparativ e sort and heap, the proposed method does not. This realizes a significantly mor e efficient search. We have succeeded in reducing computation times by one-third compared with the state-of-theart on an experiment using 100 million SIFT featu res.
****

Fast High Dimensional Vector Multiplication Face Recognition
Oren Barkan, Jonathan Weill, Lior Wolf, Hagai Aronowitz; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1960-1967
This paper advances descriptor-based face recognition by suggesting a novel usag e of descriptors to form an over-complete representation, and by proposing a new metric learning pipeline within the same/not-same framework. First, the Over-Co mplete Local Binary Patterns (OCLBP) face representation scheme is introduced as a multi-scale modified version of the Local Binary Patterns (LBP) scheme. Secon d, we propose an efficient matrix-vector multiplication-based recognition system . The system is based on Linear Discriminant Analysis (LDA) coupled with Within Class Covariance Normalization (WCCN). This is further extended to the unsupervi sed case by proposing an unsupervised variant of WCCN. Lastly, we introduce Diff usion Maps (DM) for non-linear dimensionality reduction as an alternative to the Whitened Principal Component Analysis (WPCA) method which is often used in face recognition. We evaluate the proposed framework on the LFW face recognition dat aset under the restricted, unrestricted and unsupervised protocols. In all three cases we achieve very competitive results.
****

Group Norm for Learning Structured SVMs with Unstructured Latent Variables
Daozheng Chen, Dhruv Batra, William T. Freeman; Proceedings of the IEEE Internat ional Conference on Computer Vision (ICCV), 2013, pp. 409-416
Latent variables models have been applied to a number of computer vision problem s. However, the complexity of the latent space is typically left as a free desig n choice. A larger latent space results in a more expressive model, but such mod els are prone to overfitting and are slower to perform inference with. The goal of this paper is to regularize the complexity of the latent space and learn whic h hidden states are really relevant for prediction. Specifically, we propose usi ng group-sparsity-inducing regularizers such as 1 2 to estimate the parameters o f Structured SVMs with unstructured latent variables. Our experiments on digit r ecognition and object detection show that our approach is indeed able to control the complexity of latent space without any significant loss in accuracy of the learnt model.

****
New Graph Structured Sparsity Model for Multi-label Image Annotations
Xiao Cai, Feiping Nie, Weidong Cai, Heng Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 801-808
In multi-label image annotations, because each image is associated to multiple categories, the semantic terms (label classes) are not mutually exclusive. Previous research showed that such label correlations can largely boost the annotation accuracy. However, all existing methods only directly apply the label correlation matrix to enhance the label inference and assignment without further learning the structural information among classes. In this paper, we model the label correlations using the relational graph, and propose a novel graph structured sparse learning model to incorporate the topological constraints of relation graph in multi-label classifications. As a result, our new method will capture and utilize the hidden class structures in relational graph to improve the annotation results. In proposed objective, a large number of structured sparsity-inducing norms are utilized, thus the optimization becomes difficult. To solve this problem, we derive an efficient optimization algorithm with proved convergence. We perform extensive experiments on six multi-label image annotation benchmark data sets. In all empirical results, our new method shows better annotation results than the state-of-the-art approaches.
****
Real-Time Body Tracking with One Depth Camera and Inertial Sensors
Thomas Helten, Meinard Muller, Hans-Peter Seidel, Christian Theobalt; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1105-1112
In recent years, the availability of inexpensive depth cameras, such as the Microsoft Kinect, has boosted the research in monocular full body skeletal pose tracking. Unfortunately, existing trackers often fail to capture poses where a single camera provides insufficient data, such as non-frontal poses, and all other poses with body part occlusions. In this paper, we present a novel sensor fusion approach for real-time full body tracking that succeeds in such difficult situations. It takes inspiration from previous tracking solutions, and combines a generative tracker and a discriminative tracker retrieving closest poses in a database. In contrast to previous work, both trackers employ data from a low number of inexpensive body-worn inertial sensors. These sensors provide reliable and complementary information when the monocular depth information alone is not sufficient. We also contribute by new algorithmic solutions to best fuse depth and inertial data in both trackers. One is a new visibility model to determine global body pose, occlusions and usable depth correspondences and to decide what data modality to use for discriminative tracking. We also contribute with a new inertial-based pose retrieval, and an adapted late fusion step to calculate the final body pose.
****
Image Segmentation with Cascaded Hierarchical Models and Logistic Disjunctive Normal Networks
Mojtaba Seyedhosseini, Mehdi Sajjadi, Tolga Tasdizen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2168-2175
Contextual information plays an important role in solving vision problems such as image segmentation. However, extracting contextual information and using it in an effective way remains a difficult problem. To address this challenge, we propose a multi-resolution contextual framework, called cascaded hierarchical model (CHM), which learns contextual information in a hierarchical framework for image segmentation. At each level of the hierarchy, a classifier is trained based on downsampled input images and outputs of previous levels. Our model then incorporates the resulting multi-resolution contextual information into a classifier to segment the input image at original resolution. We repeat this procedure by cascading the hierarchical framework to improve the segmentation accuracy. Multiple classifiers are learned in the CHM; therefore, a fast and accurate classifier is required to make the training tractable. The classifier also needs to be robust against overfitting due to the large number of parameters learned during train

ing. We introduce a novel classification scheme, called logistic disjunctive normal networks (LDNN), which consists of one adaptive layer of feature detectors implemented by logistic sigmoid functions followed by two fixed layers of logical units that compute conjunctions and disjunctions, respectively. We demonstrate that LDNN outperforms state-of-theart classifiers and can be used in the CHM to improve object segmentation performance.
****

Low-Rank Sparse Coding for Image Classification
Tianzhu Zhang, Bernard Ghanem, Si Liu, Changsheng Xu, Narendra Ahuja; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 281-288
In this paper, we propose a low-rank sparse coding (LRSC) method that exploits local structure information among features in an image for the purpose of image-level classification. LRSC represents densely sampled SIFT descriptors, in a spatial neighborhood, collectively as lowrank, sparse linear combinations of codewords. As such, it casts the feature coding problem as a low-rank matrix learning problem, which is different from previous methods that encode features independently. This LRSC has a number of attractive properties. (1) It encourages sparsity in feature codes, locality in codebook construction, and low-rankness for spatial consistency. (2) LRSC encodes local features jointly by considering their low-rank structure information, and is computationally attractive. We evaluate the LRSC by comparing its performance on a set of challenging benchmarks with that of orpopular coding and other state-of-theart methods. Our experiments show that by representing local features jointly, LRSC not only outperforms the state-ofthe-art in classification accuracy but also improves the time complexity of methods that use a similar sparse linear representation model for feature coding [36].
****

Learning to Rank Using Privileged Information
Viktoriia Sharmanska, Novi Quadrianto, Christoph H. Lampert; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 825-832
Many computer vision problems have an asymmetric distribution of information between training and test time. In this work, we study the case where we are given additional information about the training data, which however will not be available at test time. This situation is called learning using privileged information (LUPI). We introduce two maximum-margin techniques that are able to make use of this additional source of information, and we show that the framework is applicable to several scenarios that have been studied in computer vision before. Experiments with attributes, bounding boxes, image tags and rationales as additional information in object classification show promising results.
****

Extrinsic Camera Calibration without a Direct View Using Spherical Mirror
Amit Agrawal; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2368-2375
We consider the problem of estimating the extrinsic parameters (pose) of a camera with respect to a reference 3D object without a direct view. Since the camera does not view the object directly, previous approaches have utilized reflections in a planar mirror to solve this problem. However, a planar mirror based approach requires a minimum of three reflections and has degenerate configurations where estimation fails. In this paper, we show that the pose can be obtained using a single reflection in a spherical mirror of known radius. This makes our approach simpler and easier in practice. In addition, unlike planar mirrors, the spherical mirror based approach does not have any degenerate configurations, leading to a robust algorithm. While a planar mirror reflection results in a virtual perspective camera, a spherical mirror reflection results in a non-perspective axial camera. The axial nature of rays allows us to compute the axis (direction of sphere center) and few pose parameters in a linear fashion. We then derive an analytical solution to obtain the distance to the sphere center and remaining pose parameters and show that it corresponds to solving a 16 eppdegree equation. We present comparisons with a recent method that use planar mirrors and show that our approach recovers more accurate pose in the presence of noise. Extensive simul

ations and results on real data validate our algorithm.
****

Two-Point Gait: Decoupling Gait from Body Shape
Stephen Lombardi, Ko Nishino, Yasushi Makihara, Yasushi Yagi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1041-1048
Human gait modeling (e.g., for person identification) largely relies on image-based representations that muddle gait with body shape. Silhouettes, for instance, inherently entangle body shape and gait. For gait analysis and recognition, decoupling these two factors is desirable. Most important, once decoupled, they can be combined for the task at hand, but not if left entangled in the first place. In this paper, we introduce Two-Point Gait, a gait representation that encodes the limb motions regardless of the body shape. Two-Point Gait is directly computed on the image sequence based on the two point statistics of optical flow fields. We demonstrate its use for exploring the space of human gait and gait recognition under large clothing variation. The results show that we can achieve state-of-the-art person recognition accuracy on a challenging dataset.
****

Robust Subspace Clustering via Half-Quadratic Minimization
Yingya Zhang, Zhenan Sun, Ran He, Tieniu Tan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3096-3103
Subspace clustering has important and wide applications in computer vision and pattern recognition. It is a challenging task to learn low-dimensional subspace structures due to the possible errors (e.g., noise and corruptions) existing in high-dimensional data. Recent subspace clustering methods usually assume a sparse representation of corrupted errors and correct the errors iteratively. However large corruptions in real-world applications can not be well addressed by these methods. A novel optimization model for robust subspace clustering is proposed in this paper. The objective function of our model mainly includes two parts. The first part aims to achieve a sparse representation of each high-dimensional data point with other data points. The second part aims to maximize the correntropy between a given data point and its low-dimensional representation with other points. Correntropy is a robust measure so that the influence of large corruptions on subspace clustering can be greatly suppressed. An extension of our
****

Category-Independent Object-Level Saliency Detection
Yangqing Jia, Mei Han; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1761-1768
It is known that purely low-level saliency cues such as frequency does not lead to a good salient object detection result, requiring high-level knowledge to be adopted for successful discovery of task-independent salient objects. In this paper, we propose an efficient way to combine such high-level saliency priors and low-level appearance models. We obtain the high-level saliency prior with the objectness algorithm to find potential object candidates without the need of category information, and then enforce the consistency among the salient regions using a Gaussian MRF with the weights scaled by diverse density that emphasizes the influence of potential foreground pixels. Our model obtains saliency maps that assign high scores for the whole salient object, and achieves state-of-the-art performance on benchmark datasets covering various foreground statistics.
****

A Method of Perceptual-Based Shape Decomposition
Chang Ma, Zhongqian Dong, Tingting Jiang, Yizhou Wang, Wen Gao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 873-880
In this paper, we propose a novel perception-based shape decomposition method which aims to decompose a shape into semantically meaningful parts. In addition to three popular perception rules (the Minima rule, the Short-cut rule and the Convexity rule) in shape decomposition, we propose a new rule named part-similarity rule to encourage consistent partition of similar parts. The problem is formulated as a quadratically constrained quadratic program (QCQP) problem and is solved by a trust-region method. Experiment results on MPEG-7 dataset show that we can get a more consistent shape decomposition with human perception compared with

other state-of-the-art methods both qualitatively and quantitatively. Finally, w
e show the advantage of semantic parts over non-meaningful parts in object detec
tion on the ETHZ dataset.
****

Bayesian Robust Matrix Factorization for Image and Video Processing
Naiyan Wang, Dit-Yan Yeung; Proceedings of the IEEE International Conference on
Computer Vision (ICCV), 2013, pp. 1785-1792
Matrix factorization is a fundamental problem that is often encountered in many
computer vision and machine learning tasks. In recent years, enhancing the robus
tness of matrix factorization methods has attracted much attention in the resear
ch community. To benefit from the strengths of full Bayesian treatment over poin
t estimation, we propose here a full Bayesian approach to robust matrix factoriz
ation. For the generative process, the model parameters have conjugate priors an
d the likelihood (or noise model) takes the form of a Laplace mixture. For Bayes
ian inference, we devise an efficient sampling algorithm by exploiting a hierarc
hical view of the Laplace distribution. Besides the basic model, we also propose
 an extension which assumes that the outliers exhibit spatial or temporal proxim
ity as encountered in many computer vision applications. The proposed methods gi
ve competitive experimental results when compared with several state-of-the-art
methods on some benchmark image and video processing tasks.
****

Measuring Flow Complexity in Videos
Saad Ali; Proceedings of the IEEE International Conference on Computer Vision (I
CCV), 2013, pp. 1097-1104
In this paper a notion of flow complexity that measures the amount of interactio
n among objects is introduced and an approach to compute it directly from a vide
o sequence is proposed. The approach employs particle trajectories as the input
representation of motion and maps it into a 'braid' based representation. The ma
pping is based on the observation that 2D trajectories of particles take the for
m of a braid in space-time due to the intermingling among particles over time. A
s a result of this mapping, the problem of estimating the flow complexity from p
article trajectories becomes the problem of estimating braid complexity, which i
n turn can be computed by measuring the topological entropy of a braid. For this
 purpose recently developed mathematical tools from braid theory are employed wh
ich allow rapid computation of topological entropy of braids. The approach is ev
aluated on a dataset consisting of open source videos depicting variations in te
rms of types of moving objects, scene layout, camera view angle, motion patterns
, and object densities. The results show that the proposed approach is able to q
uantify the complexity of the flow, and at the same time provides useful insight
s about the sources of the complexity.
****

DeepFlow: Large Displacement Optical Flow with Deep Matching
Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, Cordelia Schmid; Proceeding
s of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1385
-1392
Optical flow computation is a key component in many computer vision systems desi
gned for tasks such as action detection or activity recognition. However, despit
e several major advances over the last decade, handling large displacement in op
tical flow remains an open problem. Inspired by the large displacement optical f
low of Brox & Malik [6], our approach, termed DeepFlow, blends a matching algori
thm with a variational approach for optical flow. We propose a descriptor matchi
ng algorithm, tailored to the optical flow problem, that allows to boost perform
ance on fast motions. The matching algorithm builds upon a multi-stage architect
ure with 6 layers, interleaving convolutions and max-pooling, a construction aki
n to deep convolutional nets. Using dense sampling, it allows to efficiently ret
rieve quasi-dense correspondences, and enjoys a built-in smoothing effect on des
criptors matches, a valuable asset for integration into an energy minimization f
ramework for optical flow estimation. DeepFlow efficiently handles large displac
ements occurring in realistic videos, and shows competitive performance on optic
al flow benchmarks. Furthermore, it sets a new state-of-the-art on the MPI-Sinte

l dataset [8].
****
The Way They Move: Tracking Multiple Targets with Similar Appearance
Caglayan Dicle, Octavia I. Camps, Mario Sznaier; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2304-2311
We introduce a computationally efficient algorithm for multi-object tracking by detection that addresses four main challenges: appearance similarity among targets, missing data due to targets being out of the field of view or occluded behind other objects, crossing trajectories, and camera motion. The proposed method uses motion dynamics as a cue to distinguish targets with similar appearance, minimize target mis-identification and recover missing data. Computational efficiency is achieved by using a Generalized Linear Assignment (GLA) coupled with efficient procedures to recover missing data and estimate the complexity of the underlying dynamics. The proposed approach works with tracklets of arbitrary length and does not assume a dynamical model a priori, yet it captures the overall motion dynamics of the targets. Experiments using challenging videos show that this framework can handle complex target motions, non-stationary cameras and long occlusions, on scenarios where appearance cues are not available or poor.
****
What Do You Do? Occupation Recognition in a Photo via Social Context
Ming Shao, Liangyue Li, Yun Fu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3631-3638
In this paper, we investigate the problem of recognizing occupations of multiple people with arbitrary poses in a photo. Previous work utilizing single person's nearly frontal clothing information and fore/background context preliminarily proves that occupation recognition is computationally feasible in computer vision. However, in practice, multiple people with arbitrary poses are common in a photo, and recognizing their occupations is even more challenging. We argue that with appropriately built visual attributes, co-occurrence, and spatial configuration model that is learned through structure SVM, we can recognize multiple people's occupations in a photo simultaneously. To evaluate our method's performance, we conduct extensive experiments on a new well-labeled occupation database with 14 representative occupations and over 7K images. Results on this database validate our method's effectiveness and show that occupation recognition is solvable in a more general case.
****
Pose Estimation and Segmentation of People in 3D Movies
Karteek Alahari, Guillaume Seguin, Josef Sivic, Ivan Laptev; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2112-2119
We seek to obtain a pixel-wise segmentation and pose estimation of multiple people in a stereoscopic video. This involves challenges such as dealing with unconstrained stereoscopic video, non-stationary cameras, and complex indoor and outdoor dynamic scenes. The contributions of our work are two-fold: First, we develop a segmentation model incorporating person detection, pose estimation, as well as colour, motion, and disparity cues. Our new model explicitly represents depth ordering and occlusion. Second, we introduce a stereoscopic dataset with frames extracted from feature-length movies "StreetDance 3D" and "Pina". The dataset contains 2727 realistic stereo pairs and includes annotation of human poses, person bounding boxes, and pixel-wise segmentations for hundreds of people. The dataset is composed of indoor and outdoor scenes depicting multiple people with frequent occlusions. We demonstrate results on our new challenging dataset, as well as on the H2view dataset from (Sheasby et al. ACCV 2012).
****
Calibration-Free Gaze Estimation Using Human Gaze Patterns
Fares Alnajar, Theo Gevers, Roberto Valenti, Sennay Ghebreab; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 137-144
We present a novel method to auto-calibrate gaze estimators based on gaze patterns obtained from other viewers. Our method is based on the observation that the gaze patterns of humans are indicative of where a new viewer will look at [12]. When a new viewer is looking at a stimulus, we first estimate a topology of gaze

points (initial gaze points). Next, these points are transformed so that they match the gaze patterns of other humans to find the correct gaze points. In a flexible uncalibrated setup with a web camera and no chin rest, the proposed method was tested on ten subjects and ten images. The method estimates the gaze points after looking at a stimulus for a few seconds with an average accuracy of 4.3 im. Although the reported performance is lower than what could be achieved with dedicated hardware or calibrated setup, the proposed method still provides a sufficient accuracy to trace the viewer attention. This is promising considering the fact that auto-calibration is done in a flexible setup , without the use of a chin rest, and based only on a few seconds of gaze initialization data. To the best of our knowledge, this is the first work to use human gaze patterns in order to auto-calibrate gaze estimators.

****

Lifting 3D Manhattan Lines from a Single Image

Srikumar Ramalingam, Matthew Brand; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 497-504

We propose a novel and an efficient method for reconstructing the 3D arrangement of lines extracted from a single image, using vanishing points, orthogonal structure, and an optimization procedure that considers all plausible connectivity constraints between lines. Line detection identifies a large number of salient lines that intersect or nearly intersect in an image, but relatively a few of these apparent junctions correspond to real intersections in the 3D scene. We use linear programming (LP) to identify a minimal set of least-violated connectivity constraints that are sufficient to unambiguously reconstruct the 3D lines. In contrast to prior solutions that primarily focused on well-behaved synthetic line drawings with severely restricting assumptions, we develop an algorithm that can work on real images. The algorithm produces line reconstruction by identifying 95% correct connectivity constraints in York Urban database, with a total computation time of 1 second per image.

****

A Framework for Shape Analysis via Hilbert Space Embedding

Sadeep Jayasumana, Mathieu Salzmann, Hongdong Li, Mehrtash Harandi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1249-1256

We propose a framework for 2D shape analysis using positive definite kernels defined on Kendall's shape manifold. Different representations of 2D shapes are known to generate different nonlinear spaces. Due to the nonlinearity of these spaces, most existing shape classification algorithms resort to nearest neighbor methods and to learning distances on shape spaces. Here, we propose to map shapes on Kendall's shape manifold to a high dimensional Hilbert space where Euclidean geometry applies. To this end, we introduce a kernel on this manifold that permits such a mapping, and prove its positive definiteness. This kernel lets us extend kernel-based algorithms developed for Euclidean spaces, such as SVM, MKL and kernel PCA, to the shape manifold. We demonstrate the benefits of our approach over the state-of-the-art methods on shape classification, clustering and retrieval.

****

Structured Forests for Fast Edge Detection

Piotr Dollar, C. L. Zitnick; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1841-1848

Edge detection is a critical component of many vision systems, including object detectors and image segmentation algorithms. Patches of edges exhibit well-known forms of local structure, such as straight lines or T-junctions. In this paper we take advantage of the structure present in local image patches to learn both an accurate and computationally efficient edge detector. We formulate the problem of predicting local edge masks in a structured learning framework applied to random decision forests. Our novel approach to learning decision trees robustly maps the structured labels to a discrete space on which standard information gain measures may be evaluated. The result is an approach that obtains realtime performance that is orders of magnitude faster than many competing state-of-the-art

approaches, while also achieving state-of-the-art edge detection results on the BSDS500 Segmentation dataset and NYU Depth dataset. Finally, we show the potential of our approach as a general purpose edge detector by showing our learned edge models generalize well across datasets.
****

Scene Text Localization and Recognition with Oriented Stroke Detection
Lukas Neumann, Jiri Matas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 97-104
An unconstrained end-to-end text localization and recognition method is presented. The method introduces a novel approach for character detection and recognition which combines the advantages of sliding-window and connected component methods. Characters are detected and recognized as image regions which contain strokes of specific orientations in a specific relative position, where the strokes are efficiently detected by convolving the image gradient field with a set of oriented bar filters. Additionally, a novel character representation efficiently calculated from the values obtained in the stroke detection phase is introduced. The representation is robust to shift at the stroke level, which makes it less sensitive to intra-class variations and the noise induced by normalizing character size and positioning. The effectiveness of the representation is demonstrated by the results achieved in the classification of real-world characters using an euclidian nearestneighbor classifier trained on synthetic data in a plain form. The method was evaluated on a standard dataset, where it achieves state-of-the-art results in both text localization and recognition.
****

CoDeL: A Human Co-detection and Labeling Framework
Jianping Shi, Renjie Liao, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2096-2103
We propose a co-detection and labeling (CoDeL) framework to identify persons that contain self-consistent appearance in multiple images. Our CoDeL model builds upon the deformable part-based model to detect human hypotheses and exploits cross-image correspondence via a matching classifier. Relying on a Gaussian process, this matching classifier models the similarity of two hypotheses and efficiently captures the relative importance contributed by various visual features, reducing the adverse effect of scattered occlusion. Further, the detector and matching classifier together make our model fit into a semi-supervised co-training framework, which can get enhanced results with a small amount of labeled training data. Our CoDeL model achieves decent performance on existing and new benchmark datasets.
****

Exploiting Reflection Change for Automatic Reflection Removal
Yu Li, Michael S. Brown; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2432-2439
This paper introduces an automatic method for removing reflection interference when imaging a scene behind a glass surface. Our approach exploits the subtle changes in the reflection with respect to the background in a small set of images taken at slightly different view points. Key to this idea is the use of SIFT-flow to align the images such that a pixel-wise comparison can be made across the input set. Gradients with variation across the image set are assumed to belong to the reflected scenes while constant gradients are assumed to belong to the desired background scene. By correctly labelling gradients belonging to reflection or background, the background scene can be separated from the reflection interference. Unlike previous approaches that exploit motion, our approach does not make any assumptions regarding the background or reflected scenes' geometry, nor requires the reflection to be static. This makes our approach practical for use in casual imaging scenarios. Our approach is straight forward and produces good results compared with existing methods.
****

Elastic Net Constraints for Shape Matching
Emanuele Rodola, Andrea Torsello, Tatsuya Harada, Yasuo Kuniyoshi, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2

013, pp. 1169-1176
We consider a parametrized relaxation of the widely adopted quadratic assignment problem (QAP) formulation for minimum distortion correspondence between deformable shapes. In order to control the accuracy/sparsity trade-off we introduce a weighting parameter on the combination of two existing relaxations, namely spectral and
****

A New Adaptive Segmental Matching Measure for Human Activity Recognition
Shahriar Shariat, Vladimir Pavlovic; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3583-3590
The problem of human activity recognition is a central problem in many real-world applications. In this paper we propose a fast and effective segmental alignmentbased method that is able to classify activities and interactions in complex environments. We empirically show that such model is able to recover the alignment that leads to improved similarity measures within sequence classes and hence, raises the classification performance. We also apply a bounding technique on the histogram distances to reduce the computation of the otherwise exhaustive search.
****

A Generalized Iterated Shrinkage Algorithm for Non-convex Sparse Coding
Wangmeng Zuo, Deyu Meng, Lei Zhang, Xiangchu Feng, David Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 217-224
In many sparse coding based image restoration and image classification problems, using non-convex p -norm minimization (0 top po1) can often obtain better results than the convex 1 -norm minimization. A number of algorithms, e.g., iteratively reweighted least squares (IRLS), iteratively thresholding method (ITMp), and look-up table (LUT), have been proposed for non-convex p -norm sparse coding, while some analytic solutions have been suggested for some specific values of p. In this paper, by extending the popular soft-thresholding operator, we propose a generalized iterated shrinkage algorithm (GISA) for p -norm non-convex sparse coding. Unlike the analytic solutions, the proposed GISA algorithm is easy to implement, and can be adopted for solving non-convex sparse coding problems with arbitrary p values. Compared with LUT, GISA is more general and does not need to compute and store the look-up tables. Compared with IRLS and ITMp , GISA is theoretically more solid and can achieve more accurate solutions. Experiments on image restoration and sparse coding based face recognition are conducted to validate the performance of GISA.
****

Robust Tucker Tensor Decomposition for Effective Image Representation
Miao Zhang, Chris Ding; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2448-2455
Many tensor based algorithms have been proposed for the study of high dimensional data in a large variety of computer vision and machine learning applications. However, most of the existing tensor analysis approaches are based on Frobenius norm, which makes them sensitive to outliers, because they minimize the sum of squared errors and enlarge the influence of both outliers and large feature noises. In this paper, we propose a robust Tucker tensor decomposition model (RTD) to suppress the influence of outliers, which uses L 1 -norm loss function. Yet, the optimization on L 1 -norm based tensor analysis is much harder than standard tensor decomposition. In this paper, we propose a simple and efficient algorithm to solve our RTD model. Moreover, tensor factorization-based image storage needs much less space than PCA based methods. We carry out extensive experiments to evaluate the proposed algorithm, and verify the robustness against image occlusions. Both numerical and visual results show that our RTD model is consistently better against the existence of outliers than previous tensor and PCA methods.
****

Learning Graphs to Match
Minsu Cho, Karteek Alahari, Jean Ponce; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 25-32
Many tasks in computer vision are formulated as graph matching problems. Despite

the NP-hard nature of the problem, fast and accurate approximations have led to significant progress in a wide range of applications. Learning graph models from observed data, however, still remains a challenging issue. This paper presents an effective scheme to parameterize a graph model, and learn its structural attributes for visual object matching. For this, we propose a graph representation with histogram-based attributes, and optimize them to increase the matching accuracy. Experimental evaluations on synthetic and real image datasets demonstrate the effectiveness of our approach, and show significant improvement in matching accuracy over graphs with pre-defined structures.
****

SGTD: Structure Gradient and Texture Decorrelating Regularization for Image Decomposition
Qiegen Liu, Jianbo Liu, Pei Dong, Dong Liang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1081-1088
This paper presents a novel structure gradient and texture decorrelating regularization (SGTD) for image decomposition. The motivation of the idea is under the assumption that the structure gradient and texture components should be properly decorrelated for a successful decomposition. The proposed model consists of the data fidelity term, total variation regularization and the SGTD regularization. An augmented Lagrangian method is proposed to address this optimization issue, by first transforming the unconstrained problem to an equivalent constrained problem and then applying an alternating direction method to iteratively solve the subproblems. Experimental results demonstrate that the proposed method presents better or comparable performance as state-of-the-art methods do.
****

Discovering Details and Scene Structure with Hierarchical Iconoid Shift
Tobias Weyand, Bastian Leibe; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3479-3486
Current landmark recognition engines are typically aimed at recognizing building-scale landmarks, but miss interesting details like portals, statues or windows. This is because they use a flat clustering that summarizes all photos of a building facade in one cluster. We propose Hierarchical Iconoid Shift, a novel landmark clustering algorithm capable of discovering such details. Instead of just a collection of clusters, the output of HIS is a set of dendrograms describing the detail hierarchy of a landmark. HIS is based on the novel Hierarchical Medoid Shift clustering algorithm that performs a continuous mode search over the complete scale space. HMS is completely parameter-free, has the same complexity as Medoid Shift and is easy to parallelize. We evaluate HIS on 800k images of 34 landmarks and show that it can extract an often surprising amount of detail and structure that can be applied, e.g., to provide a mobile user with more detailed information on a landmark or even to extend the landmark's Wikipedia article.
****

Single-Patch Low-Rank Prior for Non-pointwise Impulse Noise Removal
Ruixuan Wang, Emanuele Trucco; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1073-1080
This paper introduces a 'low-rank prior' for small oriented noise-free image patches: considering an oriented patch as a matrix, a low-rank matrix approximation is enough to preserve the texture details in the properly oriented patch. Based on this prior, we propose a single-patch method within a generalized joint low-rank and sparse matrix recovery framework to simultaneously detect and remove non-pointwise random-valued impulse noise (e.g., very small blobs). A weighting matrix is incorporated in the framework to encode an initial estimate of the spatial noise distribution. An accelerated proximal gradient method is adapted to estimate the optimal noise-free image patches. Experiments show the effectiveness of our framework in removing non-pointwise random-valued impulse noise.
****

Separating Reflective and Fluorescent Components Using High Frequency Illumination in the Spectral Domain
Ying Fu, Antony Lam, Imari Sato, Takahiro Okabe, Yoichi Sato; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 457-464

Hyperspectral imaging is beneficial to many applications but current methods do not consider fluorescent effects which are present in everyday items ranging from paper, to clothing, to even our food. Furthermore, everyday fluorescent items exhibit a mix of reflectance and fluorescence. So proper separation of these components is necessary for analyzing them. In this paper, we demonstrate efficient separation and recovery of reflective and fluorescent emission spectra through the use of high frequency illumination in the spectral domain. With the obtained fluorescent emission spectra from our high frequency illuminants, we then present to our knowledge, the first method for estimating the fluorescent absorption spectrum of a material given its emission spectrum. Conventional bispectral measurement of absorption and emission spectra needs to examine all combinations of incident and observed light wavelengths. In contrast, our method requires only two hyperspectral images. The effectiveness of our proposed methods are then evaluated through a combination of simulation and real experiments. We also demonstrate an application of our method to synthetic relighting of real scenes.
****

Learning to Predict Gaze in Egocentric Video
Yin Li, Alireza Fathi, James M. Rehg; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3216-3223
We present a model for gaze prediction in egocentric video by leveraging the implicit cues that exist in camera wearer's behaviors. Specifically, we compute the camera wearer's head motion and hand location from the video and combine them to estimate where the eyes look. We further model the dynamic behavior of the gaze, in particular fixations, as latent variables to improve the gaze prediction. Our gaze prediction results outperform the state-of-the-art algorithms by a large margin on publicly available egocentric vision datasets. In addition, we demonstrate that we get a significant performance boost in recognizing daily actions and segmenting foreground objects by plugging in our gaze predictions into state-of-the-art methods.
****

Fine-Grained Categorization by Alignments
E. Gavves, B. Fernando, C.G.M. Snoek, A.W.M. Smeulders, T. Tuytelaars; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1713-1720
The aim of this paper is fine-grained categorization without human interaction. Different from prior work, which relies on detectors for specific object parts, we propose to localize distinctive details by roughly aligning the objects using just the overall shape, since implicit to fine-grained categorization is the existence of a super-class shape shared among all classes. The alignments are then used to transfer part annotations from training images to test images (supervised alignment), or to blindly yet consistently segment the object in a number of regions (unsupervised alignment). We furthermore argue that in the distinction of finegrained sub-categories, classification-oriented encodings like Fisher vectors are better suited for describing localized information than popular matching oriented features like HOG. We evaluate the method on the CU-2011 Birds and Stanford Dogs fine-grained datasets, outperforming the state-of-the-art.
****

Symbiotic Segmentation and Part Localization for Fine-Grained Categorization
Yuning Chai, Victor Lempitsky, Andrew Zisserman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 321-328
We propose a new method for the task of fine-grained visual categorization. The method builds a model of the baselevel category that can be fitted to images, producing highquality foreground segmentation and mid-level part localizations. The model can be learnt from the typical datasets available for fine-grained categorization, where the only annotation provided is a loose bounding box around the instance (e.g. bird) in each image. Both segmentation and part localizations are then used to encode the image content into a highly-discriminative visual signature. The model is symbiotic in that part discovery/localization is helped by segmentation and, conversely, the segmentation is helped by the detection (e.g. part layout). Our model builds on top of the part-based object category detector

of Felzenszwalb et al., and also on the powerful GrabCut segmentation algorithm of Rother et al., and adds a simple spatial saliency coupling between them. In our evaluation, the model improves the categorization accuracy over the state-of-the-art. It also improves over what can be achieved with an analogous system that runs segmentation and part-localization independently.
****

Learning People Detectors for Tracking in Crowded Scenes
Siyu Tang, Mykhaylo Andriluka, Anton Milan, Konrad Schindler, Stefan Roth, Bernt Schiele; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1049-1056
People tracking in crowded real-world scenes is challenging due to frequent and long-term occlusions. Recent tracking methods obtain the image evidence from object (people) detectors, but typically use off-the-shelf detectors and treat them as black box components. In this paper we argue that for best performance one should explicitly train people detectors on failure cases of the overall tracker instead. To that end, we first propose a novel joint people detector that combines a state-of-the-art single person detector with a detector for pairs of people, which explicitly exploits common patterns of person-person occlusions across multiple viewpoints that are a frequent failure case for tracking in crowded scenes. To explicitly address remaining failure modes of the tracker we explore two methods. First, we analyze typical failures of trackers and train a detector explicitly on these cases. And second, we train the detector with the people tracker in the loop, focusing on the most common tracker failures. We show that our joint multi-person detector significantly improves both detection accuracy as well as tracker performance, improving the state-of-the-art on standard benchmarks.
****

SIFTpack: A Compact Representation for Efficient SIFT Matching
Alexandra Gilinsky, Lihi Zelnik Manor; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 777-784
Computing distances between large sets of SIFT descriptors is a basic step in numerous algorithms in computer vision. When the number of descriptors is large, as is often the case, computing these distances can be extremely time consuming. In this paper we propose the SIFTpack: a compact way of storing SIFT descriptors, which enables significantly faster calculations between sets of SIFTs than the current solutions. SIFTpack can be used to represent SIFTs densely extracted from a single image or sparsely from multiple different images. We show that the SIFTpack representation saves both storage space and run time, for both finding nearest neighbors and for computing all distances between all descriptors. The usefulness of SIFTpack is also demonstrated as an alternative implementation for K-means dictionaries of visual words.
****

Forward Motion Deblurring
Shicheng Zheng, Li Xu, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1465-1472
We handle a special type of motion blur considering that cameras move primarily forward or backward. Solving this type of blur is of unique practical importance since nearly all car, traffic and bike-mounted cameras follow out-ofplane translational motion. We start with the study of geometric models and analyze the difficulty of existing methods to deal with them. We also propose a solution accounting for depth variation. Homographies associated with different 3D planes are considered and solved for in an optimization framework. Our method is verified on several natural image examples that cannot be satisfyingly dealt with by previous methods.
****

HOGgles: Visualizing Object Detection Features
Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, Antonio Torralba; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1-8
We introduce algorithms to visualize feature spaces used by object detectors. The tools in this paper allow a human to put on 'HOG goggles' and perceive the visual world as a HOG based object detector sees it. We found that these visualizat

ions allow us to analyze object detection systems in new ways and gain new insig
ht into the detector's failures. For example, when we visualize the features for
 high scoring false alarms, we discovered that, although they are clearly wrong
in image space, they do look deceptively similar to true positives in feature sp
ace. This result suggests that many of these false alarms are caused by our choi
ce of feature space, and indicates that creating a better learning algorithm or
building bigger datasets is unlikely to correct these errors. By visualizing fea
ture spaces, we can gain a more intuitive understanding of our detection systems
.
****

Pictorial Human Spaces: How Well Do Humans Perceive a 3D Articulated Pose?
Elisabeta Marinoiu, Dragos Papava, Cristian Sminchisescu; Proceedings of the IEE
E International Conference on Computer Vision (ICCV), 2013, pp. 1289-1296
Human motion analysis in images and video is a central computer vision problem.
Yet, there are no studies that reveal how humans perceive other people in images
 and how accurate they are. In this paper we aim to unveil some of the processin
g-as well as the levels of accuracy-involved in the 3D perception of people from
 images by assessing the human performance. Our contributions are: (1) the const
ruction of an experimental apparatus that relates perception and measurement, in
 particular the visual and kinematic performance with respect to 3D ground truth
 when the human subject is presented an image of a person in a given pose; (2) t
he creation of a dataset containing images, articulated 2D and 3D pose ground tr
uth, as well as synchronized eye movement recordings of human subjects, shown a
variety of human body configurations, both easy and difficult, as well as their
're-enacted' 3D poses; (3) quantitative analysis revealing the human performance
 in 3D pose reenactment tasks, the degree of stability in the visual fixation pa
tterns of human subjects, and the way it correlates with different poses. We als
o discuss the implications of our findings for the construction of visual human
sensing systems.
****

EVSAC: Accelerating Hypotheses Generation by Modeling Matching Scores with Extre
me Value Theory
Victor Fragoso, Pradeep Sen, Sergio Rodriguez, Matthew Turk; Proceedings of the
IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2472-2479
Algorithms based on RANSAC that estimate models using feature correspondences be
tween images can slow down tremendously when the percentage of correct correspon
dences (inliers) is small. In this paper, we present a probabilistic parametric
model that allows us to assign confidence values for each matching correspondenc
e and therefore accelerates the generation of hypothesis models for RANSAC under
 these conditions. Our framework leverages Extreme Value Theory to accurately mo
del the statistics of matching scores produced by a nearest-neighbor feature mat
cher. Using a new algorithm based on this model, we are able to estimate accurat
e hypotheses with RANSAC at low inlier ratios significantly faster than previous
 stateof-the-art approaches, while still performing comparably when the number o
f inliers is large. We present results of homography and fundamental matrix esti
mation experiments for both SIFT and SURF matches that demonstrate that our meth
od leads to accurate and fast model estimations.
****

Conservation Tracking
Martin Schiegg, Philipp Hanslovsky, Bernhard X. Kausler, Lars Hufnagel, Fred A.
Hamprecht; Proceedings of the IEEE International Conference on Computer Vision (
ICCV), 2013, pp. 2928-2935
The quality of any tracking-by-assignment hinges on the accuracy of the foregoin
g target detection / segmentation step. In many kinds of images, errors in this
first stage are unavoidable. These errors then propagate to, and corrupt, the tr
acking result. Our main contribution is the first probabilistic graphical model
that can explicitly account for overand undersegmentation errors even when the n
umber of tracking targets is unknown and when they may divide, as in cell cultur
es. The tracking model we present implements global consistency constraints for
the number of targets comprised by each detection and is solved to global optima

lity on reasonably large 2D+t and 3D+t datasets. In addition, we empirically dem
onstrate the effectiveness of a postprocessing that allows to establish target i
dentity even across occlusion / undersegmentation. The usefulness and efficiency
 of this new tracking method is demonstrated on three different and challenging
2D+t and 3D+t datasets from developmental biology.
****

Multiview Photometric Stereo Using Planar Mesh Parameterization
Jaesik Park, Sudipta N. Sinha, Yasuyuki Matsushita, Yu-Wing Tai, In So Kweon; Pr
oceedings of the IEEE International Conference on Computer Vision (ICCV), 2013,
pp. 1161-1168
We propose a method for accurate 3D shape reconstruction using uncalibrated mult
iview photometric stereo. A coarse mesh reconstructed using multiview stereo is
first parameterized using a planar mesh parameterization technique. Subsequently
, multiview photometric stereo is performed in the 2D parameter domain of the me
sh, where all geometric and photometric cues from multiple images can be treated
 uniformly. Unlike traditional methods, there is no need for merging view-depend
ent surface normal maps. Our key contribution is a new photometric stereo based
mesh refinement technique that can efficiently reconstruct meshes with extremely
 fine geometric details by directly estimating a displacement texture map in the
 2D parameter domain. We demonstrate that intricate surface geometry can be reco
nstructed using several challenging datasets containing surfaces with specular r
eflections, multiple albedos and complex topologies.
****

Handling Uncertain Tags in Visual Recognition
Arash Vahdat, Greg Mori; Proceedings of the IEEE International Conference on Com
puter Vision (ICCV), 2013, pp. 737-744
Gathering accurate training data for recognizing a set of attributes or tags on
images or videos is a challenge. Obtaining labels via manual effort or from weak
ly-supervised data typically results in noisy training labels. We develop the Fl
ipSVM, a novel algorithm for handling these noisy, structured labels. The FlipSV
M models label noise by "flipping" labels on training examples. We show empirica
lly that the FlipSVM is effective on images-and-attributes and video tagging dat
asets.
****

Finding Causal Interactions in Video Sequences
Mustafa Ayazoglu, Burak Yilmaz, Mario Sznaier, Octavia Camps; Proceedings of the
 IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3575-3582
This paper considers the problem of detecting causal interactions in video clips
. Specifically, the goal is to detect whether the actions of a given target can
be explained in terms of the past actions of a collection of other agents. We pr
opose to solve this problem by recasting it into a directed graph topology ident
ification, where each node corresponds to the observed motion of a given target,
 and each link indicates the presence of a causal correlation. As shown in the p
aper, this leads to a block-sparsification problem that can be efficiently solve
d using a modified Group-Lasso type approach, capable of handling missing data a
nd outliers (due for instance to occlusion and mis-identified correspondences).
Moreover, this approach also identifies time instants where the interactions bet
ween agents change, thus providing event detection capabilities. These results a
re illustrated with several examples involving non-trivial interactions amongst
several human subjects.
****

A Non-parametric Bayesian Network Prior of Human Pose
Andreas M. Lehrmann, Peter V. Gehler, Sebastian Nowozin; Proceedings of the IEEE
 International Conference on Computer Vision (ICCV), 2013, pp. 1281-1288
Having a sensible prior of human pose is a vital ingredient for many computer vi
sion applications, including tracking and pose estimation. While the application
 of global non-parametric approaches and parametric models has led to some succe
ss, finding the right balance in terms of flexibility and tractability, as well
as estimating model parameters from data has turned out to be challenging. In th
is work, we introduce a sparse Bayesian network model of human pose that is non-

parametric with respect to the estimation of both its graph structure and its lo cal distributions. We describe an efficient sampling scheme for our model and sh ow its tractability for the computation of exact log-likelihoods. We empirically validate our approach on the Human 3.6M dataset and demonstrate superior perfor mance to global models and parametric networks. We further illustrate our model' s ability to represent and compose poses not present in the training set (compos itionality) and describe a speed-accuracy trade-off that allows realtime scoring of poses.
****

Topology-Constrained Layered Tracking with Latent Flow
Jason Chang, John W. Fisher III; Proceedings of the IEEE International Conferenc e on Computer Vision (ICCV), 2013, pp. 161-168
We present an integrated probabilistic model for layered object tracking that co mbines dynamics on implicit shape representations, topological shape constraints , adaptive appearance models, and layered flow. The generative model combines th e evolution of appearances and layer shapes with a Gaussian process flow and exp licit layer ordering. Efficient MCMC sampling algorithms are developed to enable a particle filtering approach while reasoning about the distribution of object boundaries in video. We demonstrate the utility of the proposed tracking algorit hm on a wide variety of video sources while achieving state-of-the-art results o n a boundary-accurate tracking dataset.
****

Saliency Detection via Dense and Sparse Reconstruction
Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2976-2983
In this paper, we propose a visual saliency detection algorithm from the perspec tive of reconstruction errors. The image boundaries are first extracted via supe rpixels as likely cues for background templates, from which dense and sparse app earance models are constructed. For each image region, we first compute dense an d sparse reconstruction errors. Second, the reconstruction errors are propagated based on the contexts obtained from K-means clustering. Third, pixel-level sali ency is computed by an integration of multi-scale reconstruction errors and refi ned by an object-biased Gaussian model. We apply the Bayes formula to integrate saliency measures based on dense and sparse reconstruction errors. Experimental results show that the proposed algorithm performs favorably against seventeen st ate-of-the-art methods in terms of precision and recall. In addition, the propos ed algorithm is demonstrated to be more effective in highlighting salient object s uniformly and robust to background noise.
****

Style-Aware Mid-level Representation for Discovering Visual Connections in Space and Time
Yong Jae Lee, Alexei A. Efros, Martial Hebert; Proceedings of the IEEE Internati onal Conference on Computer Vision (ICCV), 2013, pp. 1857-1864
We present a weakly-supervised visual data mining approach that discovers connec tions between recurring midlevel visual elements in historic (temporal) and geog raphic (spatial) image collections, and attempts to capture the underlying visua l style. In contrast to existing discovery methods that mine for patterns that r emain visually consistent throughout the dataset, our goal is to discover visual elements whose appearance changes due to change in time or location; i.e., exhi bit consistent stylistic variations across the label space (date or geo-location). To discover these elements, we first identify groups of patches that are styl esensitive. We then incrementally build correspondences to find the same element across the entire dataset. Finally, we train style-aware regressors that model each element's range of stylistic differences. We apply our approach to date and geo-location prediction and show substantial improvement over several baselines that do not model visual style. We also demonstrate the method's effectiveness on the related task of fine-grained classification.
****

Synergistic Clustering of Image and Segment Descriptors for Unsupervised Scene U nderstanding

Daniel M. Steinberg, Oscar Pizarro, Stefan B. Williams; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3463-3470
With the advent of cheap, high fidelity, digital imaging systems, the quantity and rate of generation of visual data can dramatically outpace a humans ability to label or annotate it. In these situations there is scope for the use of unsupervised approaches that can model these datasets and automatically summarise their content. To this end, we present a totally unsupervised, and annotation-less, model for scene understanding. This model can simultaneously cluster whole-image and segment descriptors, thereby forming an unsupervised model of scenes and objects. We show that this model outperforms other unsupervised models that can only cluster one source of information (image or segment) at once. We are able to compare unsupervised and supervised techniques using standard measures derived from confusion matrices and contingency tables. This shows that our unsupervised model is competitive with current supervised and weakly-supervised models for scene understanding on standard datasets. We also demonstrate our model operating on a dataset with more than 100,000 images collected by an autonomous underwater vehicle.
****

Multi-channel Correlation Filters
Hamed Kiani Galoogahi, Terence Sim, Simon Lucey; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3072-3079
Modern descriptors like HOG and SIFT are now commonly used in vision for pattern detection within image and video. From a signal processing perspective, this detection process can be efficiently posed as a correlation/convolution between a multi-channel image and a multi-channel detector/filter which results in a singlechannel response map indicating where the pattern (e.g. object) has occurred. In this paper, we propose a novel framework for learning a multi-channel detector/filter efficiently in the frequency domain, both in terms of training time and memory footprint, which we refer to as a multichannel correlation filter. To demonstrate the effectiveness of our strategy, we evaluate it across a number of visual detection/localization tasks where we: (i) exhibit superior performance to current state of the art correlation filters, and (ii) superior computational and memory efficiencies compared to state of the art spatial detectors.
****

Image Set Classification Using Holistic Multiple Order Statistics Features and Localized Multi-kernel Metric Learning
Jiwen Lu, Gang Wang, Pierre Moulin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 329-336
This paper presents a new approach for image set classification, where each training and testing example contains a set of image instances of an object captured from varying viewpoints or under varying illuminations. While a number of image set classification methods have been proposed in recent years, most of them model each image set as a single linear subspace or mixture of linear subspaces, which may lose some discriminative information for classification. To address this, we propose exploring multiple order statistics as features of image sets, and develop a localized multikernel metric learning (LMKML) algorithm to effectively combine different order statistics information for classification. Our method achieves the state-of-the-art performance on four widely used databases including the Honda/UCSD, CMU Mobo, and Youtube face datasets, and the ETH-80 object dataset.
****

Modeling Occlusion by Discriminative AND-OR Structures
Bo Li, Wenze Hu, Tianfu Wu, Song-Chun Zhu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2560-2567
Occlusion presents a challenge for detecting objects in real world applications. To address this issue, this paper models object occlusion with an AND-OR structure which (i) represents occlusion at semantic part level, and (ii) captures the regularities of different occlusion configurations (i.e., the different combinations of object part visibilities). This paper focuses on car detection on street. Since annotating part occlusion on real images is time-consuming and error-pr

one, we propose to learn the the AND-OR structure automatically using synthetic images of CAD models placed at different relative positions. The model parameters are learned from real images under the latent structural SVM (LSSVM) framework. In inference, an efficient dynamic programming (DP) algorithm is utilized. In experiments, we test our method on both car detection and car view estimation. Experimental results show that (i) Our CAD simulation strategy is capable of generating occlusion patterns for real scenarios, (ii) The proposed AND-OR structure model is effective for modeling occlusions, which outperforms the deformable part-based model (DPM) [6, 10] in car detection on both our self-collected street parking dataset and the Pascal VOC 2007 car dataset [4], (iii) The learned model is on-par with the state-of-the-art methods on car view estimation tested on two public datasets.

****

Breaking the Chain: Liberation from the Temporal Markov Assumption for Tracking Human Poses

Ryan Tokola, Wongun Choi, Silvio Savarese; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2424-2431

We present an approach to multi-target tracking that has expressive potential beyond the capabilities of chainshaped hidden Markov models, yet has significantly reduced complexity. Our framework, which we call tracking-byselection, is similar to tracking-by-detection in that it separates the tasks of detection and tracking, but it shifts temporal reasoning from the tracking stage to the detection stage. The core feature of tracking-by-selection is that it reasons about path hypotheses that traverse the entire video instead of a chain of single-frame object hypotheses. A traditional chain-shaped tracking-by-detection model is only able to promote consistency between one frame and the next. In tracking-by-selection, path hypotheses exist across time, and encouraging long-term temporal consistency is as simple as rewarding path hypotheses with consistent image features. One additional advantage of tracking-by-selection is that it results in a dramatically simplified model that can be solved exactly. We adapt an existing tracking-by-detection model to the tracking-by-selection framework, and show improved performance on a challenging dataset (introduced in [18]).

****

ACTIVE: Activity Concept Transitions in Video Event Classification

Chen Sun, Ram Nevatia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 913-920

The goal of high level event classification from videos is to assign a single, high level event label to each query video. Traditional approaches represent each video as a set of low level features and encode it into a fixed length feature vector (e.g. Bag-of-Words), which leave a big gap between low level visual features and high level events. Our paper tries to address this problem by exploiting activity concept transitions in video events (ACTIVE). A video is treated as a sequence of short clips, all of which are observations corresponding to latent activity concept variables in a Hidden Markov Model (HMM). We propose to apply Fisher Kernel techniques so that the concept transitions over time can be encoded into a compact and fixed length feature vector very efficiently. Our approach can utilize concept annotations from independent datasets, and works well even with a very small number of training samples. Experiments on the challenging NIST TRECVID Multimedia Event Detection (MED) dataset shows our approach performs favorably over the state-of-the-art.

****

A Joint Intensity and Depth Co-sparse Analysis Model for Depth Map Super-resolution

Martin Kiechle, Simon Hawe, Martin Kleinsteuber; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1545-1552

High-resolution depth maps can be inferred from lowresolution depth measurements and an additional highresolution intensity image of the same scene. To that end, we introduce a bimodal co-sparse analysis model, which is able to capture the interdependency of registered intensity and depth information. This model is based on the assumption that the co-supports of corresponding bimodal image structu

res are aligned when computed by a suitable pair of analysis operators. No analy tic form of such operators exist and we propose a method for learning them from a set of registered training signals. This learning process is done offline and returns a bimodal analysis operator that is universally applicable to natural sc enes. We use this to exploit the bimodal co-sparse analysis model as a prior for solving inverse problems, which leads to an efficient algorithm for depth map s uper-resolution.
****

Towards Understanding Action Recognition
Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, Michael J. Black; P roceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3192-3199
Although action recognition in videos is widely studied, current methods often f ail on real-world datasets. Many recent approaches improve accuracy and robustne ss to cope with challenging video sequences, but it is often unclear what affect s the results most. This paper attempts to provide insights based on a systemati c performance evaluation using thoroughly-annotated data of human actions. We an notate human Joints for the HMDB dataset (J-HMDB). This annotation can be used t o derive ground truth optical flow and segmentation. We evaluate current methods using this dataset and systematically replace the output of various algorithms with ground truth. This enables us to discover what is important for example, sh ould we work on improving flow algorithms, estimating human bounding boxes, or e nabling pose estimation? In summary, we find that highlevel pose features greatl y outperform low/mid level features; in particular, pose over time is critical. While current pose estimation algorithms are far from perfect, features extracte d from estimated pose on a subset of J-HMDB, in which the full body is visible, outperform low/mid-level features. We also find that the accuracy of the action recognition framework can be greatly increased by refining the underlying low/mi d level features; this suggests it is important to improve optical flow and huma n detection algorithms. Our analysis and J-HMDB dataset should facilitate a deep er understanding of action recognition algorithms.
****

Go-ICP: Solving 3D Registration Efficiently and Globally Optimally
Jiaolong Yang, Hongdong Li, Yunde Jia; Proceedings of the IEEE International Con ference on Computer Vision (ICCV), 2013, pp. 1457-1464
Registration is a fundamental task in computer vision. The Iterative Closest Poi nt (ICP) algorithm is one of the widely-used methods for solving the registratio n problem. Based on local iteration, ICP is however well-known to suffer from lo cal minima. Its performance critically relies on the quality of initialization, and only local optimality is guaranteed. This paper provides the very first glob ally optimal solution to Euclidean registration of two 3D pointsets or two 3D su rfaces under the L 2 error. Our method is built upon ICP, but combines it with a branch-and-bound (BnB) scheme which searches the 3D motion space SE(3) efficien tly. By exploiting the special structure of the underlying geometry, we derive n ovel upper and lower bounds for the ICP error function. The integration of local ICP and global BnB enables the new method to run efficiently in practice, and i ts optimality is exactly guaranteed. We also discuss extensions, addressing the issue of outlier robustness.
****

Geometric Registration Based on Distortion Estimation
Wei Zeng, Mayank Goswami, Feng Luo, Xianfeng Gu; Proceedings of the IEEE Interna tional Conference on Computer Vision (ICCV), 2013, pp. 2632-2639
Surface registration plays a fundamental role in many applications in computer v ision and aims at finding a oneto-one correspondence between surfaces. Conformal mapping based surface registration methods conformally map 2D/3D surfaces onto 2D canonical domains and perform the matching on the 2D plane. This registration framework reduces dimensionality, and the result is intrinsic to Riemannian met ric and invariant under isometric deformation. However, conformal mapping will b e affected by inconsistent boundaries and non-isometric deformations of surfaces . In this work, we quantify the effects of boundary variation and non-isometric

deformation to conformal mappings, and give the theoretical upper bounds for the distortions of conformal mappings under these two factors. Besides giving the thorough theoretical proofs of the theorems, we verified them by concrete experiments using 3D human facial scans with dynamic expressions and varying boundaries. Furthermore, we used the distortion estimates for reducing search range in feature matching of surface registration applications. The experimental results are consistent with the theoretical predictions and also demonstrate the performance improvements in feature tracking.
****

Handwritten Word Spotting with Corrected Attributes
Jon Almazan, Albert Gordo, Alicia Fornes, Ernest Valveny; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1017-1024
We propose an approach to multi-writer word spotting, where the goal is to find a query word in a dataset comprised of document images. We propose an attributes-based approach that leads to a low-dimensional, fixed-length representation of the word images that is fast to compute and, especially, fast to compare. This approach naturally leads to an unified representation of word images and strings, which seamlessly allows one to indistinctly perform queryby-example, where the query is an image, and query-bystring, where the query is a string. We also propose a calibration scheme to correct the attributes scores based on Canonical Correlation Analysis that greatly improves the results on a challenging dataset. We test our approach on two public datasets showing state-of-the-art results.
****

Interactive Markerless Articulated Hand Motion Tracking Using RGB and Depth Data
Srinath Sridhar, Antti Oulasvirta, Christian Theobalt; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2456-2463
Tracking the articulated 3D motion of the hand has important applications, for example, in human-computer interaction and teleoperation. We present a novel method that can capture a broad range of articulated hand motions at interactive rates. Our hybrid approach combines, in a voting scheme, a discriminative, part-based pose retrieval method with a generative pose estimation method based on local optimization. Color information from a multiview RGB camera setup along with a person-specific hand model are used by the generative method to find the pose that best explains the observed images. In parallel, our discriminative pose estimation method uses fingertips detected on depth data to estimate a complete or partial pose of the hand by adopting a part-based pose retrieval strategy. This part-based strategy helps reduce the search space drastically in comparison to a global pose retrieval strategy. Quantitative results show that our method achieves state-of-the-art accuracy on challenging sequences and a near-realtime performance of 10 fps on a desktop computer.
****

Network Principles for SfM: Disambiguating Repeated Structures with Local Context
Kyle Wilson, Noah Snavely; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 513-520
Repeated features are common in urban scenes. Many objects, such as clock towers with nearly identical sides, or domes with strong radial symmetries, pose challenges for structure from motion. When similar but distinct features are mistakenly equated, the resulting 3D reconstructions can have errors ranging from phantom walls and superimposed structures to a complete failure to reconstruct. We present a new approach to solving such problems by considering the local visibility structure of such repeated features. Drawing upon network theory, we present a new way of scoring features using a measure of local clustering. Our model leads to a simple, fast, and highly scalable technique for disambiguating repeated features based on an analysis of an underlying visibility graph, without relying on explicit geometric reasoning. We demonstrate our method on several very large datasets drawn from Internet photo collections, and compare it to a more traditional geometry-based disambiguation technique.
****

Improving Graph Matching via Density Maximization

Chao Wang, Lei Wang, Lingqiao Liu; Proceedings of the IEEE International Confere nce on Computer Vision (ICCV), 2013, pp. 3424-3431
Graph matching has been widely used in various applications in computer vision d ue to its powerful performance. However, it poses three challenges to image spar se feature matching: (1) The combinatorial nature limits the size of the possibl e matches; (2) It is sensitive to outliers because the objective function prefer s more matches; (3) It works poorly when handling many-to-many object correspond ences, due to its assumption of one single cluster for each graph. In this paper , we address these problems with a unified framework--Density Maximization. We p ropose a graph density local estimator (DLE) to measure the quality of matches. Density Maximization aims to maximize the DLE values both locally and globally. The local maximization of auebfinds the clusters of nodes as well as eliminates the outliers. The global maximization of ela efficiently refines the matches by exploring a much larger matching space. Our Density Maximization is orthogonal t o specific graph matching algorithms. Experimental evaluation demonstrates that it significantly boosts the true matches and enables graph matching to handle bo th outliers and many-to-many object correspondences.
****

A Unified Rolling Shutter and Motion Blur Model for 3D Visual Registration
Maxime Meilland, Tom Drummond, Andrew I. Comport; Proceedings of the IEEE Intern ational Conference on Computer Vision (ICCV), 2013, pp. 2016-2023
Motion blur and rolling shutter deformations both inhibit visual motion registra tion, whether it be due to a moving sensor or a moving target. Whilst both defor mations exist simultaneously, no models have been proposed to handle them togeth er. Furthermore, neither deformation has been considered previously in the conte xt of monocular fullimage 6 degrees of freedom registration or RGB-D structure a nd motion. As will be shown, rolling shutter deformation is observed when a came ra moves faster than a single pixel in parallax between subsequent scan-lines. B lur is a function of the pixel exposure time and the motion vector. In this pape r a complete dense 3D registration model will be derived to account for both mot ion blur and rolling shutter deformations simultaneously. Various approaches wil l be compared with respect to ground truth and live real-time performance will b e demonstrated for complex scenarios where both blur and shutter deformations ar e dominant.
****

Fibonacci Exposure Bracketing for High Dynamic Range Imaging
Mohit Gupta, Daisuke Iso, Shree K. Nayar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1473-1480
Exposure bracketing for high dynamic range (HDR) imaging involves capturing seve ral images of the scene at different exposures. If either the camera or the scen e moves during capture, the captured images must be registered. Large exposure d ifferences between bracketed images lead to inaccurate registration, resulting i n artifacts such as ghosting (multiple copies of scene objects) and blur. We pre sent two techniques, one for image capture (Fibonacci exposure bracketing) and o ne for image registration (generalized registration), to prevent such motion-rel ated artifacts. Fibonacci bracketing involves capturing a sequence of images suc h that each exposure time is the sum of the previous N (N > 1) exposures. Genera lized registration involves estimating motion between sums of contiguous sets of frames, instead of between individual frames. Together, the two techniques ensu re that motion is always estimated between frames of the same total exposure tim e. This results in HDR images and videos which have both a large dynamic range a nd minimal motion-related artifacts. We show, by results for several real-world indoor and outdoor scenes, that the proposed approach significantly outperforms several existing bracketing schemes.
****

Potts Model, Parametric Maxflow and K-Submodular Functions
Igor Gridchyn, Vladimir Kolmogorov; Proceedings of the IEEE International Confer ence on Computer Vision (ICCV), 2013, pp. 2320-2327
The problem of minimizing the Potts energy function frequently occurs in compute r vision applications. One way to tackle this NP-hard problem was proposed by Ko

vtun [20, 21]. It identifies a part of an optimal solution by running k maxflow computations, where k is the number of labels. The number of "labeled" pixels can be significant in some applications, e.g. 50-93% in our tests for stereo. We show how to reduce the runtime to O(log k) maxflow computations (or one parametric maxflow computation). Furthermore, the output of our algorithm allows to speed-up the subsequent alpha expansion for the unlabeled part, or can be used as it is for time-critical applications. To derive our technique, we generalize the algorithm of Felzenszwalb et al. [7] for Tree Metrics. We also show a connection to k-submodular functions from combinatorial optimization, and discuss k-submodular relaxations for general energy functions.
****

Target-Driven Moire Pattern Synthesis by Phase Modulation
Pei-Hen Tsai, Yung-Yu Chuang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1912-1919
This paper investigates an approach for generating two grating images so that the moir?? pattern of their superposition resembles the target image. Our method is grounded on the fundamental moir?? theorem. By focusing on the visually most dominant (1, -1)-moir?? component, we obtain the phase modulation constraint on the phase shifts between the two grating images. For improving visual appearance of the grating images and hiding capability the embedded image, a smoothness term is added to spread information between the two grating images and an appearance phase function is used to add irregular structures into grating images. The grating images can be printed on transparencies and the hidden image decoding can be performed optically by overlaying them together. The proposed method enables the creation of moir?? art and allows visual decoding without computers.
****

Implied Feedback: Learning Nuances of User Behavior in Image Search
Devi Parikh, Kristen Grauman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 745-752
User feedback helps an image search system refine its relevance predictions, tailoring the search towards the user's preferences. Existing methods simply take feedback at face value: clicking on an image means the user wants things like it; commenting that an image lacks a specific attribute means the user wants things that have it. However, we expect there is actually more information behind the user's literal feedback. In particular, a user's (possibly subconscious) search strategy leads him to comment on certain images rather than others, based on how any of the visible candidate images compare to the desired content. For example, he may be more likely to give negative feedback on an irrelevant image that is relatively close to his target, as opposed to bothering with one that is altogether different. We introduce novel features to capitalize on such implied feedback cues, and learn a ranking function that uses them to improve the system's relevance estimates. We validate the approach with real users searching for shoes, faces, or scenes using two different modes of feedback: binary relevance feedback and relative attributes-based feedback. The results show that retrieval improves significantly when the system accounts for the learned behaviors. We show that the nuances learned are domain-invariant, and useful for both generic user-independent search as well as personalized user-specific search.
****

How Related Exemplars Help Complex Event Detection in Web Videos?
Yi Yang, Zhigang Ma, Zhongwen Xu, Shuicheng Yan, Alexander G. Hauptmann; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2104-2111
Compared to visual concepts such as actions, scenes and objects, complex event is a higher level abstraction of longer video sequences. For example, a "marriage proposal" event is described by multiple objects (e.g., ring, faces), scenes (e.g., in a restaurant, outdoor) and actions (e.g., kneeling down). The positive exemplars which exactly convey the precise semantic of an event are hard to obtain. It would be beneficial to utilize the related exemplars for complex event detection. However, the semantic correlations between related exemplars and the target event vary substantially as relatedness assessment is subjective. Two relate

d exemplars can be about completely different events, e.g., in the TRECVID MED dataset, both bicycle riding and equestrianism are labeled as related to "attempting a bike trick" event. To tackle the subjectiveness of human assessment, our algorithm automatically evaluates how positive the related exemplars are for the detection of an event and uses them on an exemplar-specific basis. Experiments demonstrate that our algorithm is able to utilize related exemplars adaptively, and the algorithm gains good performance for complex event detection.
****

Decomposing Bag of Words Histograms
Ankit Gandhi, Karteek Alahari, C.V. Jawahar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 305-312
We aim to decompose a global histogram representation of an image into histograms of its associated objects and regions. This task is formulated as an optimization problem, given a set of linear classifiers, which can effectively discriminate the object categories present in the image. Our decomposition bypasses harder problems associated with accurately localizing and segmenting objects. We evaluate our method on a wide variety of composite histograms, and also compare it with MRF -based solutions. In addition to merely measuring the accuracy of decomposition, we also show the utility of the estimated object and background histograms for the task of image classification on the PASCAL VOC 2007 dataset.
****

Higher Order Matching for Consistent Multiple Target Tracking
Chetan Arora, Amir Globerson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 177-184
This paper addresses the data assignment problem in multi frame multi object tracking in video sequences. Traditional methods employing maximum weight bipartite matching offer limited temporal modeling. It has recently been shown [6, 8, 24] that incorporating higher order temporal constraints improves the assignment solution. Finding maximum weight matching with higher order constraints is however NP-hard and the solutions proposed until now have either been greedy [8] or rely on greedy rounding of the solution obtained from spectral techniques [15]. We propose a novel algorithm to find the approximate solution to data assignment problem with higher order temporal constraints using the method of dual decomposition and the MPLP message passing algorithm [21]. We compare the proposed algorithm with an implementation of [8] and [15] and show that proposed technique provides better solution with a bound on approximation factor for each inferred solution.
****

Complementary Projection Hashing
Zhongming Jin, Yao Hu, Yue Lin, Debing Zhang, Shiding Lin, Deng Cai, Xuelong Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 257-264
Recently, hashing techniques have been widely applied to solve the approximate nearest neighbors search problem in many vision applications. Generally, these hashing approaches generate 2 c buckets, where c is the length of the hash code. A good hashing method should satisfy the following two requirements: 1) mapping the nearby data points into the same bucket or nearby (measured by the Hamming distance) buckets. 2) all the data points are evenly distributed among all the buckets. In this paper, we propose a novel algorithm named Complementary Projection Hashing (CPH) to find the optimal hashing functions which explicitly considers the above two requirements. Specifically, CPH aims at sequentially finding a series of hyperplanes (hashing functions) which cross the sparse region of the data. At the same time, the data points are evenly distributed in the hypercubes generated by these hyperplanes. The experiments comparing with the state-of-the-art hashing methods demonstrate the effectiveness of the proposed method.
****

Super-resolution via Transform-Invariant Group-Sparse Regularization
Carlos Fernandez-Granda, Emmanuel J. CandÃ¨s; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3336-3343
We present a framework to super-resolve planar regions found in urban scenes and

other man-made environments by taking into account their 3D geometry. Such regi ons have highly structured straight edges, but this prior is challenging to expl oit due to deformations induced by the projection onto the imaging plane. Our me thod factors out such deformations by using recently developed tools based on co nvex optimization to learn a transform that maps the image to a domain where its gradient has a simple group-sparse structure. This allows to obtain a novel con vex regularizer that enforces global consistency constraints between the edges o f the image. Computational experiments with real images show that this data-driv en approach to the design of regularizers promoting transform-invariant group sp arsity is very effective at high super-resolution factors. We view our approach as complementary to most recent superresolution methods, which tend to focus on hallucinating high-frequency textures.

****

Inferring "Dark Matter" and "Dark Energy" from Videos

Dan Xie, Sinisa Todorovic, Song-Chun Zhu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2224-2231

This paper presents an approach to localizing functional objects in surveillance videos without domain knowledge about semantic object classes that may appear i n the scene. Functional objects do not have discriminative appearance and shape, but they affect behavior of people in the scene. For example, they "attract" pe ople to approach them for satisfying certain needs (e.g., vending machines could quench thirst), or "repel" people to avoid them (e.g., grass lawns). Therefore, functional objects can be viewed as "dark matter", emanating "dark energy" that affects people's trajectories in the video. To detect "dark matter" and infer t heir "dark energy" field, we extend the Lagrangian mechanics. People are treated as particle-agents with latent intents to approach "dark matter" and thus satis fy their needs, where their motions are subject to a composite "dark energy" fie ld of all functional objects in the scene. We make the assumption that people ta ke globally optimal paths toward the intended "dark matter" while avoiding laten t obstacles. A Bayesian framework is used to probabilistically model: people's t rajectories and intents, constraint map of the scene, and locations of functiona l objects. A data-driven Markov Chain Monte Carlo (MCMC) process is used for inf erence. Our evaluation on videos of public squares and courtyards demonstrates o ur effectiveness in localizing functional objects and predicting people's trajec tories in unobserved parts of the video footage.

****

Optimal Orthogonal Basis and Image Assimilation: Motion Modeling

Etienne Huot, Giuseppe Papari, Isabelle Herlin; Proceedings of the IEEE Internat ional Conference on Computer Vision (ICCV), 2013, pp. 3352-3359

This paper describes modeling and numerical computation of orthogonal bases, whi ch are used to describe images and motion fields. Motion estimation from image d ata is then studied on subspaces spanned by these bases. A reduced model is obta ined as the Galerkin projection on these subspaces of a physical model, based on Euler and optical flow equations. A data assimilation method is studied, which assimilates coefficients of image data in the reduced model in order to estimate motion coefficients. The approach is first quantified on synthetic data: it dem onstrates the interest of model reduction as a compromise between results qualit y and computational cost. Results obtained on real data are then displayed so as to illustrate the method.

****

Detecting Avocados to Zucchinis: What Have We Done, and Where Are We Going?

Olga Russakovsky, Jia Deng, Zhiheng Huang, Alexander C. Berg, Li Fei-Fei; Procee dings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2064-2071

The growth of detection datasets and the multiple directions of object detection research provide both an unprecedented need and a great opportunity for a thoro ugh evaluation of the current state of the field of categorical object detection . In this paper we strive to answer two key questions. First, where are we curre ntly as a field: what have we done right, what still needs to be improved? Secon d, where should we be going in designing the next generation of object detectors

? Inspired by the recent work of Hoiem et al. [10] on the standard PASCAL VOC de tection dataset, we perform a large-scale study on the ImageNet Large Scale Visu al Recognition Challenge (ILSVRC) data. First, we quantitatively demonstrate tha t this dataset provides many of the same detection challenges as the PASCAL VOC. Due to its scale of 1000 object categories, ILSVRC also provides an excellent t estbed for understanding the performance of detectors as a function of several k ey properties of the object classes. We conduct a series of analyses looking at how different detection methods perform on a number of imagelevel and object-cla ss-level properties such as texture, color, deformation, and clutter. We learn i mportant lessons of the current object detection methods and propose a number of insights for designing the next generation object detectors.
****

Neighbor-to-Neighbor Search for Fast Coding of Feature Vectors
Nakamasa Inoue, Koichi Shinoda; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1233-1240
Assigning a visual code to a low-level image descriptor, which we call code assi gnment, is the most computationally expensive part of image classification algor ithms based on the bag of visual word (BoW) framework. This paper proposes a fas t computation method, Neighbor-toNeighbor (NTN) search, for this code assignment . Based on the fact that image features from an adjacent region are usually simi lar to each other, this algorithm effectively reduces the cost of calculating th e distance between a codeword and a feature vector. This method can be applied n ot only to a hard codebook constructed by vector quantization (NTN-VQ), but also to a soft codebook, a Gaussian mixture model (NTN-GMM). We evaluated this metho d on the PASCAL VOC 2007 classification challenge task. NTN-VQ reduced the assig nment cost by 77.4% in super-vector coding, and NTN-GMM reduced it by 89.3% in F isher-vector coding, without any significant degradation in classification perfo rmance.
****

Flattening Supervoxel Hierarchies by the Uniform Entropy Slice
Chenliang Xu, Spencer Whitt, Jason J. Corso; Proceedings of the IEEE Internation al Conference on Computer Vision (ICCV), 2013, pp. 2240-2247
Supervoxel hierarchies provide a rich multiscale decomposition of a given video suitable for subsequent processing in video analysis. The hierarchies are typica lly computed by an unsupervised process that is susceptible to undersegmentation at coarse levels and over-segmentation at fine levels, which make it a challeng e to adopt the hierarchies for later use. In this paper, we propose the first me thod to overcome this limitation and flatten the hierarchy into a single segment ation. Our method, called the uniform entropy slice, seeks a selection of superv oxels that balances the relative level of information in the selected supervoxel s based on some post hoc feature criterion such as objectness. For example, with this criterion, in regions nearby objects, our method prefers finer supervoxels to capture the local details, but in regions away from any objects we prefer co arser supervoxels. We formulate the uniform entropy slice as a binary quadratic program and implement four different feature criteria, both unsupervised and sup ervised, to drive the flattening. Although we apply it only to supervoxel hierar chies in this paper, our method is generally applicable to segmentation tree hie rarchies. Our experiments demonstrate both strong qualitative performance and su perior quantitative performance to state of the art baselines on benchmark inter net videos.
****

Find the Best Path: An Efficient and Accurate Classifier for Image Hierarchies
Min Sun, Wan Huang, Silvio Savarese; Proceedings of the IEEE International Confe rence on Computer Vision (ICCV), 2013, pp. 265-272
Many methods have been proposed to solve the image classification problem for a large number of categories. Among them, methods based on tree-based representati ons achieve good trade-off between accuracy and test time efficiency. While focu sing on learning a tree-shaped hierarchy and the corresponding set of classifier s, most of them [11, 2, 14] use a greedy prediction algorithm for test time effi ciency. We argue that the dramatic decrease in accuracy at high efficiency is ca

used by the specific design choice of the learning and greedy prediction algorithms. In this work, we propose a classifier which achieves a better trade-off between efficiency and accuracy with a given tree-shaped hierarchy. First, we convert the classification problem as finding the best path in the hierarchy, and a novel branchand-bound-like algorithm is introduced to efficiently search for the best path. Second, we jointly train the classifiers using a novel Structured SVM (SSVM) formulation with additional bound constraints. As a result, our method achieves a significant 4.65%, 5.43%, and 4.07% (relative 24.82%, 41.64%, and 109.79%) improvement in accuracy at high efficiency compared to state-of-the-art greedy "tree-based" methods [14] on Caltech-256 [15], SUN [32] and ImageNet 1K [9] dataset, respectively. Finally, we show that our branch-and-bound-like algorithm naturally ranks the paths in the hierarchy (Fig. 8) so that users can further process them.

****

Model Recommendation with Virtual Probes for Egocentric Hand Detection

Cheng Li, Kris M. Kitani; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2624-2631

Egocentric cameras can be used to benefit such tasks as analyzing fine motor skills, recognizing gestures and learning about hand-object manipulation. To enable such technology, we believe that the hands must detected on the pixellevel to gain important information about the shape of the hands and fingers. We show that the problem of pixel-wise hand detection can be effectively solved, by posing the problem as a model recommendation task. As such, the goal of a recommendation system is to recommend the n-best hand detectors based on the probe set a small amount of labeled data from the test distribution. This requirement of a probe set is a serious limitation in many applications, such as ego-centric hand detection, where the test distribution may be continually changing. To address this limitation, we propose the use of virtual probes which can be automatically extracted from the test distribution. The key idea is that many features, such as the color distribution or relative performance between two detectors, can be used as a proxy to the probe set. In our experiments we show that the recommendation paradigm is well-equipped to handle complex changes in the appearance of the hands in firstperson vision. In particular, we show how our system is able to generalize to new scenarios by testing our model across multiple users.

****

Video Motion for Every Visible Point

Susanna Ricco, Carlo Tomasi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2464-2471

Dense motion of image points over many video frames can provide important information about the world. However, occlusions and drift make it impossible to compute long motion paths by merely concatenating optical flow vectors between consecutive frames. Instead, we solve for entire paths directly, and flag the frames in which each is visible. As in previous work, we anchor each path to a unique pixel which guarantees an even spatial distribution of paths. Unlike earlier methods, we allow paths to be anchored in any frame. By explicitly requiring that at least one visible path passes within a small neighborhood of every pixel, we guarantee complete coverage of all visible points in all frames. We achieve state-of-the-art results on real sequences including both rigid and non-rigid motions with significant occlusions.

****

Camera Alignment Using Trajectory Intersections in Unsynchronized Videos

Thomas Kuo, Santhoshkumar Sunderrajan, B.S. Manjunath; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1121-1128

This paper addresses the novel and challenging problem of aligning camera views that are unsynchronized by low and/or variable frame rates using object trajectories. Unlike existing trajectory-based alignment methods, our method does not require frame-to-frame synchronization. Instead, we propose using the intersections of corresponding object trajectories to match views. To find these intersections, we introduce a novel trajectory matching algorithm based on matching Spatio-Temporal Context Graphs (STCGs). These graphs represent the distances between tr

ajectories in time and space within a view, and are matched to an STCG from anot
her view to find the corresponding trajectories. To the best of our knowledge, t
his is one of the first attempts to align views that are unsynchronized with var
iable frame rates. The results on simulated and real-world datasets show traject
ory intersections are a viable feature for camera alignment, and that the trajec
tory matching method performs well in real-world scenarios.
****

A Unified Video Segmentation Benchmark: Annotation, Metrics and Analysis
Fabio Galasso, Naveen Shankar Nagaraja, Tatiana Jimenez Cardenas, Thomas Brox, B
ernt Schiele; Proceedings of the IEEE International Conference on Computer Visio
n (ICCV), 2013, pp. 3527-3534
Video segmentation research is currently limited by the lack of a benchmark data
set that covers the large variety of subproblems appearing in video segmentation
 and that is large enough to avoid overfitting. Consequently, there is little an
alysis of video segmentation which generalizes across subtasks, and it is not ye
t clear which and how video segmentation should leverage the information from th
e still-frames, as previously studied in image segmentation, alongside video spe
cific information, such as temporal volume, motion and occlusion. In this work w
e provide such an analysis based on annotations of a large video dataset, where
each video is manually segmented by multiple persons. Moreover, we introduce a n
ew volume-based metric that includes the important aspect of temporal consistenc
y, that can deal with segmentation hierarchies, and that reflects the tradeoff b
etween over-segmentation and segmentation accuracy.
****

Optical Flow via Locally Adaptive Fusion of Complementary Data Costs
Tae Hyun Kim, Hee Seok Lee, Kyoung Mu Lee; Proceedings of the IEEE International
 Conference on Computer Vision (ICCV), 2013, pp. 3344-3351
Many state-of-the-art optical flow estimation algorithms optimize the data and r
egularization terms to solve ill-posed problems. In this paper, in contrast to t
he conventional optical flow framework that uses a single or fixed data model, w
e study a novel framework that employs locally varying data term that adaptively
 combines different multiple types of data models. The locally adaptive data ter
m greatly reduces the matching ambiguity due to the complementary nature of the
multiple data models. The optimal number of complementary data models is learnt
by minimizing the redundancy among them under the minimum description length con
straint (MDL). From these chosen data models, a new optical flow estimation ener
gy model is designed with the weighted sum of the multiple data models, and a co
nvex optimization-based highly effective and practical solution that finds the o
ptical flow, as well as the weights is proposed. Comparative experimental result
s on the Middlebury optical flow benchmark show that the proposed method using t
he complementary data models outperforms the state-ofthe art methods.
****

Shortest Paths with Curvature and Torsion
Petter Strandmark, Johannes Ulen, Fredrik Kahl, Leo Grady; Proceedings of the IE
EE International Conference on Computer Vision (ICCV), 2013, pp. 2024-2031
This paper describes a method of finding thin, elongated structures in images an
d volumes. We use shortest paths to minimize very general functionals of higher-
order curve properties, such as curvature and torsion. Our globally optimal meth
od uses line graphs and its runtime is polynomial in the size of the discretizat
ion, often in the order of seconds on a single computer. To our knowledge, we ar
e the first to perform experiments in three dimensions with curvature and torsio
n regularization. The largest graphs we process have almost one hundred billion
arcs. Experiments on medical images and in multi-view reconstruction show the si
gnificance and practical usefulness of regularization based on curvature while t
orsion is still only tractable for small-scale problems.
****

Multi-view 3D Reconstruction from Uncalibrated Radially-Symmetric Cameras
Jae-Hak Kim, Yuchao Dai, Hongdong Li, Xin Du, Jonghyuk Kim; Proceedings of the I
EEE International Conference on Computer Vision (ICCV), 2013, pp. 1896-1903
We present a new multi-view 3D Euclidean reconstruction method for arbitrary unc

alibrated radially-symmetric cameras, which needs no calibration or any camera m
odel parameters other than radial symmetry. It is built on the radial 1D camera
model [25], a unified mathematical abstraction to different types of radially-sy
mmetric cameras. We formulate the problem of multi-view reconstruction for radia
l 1D cameras as a matrix rank minimization problem. Efficient implementation bas
ed on alternating direction continuation is proposed to handle scalability issue
for real-world applications. Our method applies to a wide range of omnidirectio
nal cameras including both dioptric and catadioptric (central and non-central) c
ameras. Additionally, our method deals with complete and incomplete measurements
under a unified framework elegantly. Experiments on both synthetic and real ima
ges from various types of cameras validate the superior performance of our new m
ethod, in terms of numerical accuracy and robustness.
****

Illuminant Chromaticity from Image Sequences
Veronique Prinet, Dani Lischinski, Michael Werman; Proceedings of the IEEE Inter
national Conference on Computer Vision (ICCV), 2013, pp. 3320-3327
We estimate illuminant chromaticity from temporal sequences, for scenes illumina
ted by either one or two dominant illuminants. While there are many methods for
illuminant estimation from a single image, few works so far have focused on vide
os, and even fewer on multiple light sources. Our aim is to leverage information
provided by the temporal acquisition, where either the objects or the camera or
the light source are/is in motion in order to estimate illuminant color without
the need for user interaction or using strong assumptions and heuristics. We in
troduce a simple physically-based formulation based on the assumption that the i
ncident light chromaticity is constant over a short space-time domain. We show t
hat a deterministic approach is not sufficient for accurate and robust estimatio
n: however, a probabilistic formulation makes it possible to implicitly integrat
e away hidden factors that have been ignored by the physical model. Experimental
results are reported on a dataset of natural video sequences and on the GrayBal
l benchmark, indicating that we compare favorably with the state-of-the-art.
****

Allocentric Pose Estimation
M. Jose Antonio, Luc De Raedt, Tinne Tuytelaars; Proceedings of the IEEE Interna
tional Conference on Computer Vision (ICCV), 2013, pp. 289-296
The task of object pose estimation has been a challenge since the early days of
computer vision. To estimate the pose (or viewpoint) of an object, people have m
ostly looked at object intrinsic features, such as shape or appearance. Surprisi
ngly, informative features provided by other, external elements in the scene, ha
ve so far mostly been ignored. At the same time, contextual cues have been shown
to be of great benefit for related tasks such as object detection or action rec
ognition. In this paper, we explore how information from other objects in the sc
ene can be exploited for pose estimation. In particular, we look at object confi
gurations. We show that, starting from noisy object detections and pose estimate
s, exploiting the estimated pose and location of other objects in the scene can
help to estimate the objects' poses more accurately. We explore both a camera-ce
ntered as well as an object-centered representation for relations. Experiments o
n the challenging KITTI dataset show that object configurations can indeed be us
ed as a complementary cue to appearance-based pose estimation. In addition, obje
ct-centered relational representations can also assist object detection.
****

The Interestingness of Images
Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, Luc Van Gool
; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 20
13, pp. 1633-1640
We investigate human interest in photos. Based on our own and others' psychologi
cal experiments, we identify various cues for "interestingness", namely aestheti
cs, unusualness and general preferences. For the ranking of retrieved images, in
terestingness is more appropriate than cues proposed earlier. Interestingness is
, for example, correlated with what people believe they will remember. This is o
pposed to actual memorability, which is uncorrelated to both of them. We introdu

ce a set of features computationally capturing the three main aspects of visual interestingness that we propose and build an interestingness predictor from them . Its performance is shown on three datasets with varying context, reflecting di verse levels of prior knowledge of the viewers.
****

Learning Maximum Margin Temporal Warping for Action Recognition
Jiang Wang, Ying Wu; Proceedings of the IEEE International Conference on Compute r Vision (ICCV), 2013, pp. 2688-2695
Temporal misalignment and duration variation in video actions largely influence the performance of action recognition, but it is very difficult to specify effec tive temporal alignment on action sequences. To address this challenge, this pap er proposes a novel discriminative learning-based temporal alignment method, cal led maximum margin temporal warping (MMTW), to align two action sequences and me asure their matching score. Based on the latent structure SVM formulation, the p roposed MMTW method is able to learn a phantom action template to represent an a ction class for maximum discrimination against other classes. The recognition of this action class is based on the associated learned alignment of the input act ion. Extensive experiments on five benchmark datasets have demonstrated that thi s MMTW model is able to significantly promote the accuracy and robustness of act ion recognition under temporal misalignment and variations.
****

Rolling Shutter Stereo
Olivier Saurer, Kevin Koser, Jean-Yves Bouguet, Marc Pollefeys; Proceedings of t he IEEE International Conference on Computer Vision (ICCV), 2013, pp. 465-472
A huge fraction of cameras used nowadays is based on CMOS sensors with a rolling shutter that exposes the image line by line. For dynamic scenes/cameras this in troduces undesired effects like stretch, shear and wobble. It has been shown ear lier that rotational shake induced rolling shutter effects in hand-held cell pho ne capture can be compensated based on an estimate of the camera rotation. In co ntrast, we analyse the case of significant camera motion, e.g. where a bypassing streetlevel capture vehicle uses a rolling shutter camera in a 3D reconstructio n framework. The introduced error is depth dependent and cannot be compensated b ased on camera motion/rotation alone, invalidating also rectification for stereo camera systems. On top, significant lens distortion as often present in wide an gle cameras intertwines with rolling shutter effects as it changes the time at w hich a certain 3D point is seen. We show that naive 3D reconstructions (assuming global shutter) will deliver biased geometry already for very mild assumptions on vehicle speed and resolution. We then develop rolling shutter dense multiview stereo algorithms that solve for time of exposure and depth at the same time, e ven in the presence of lens distortion and perform an evaluation on ground truth laser scan models as well as on real street-level data.
****

Fast Sparsity-Based Orthogonal Dictionary Learning for Image Restoration
Chenglong Bao, Jian-Feng Cai, Hui Ji; Proceedings of the IEEE International Conf erence on Computer Vision (ICCV), 2013, pp. 3384-3391
In recent years, how to learn a dictionary from input images for sparse modellin g has been one very active topic in image processing and recognition. Most exist ing dictionary learning methods consider an over-complete dictionary, e.g. the K -SVD method. Often they require solving some minimization problem that is very c hallenging in terms of computational feasibility and efficiency. However, if the correlations among dictionary atoms are not well constrained, the redundancy of the dictionary does not necessarily improve the performance of sparse coding. T his paper proposed a fast orthogonal dictionary learning method for sparse image representation. With comparable performance on several image restoration tasks, the proposed method is much more computationally efficient than the over-comple te dictionary based learning methods.
****

Slice Sampling Particle Belief Propagation
Oliver Muller, Michael Ying Yang, Bodo Rosenhahn; Proceedings of the IEEE Intern ational Conference on Computer Vision (ICCV), 2013, pp. 1129-1136

Inference in continuous label Markov random fields is a challenging task. We use particle belief propagation (PBP) for solving the inference problem in continuous label space. Sampling particles from the belief distribution is typically done by using Metropolis-Hastings (MH) Markov chain Monte Carlo (MCMC) methods which involves sampling from a proposal distribution. This proposal distribution has to be carefully designed depending on the particular model and input data to achieve fast convergence. We propose to avoid dependence on a proposal distribution by introducing a slice sampling based PBP algorithm. The proposed approach shows superior convergence performance on an image denoising toy example. Our findings are validated on a challenging relational 2D feature tracking application.

****

Training Deformable Part Models with Decorrelated Features

Ross Girshick, Jitendra Malik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3016-3023

In this paper, we show how to train a deformable part model (DPM) fast--typically in less than 20 minutes, or four times faster than the current fastest method--while maintaining high average precision on the PASCAL VOC datasets. At the core of our approach is "latent LDA," a novel generalization of linear discriminant analysis for learning latent variable models. Unlike latent SVM, latent LDA uses efficient closed-form updates and does not require an expensive search for hard negative examples. Our approach also acts as a springboard for a detailed experimental study of DPM training. We isolate and quantify the impact of key training factors for the first time (e.g., How important are discriminative SVM filters? How important is joint parameter estimation? How many negative images are needed for training?). Our findings yield useful insights for researchers working with Markov random fields and partbased models, and have practical implications for speeding up tasks such as model selection.

****

Efficient Salient Region Detection with Soft Image Abstraction

Ming-Ming Cheng, Jonathan Warrell, Wen-Yan Lin, Shuai Zheng, Vibhav Vineet, Nigel Crook; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1529-1536

Detecting visually salient regions in images is one of the fundamental problems in computer vision. We propose a novel method to decompose an image into large scale perceptually homogeneous elements for efficient salient region detection, using a soft image abstraction representation. By considering both appearance similarity and spatial distribution of image pixels, the proposed representation abstracts out unnecessary image details, allowing the assignment of comparable saliency values across similar regions, and producing perceptually accurate salient region detection. We evaluate our salient region detection approach on the largest publicly available dataset with pixel accurate annotations. The experimental results show that the proposed method outperforms 18 alternate methods, reducing the mean absolute error by 25.2% compared to the previous best result, while being computationally more efficient.

****

Video Segmentation by Tracking Many Figure-Ground Segments

Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, James M. Rehg; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2192-2199

We propose an unsupervised video segmentation approach by simultaneously tracking multiple holistic figureground segments. Segment tracks are initialized from a pool of segment proposals generated from a figure-ground segmentation algorithm. Then, online non-local appearance models are trained incrementally for each track using a multi-output regularized least squares formulation. By using the same set of training examples for all segment tracks, a computational trick allows us to track hundreds of segment tracks efficiently, as well as perform optimal online updates in closed-form. Besides, a new composite statistical inference approach is proposed for refining the obtained segment tracks, which breaks down the initial segment proposals and recombines for better ones by utilizing highorde

r statistic estimates from the appearance model and enforcing temporal consisten
cy. For evaluating the algorithm, a dataset, SegTrack v2, is collected with abou
t 1,000 frames with pixel-level annotations. The proposed framework outperforms
state-of-the-art approaches in the dataset, showing its efficiency and robustnes
s to challenges in different video sequences.
****

Bayesian 3D Tracking from Monocular Video
Ernesto Brau, Jinyan Guan, Kyle Simek, Luca Del Pero, Colin Reimer Dawson, Kobus
 Barnard; Proceedings of the IEEE International Conference on Computer Vision (I
CCV), 2013, pp. 3368-3375
We develop a Bayesian modeling approach for tracking people in 3D from monocular
 video with unknown cameras. Modeling in 3D provides natural explanations for oc
clusions and smoothness discontinuities that result from projection, and allows
priors on velocity and smoothness to be grounded in physical quantities: meters
and seconds vs. pixels and frames. We pose the problem in the context of data as
sociation, in which observations are assigned to tracks. A correct application o
f Bayesian inference to multitarget tracking must address the fact that the mode
l's dimension changes as tracks are added or removed, and thus, posterior densit
ies of different hypotheses are not comparable. We address this by marginalizing
 out the trajectory parameters so the resulting posterior over data associations
 has constant dimension. This is made tractable by using (a) Gaussian process pr
iors for smooth trajectories and (b) approximately Gaussian likelihood functions
. Our approach provides a principled method for incorporating multiple sources o
f evidence; we present results using both optical flow and object detector outpu
ts. Results are comparable to recent work on 3D tracking and, unlike others, our
 method requires no pre-calibrated cameras.
****

Concurrent Action Detection with Structural Prediction
Ping Wei, Nanning Zheng, Yibiao Zhao, Song-Chun Zhu; Proceedings of the IEEE Int
ernational Conference on Computer Vision (ICCV), 2013, pp. 3136-3143
Action recognition has often been posed as a classification problem, which assum
es that a video sequence only have one action class label and different actions
are independent. However, a single human body can perform multiple concurrent ac
tions at the same time, and different actions interact with each other. This pap
er proposes a concurrent action detection model where the action detection is fo
rmulated as a structural prediction problem. In this model, an interval in a vid
eo sequence can be described by multiple action labels. An detected action inter
val is determined both by the unary local detector and the relations with other
actions. We use a wavelet feature to represent the action sequence, and design a
 composite temporal logic descriptor to describe the action relations. The model
 parameters are trained by structural SVM learning. Given a long video sequence,
 a sequential decision window search algorithm is designed to detect the actions
. Experiments on our new collected concurrent action dataset demonstrate the str
ength of our method.
****

Discriminatively Trained Templates for 3D Object Detection: A Real Time Scalable
 Approach
Reyes Rios-Cabrera, Tinne Tuytelaars; Proceedings of the IEEE International Conf
erence on Computer Vision (ICCV), 2013, pp. 2048-2055
In this paper we propose a new method for detecting multiple specific 3D objects
 in real time. We start from the template-based approach based on the LINE2D/LIN
EMOD representation introduced recently by Hinterstoisser et al., yet extend it
in two ways. First, we propose to learn the templates in a discriminative fashio
n. We show that this can be done online during the collection of the example ima
ges, in just a few milliseconds, and has a big impact on the accuracy of the det
ector. Second, we propose a scheme based on cascades that speeds up detection. S
ince detection of an object is fast, new objects can be added with very low cost
, making our approach scale well. In our experiments, we easily handle 10-30 3D
objects at frame rates above 10fps using a single CPU core. We outperform the st
ate-of-the-art both in terms of speed as well as in terms of accuracy, as valida

ted on 3 different datasets. This holds both when using monocular color images (with LINE2D) and when using RGBD images (with LINEMOD). Moreover, we propose a c hallenging new dataset made of 12 objects, for future competing methods on monoc ular color images.
****

The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Re cognition and Detection
Mihai Zanfir, Marius Leordeanu, Cristian Sminchisescu; Proceedings of the IEEE I nternational Conference on Computer Vision (ICCV), 2013, pp. 2752-2759
Human action recognition under low observational latency is receiving a growing interest in computer vision due to rapidly developing technologies in human-robo t interaction, computer gaming and surveillance. In this paper we propose a fast , simple, yet powerful non-parametric Moving Pose (MP) framework for low-latency human action and activity recognition. Central to our methodology is a moving p ose descriptor that considers both pose information as well as differential quan tities (speed and acceleration) of the human body joints within a short time win dow around the current frame. The proposed descriptor is used in conjunction wit h a modified kNN classifier that considers both the temporal location of a parti cular frame within the action sequence as well as the discrimination power of it s moving pose descriptor compared to other frames in the training set. The resul ting method is non-parametric and enables low-latency recognition, one-shot lear ning, and action detection in difficult unsegmented sequences. Moreover, the fra mework is real-time, scalable, and outperforms more sophisticated approaches on challenging benchmarks like MSR-Action3D or MSR-DailyActivities3D.
****

Learning a Dictionary of Shape Epitomes with Applications to Image Labeling
Liang-Chieh Chen, George Papandreou, Alan L. Yuille; Proceedings of the IEEE Int ernational Conference on Computer Vision (ICCV), 2013, pp. 337-344
The first main contribution of this paper is a novel method for representing ima ges based on a dictionary of shape epitomes. These shape epitomes represent the local edge structure of the image and include hidden variables to encode shift a nd rotations. They are learnt in an unsupervised manner from groundtruth edges. This dictionary is compact but is also able to capture the typical shapes of edg es in natural images. In this paper, we illustrate the shape epitomes by applyin g them to the image labeling task. In other work, described in the supplementary material, we apply them to edge detection and image modeling. We apply shape ep itomes to image labeling by using Conditional Random Field (CRF) Models. They ar e alternatives to the superpixel or pixel representations used in most CRFs. In our approach, the shape of an image patch is encoded by a shape epitome from the dictionary. Unlike the superpixel representation, our method avoids making earl y decisions which cannot be reversed. Our resulting hierarchical CRFs efficientl y capture both local and global class co-occurrence properties. We demonstrate i ts quantitative and qualitative properties of our approach with image labeling e xperiments on two standard datasets: MSRC-21 and Stanford Background.
****

Online Motion Segmentation Using Dynamic Label Propagation
Ali Elqursh, Ahmed Elgammal; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2008-2015
The vast majority of work on motion segmentation adopts the affine camera model due to its simplicity. Under the affine model, the motion segmentation problem b ecomes that of subspace separation. Due to this assumption, such methods are mai nly offline and exhibit poor performance when the assumption is not satisfied. T his is made evident in state-of-the-art methods that relax this assumption by us ing piecewise affine spaces and spectral clustering techniques to achieve better results. In this paper, we formulate the problem of motion segmentation as that of manifold separation. We then show how label propagation can be used in an on line framework to achieve manifold separation. The performance of our framework is evaluated on a benchmark dataset and achieves competitive performance while b eing online.
****

Sequential Bayesian Model Update under Structured Scene Prior for Semantic Road Scenes Labeling
Evgeny Levinkov, Mario Fritz; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1321-1328
Semantic road labeling is a key component of systems that aim at assisted or even autonomous driving. Considering that such systems continuously operate in the realworld, unforeseen conditions not represented in any conceivable training procedure are likely to occur on a regular basis. In order to equip systems with the ability to cope with such situations, we would like to enable adaptation to such new situations and conditions at runtime. Existing adaptive methods for image labeling either require labeled data from the new condition or even operate globally on a complete test set. None of this is a desirable mode of operation for a system as described above where new images arrive sequentially and conditions may vary. We study the effect of changing test conditions on scene labeling methods based on a new diverse street scene dataset. We propose a novel approach that can operate in such conditions and is based on a sequential Bayesian model update in order to robustly integrate the arriving images into the adapting procedure.
****
Directed Acyclic Graph Kernels for Action Recognition
Ling Wang, Hichem Sahbi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3168-3175
One of the trends of action recognition consists in extracting and comparing mid-level features which encode visual and motion aspects of objects into scenes. However, when scenes contain high-level semantic actions with many interacting parts, these mid-level features are not sufficient to capture high level structures as well as high order causal relationships between moving objects resulting in to a clear drop in performances. In this paper, we address this issue and we propose an alternative action recognition method based on a novel graph kernel. In the main contributions of this work, we first describe actions in videos using directed acyclic graphs (DAGs), that naturally encode pairwise interactions between moving object parts, and then we compare these DAGs by analyzing the spectrum of their sub-patterns that capture complex higher order interactions. This extraction and comparison process is computationally tractable, resulting from the acyclic property of DAGs, and it also defines a positive semi-definite kernel. When plugging the latter into support vector machines, we obtain an action recognition algorithm that overtakes related work, including graph-based methods, on a standard evaluation dataset.
****
Strong Appearance and Expressive Spatial Models for Human Pose Estimation
Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, Bernt Schiele; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3487-3494
Typical approaches to articulated pose estimation combine spatial modelling of the human body with appearance modelling of body parts. This paper aims to push the state-of-the-art in articulated pose estimation in two ways. First we explore various types of appearance representations aiming to substantially improve the body part hypotheses. And second, we draw on and combine several recently proposed powerful ideas such as more flexible spatial models as well as image-conditioned spatial models. In a series of experiments we draw several important conclusions: (1) we show that the proposed appearance representations are complementary; (2) we demonstrate that even a basic tree-structure spatial human body model achieves state-ofthe-art performance when augmented with the proper appearance representation; and (3) we show that the combination of the best performing appearance model with a flexible image-conditioned spatial model achieves the best result, significantly improving over the state of the art, on the "Leeds Sports Poses" and "Parse" benchmarks.
****
Revisiting Example Dependent Cost-Sensitive Learning with Decision Trees
Oisin Mac Aodha, Gabriel J. Brostow; Proceedings of the IEEE International Confe

rence on Computer Vision (ICCV), 2013, pp. 193-200
Typical approaches to classification treat class labels as disjoint. For each tr
aining example, it is assumed that there is only one class label that correctly
describes it, and that all other labels are equally bad. We know however, that g
ood and bad labels are too simplistic in many scenarios, hurting accuracy. In th
e realm of example dependent costsensitive learning, each label is instead a vec
tor representing a data point's affinity for each of the classes. At test time,
our goal is not to minimize the misclassification rate, but to maximize that aff
inity. We propose a novel example dependent cost-sensitive impurity measure for
decision trees. Our experiments show that this new impurity measure improves tes
t performance while still retaining the fast test times of standard classificati
on trees. We compare our approach to classification trees and other cost-sensiti
ve methods on three computer vision problems, tracking, descriptor matching, and
optical flow, and show improvements in all three domains.
****

Matching Dry to Wet Materials
Yaser Yacoob; Proceedings of the IEEE International Conference on Computer Visio
n (ICCV), 2013, pp. 2952-2959
When a translucent liquid is spilled over a rough surface it causes a significan
t change in the visual appearance of the surface. This wetting phenomenon is eas
ily detected by humans, and an early model was devised by the physicist Andres J
onas Angstrom nearly a century ago. In this paper we investigate the problem of
determining if a wet/dry relationship between two image patches explains the dif
ferences in their visual appearance. Water tends to be the typical liquid involv
ed and therefore it is the main objective. At the same time, we consider the gen
eral problem where the liquid has some of the characteristics of water (i.e., a
similar refractive index), but has an unknown spectral absorption profile (e.g.,
coffee, tea, wine, etc.). We report on several experiments using our own images
, a publicly available dataset, and images downloaded from the web.
****

On the Mean Curvature Flow on Graphs with Applications in Image and Manifold Pro
cessing
Abdallah El Chakik, Abderrahim Elmoataz, Ahcene Sadi; Proceedings of the IEEE In
ternational Conference on Computer Vision (ICCV), 2013, pp. 697-704
In this paper, we propose an adaptation and transcription of the mean curvature
level set equation on a general discrete domain (weighted graphs with arbitrary
topology). We introduce the perimeters on graph using difference operators and d
efine the curvature as the first variation of these perimeters. Our proposed app
roach of mean curvature unifies both local and non local notions of mean curvatu
re on Euclidean domains. Furthermore, it allows the extension to the processing
of manifolds and data which can be represented by graphs.
****

Example-Based Facade Texture Synthesis
Dengxin Dai, Hayko Riemenschneider, Gerhard Schmitt, Luc Van Gool; Proceedings o
f the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1065-10
72
There is an increased interest in the efficient creation of city models, be it v
irtual or as-built. We present a method for synthesizing complex, photo-realisti
c facade images, from a single example. After parsing the example image into its
semantic components, a tiling for it is generated. Novel tilings can then be cr
eated, yielding facade textures with different dimensions or with occluded parts
inpainted. A genetic algorithm guides the novel facades as well as inpainted pa
rts to be consistent with the example, both in terms of their overall structure
and their detailed textures. Promising results for multiple standard datasets in
particular for the different building styles they contain demonstrate the poten
tial of the method.
****

SYM-FISH: A Symmetry-Aware Flip Invariant Sketch Histogram Shape Descriptor
Xiaochun Cao, Hua Zhang, Si Liu, Xiaojie Guo, Liang Lin; Proceedings of the IEEE
International Conference on Computer Vision (ICCV), 2013, pp. 313-320

Recently, studies on sketch, such as sketch retrieval and sketch classification, have received more attention in the computer vision community. One of its most fundamental and essential problems is how to more effectively describe a sketch image. Many existing descriptors, such as shape context, have achieved great success. In this paper, we propose a new descriptor, namely Symmetric-aware Flip Invariant Sketch Histogram (SYM-FISH) to refine the shape context feature. Its extraction process includes three steps. First the Flip Invariant Sketch Histogram (FISH) descriptor is extracted on the input image, which is a flip-invariant version of the shape context feature. Then we explore the symmetry character of the image by calculating the kurtosis coefficient. Finally, the SYM-FISH is generated by constructing a symmetry table. The new SYM-FISH descriptor supplements the original shape context by encoding the symmetric information, which is a pervasive characteristic of natural scene and objects. We evaluate the efficacy of the novel descriptor in two applications, i.e., sketch retrieval and sketch classification. Extensive experiments on three datasets well demonstrate the effectiveness and robustness of the proposed SYM-FISH descriptor.

****

Robust Feature Set Matching for Partial Face Recognition

Renliang Weng, Jiwen Lu, Junlin Hu, Gao Yang, Yap-Peng Tan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 601-608

Over the past two decades, a number of face recognition methods have been proposed in the literature. Most of them use holistic face images to recognize people. However, human faces are easily occluded by other objects in many real-world scenarios and we have to recognize the person of interest from his/her partial faces. In this paper, we propose a new partial face recognition approach by using feature set matching, which is able to align partial face patches to holistic gallery faces automatically and is robust to occlusions and illumination changes. Given each gallery image and probe face patch, we first detect keypoints and extract their local features. Then, we propose a Metric Learned Extended Robust Point Matching (MLERPM) method to discriminatively match local feature sets of a pair of gallery and probe samples. Lastly, the similarity of two faces is converted as the distance between two feature sets. Experimental results on three public face databases are presented to show the effectiveness of the proposed approach.

****

Cross-View Action Recognition over Heterogeneous Feature Spaces

Xinxiao Wu, Han Wang, Cuiwei Liu, Yunde Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 609-616

In cross-view action recognition, "what you saw" in one view is different from "what you recognize" in another view. The data distribution even the feature space can change from one view to another due to the appearance and motion of actions drastically vary across different views. In this paper, we address the problem of transferring action models learned in one view (source view) to another different view (target view), where action instances from these two views are represented by heterogeneous features. A novel learning method, called Heterogeneous Transfer Discriminantanalysis of Canonical Correlations (HTDCC), is proposed to learn a discriminative common feature space for linking source and target views to transfer knowledge between them. Two projection matrices that respectively map data from source and target views into the common space are optimized via simultaneously minimizing the canonical correlations of inter-class samples and maximizing the intraclass canonical correlations. Our model is neither restricted to corresponding action instances in the two views nor restricted to the same type of feature, and can handle only a few or even no labeled samples available in the target view. To reduce the data distribution mismatch between the source and target views in the common feature space, a nonparametric criterion is included in the objective function. We additionally propose a joint weight learning method to fuse multiple source-view action classifiers for recognition in the target view. Different combination weights are assigned to different source views, with each weight presenting how contributive the corresponding source view is to the target view. The proposed method is evaluated on the IXMAS multi-view dataset and achieves promising results.

****
Building Part-Based Object Detectors via 3D Geometry
Abhinav Shrivastava, Abhinav Gupta; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1745-1752
This paper proposes a novel part-based representation for modeling object categories. Our representation combines the effectiveness of deformable part-based models with the richness of geometric representation by defining parts based on consistent underlying 3D geometry. Our key hypothesis is that while the appearance and the arrangement of parts might vary across the instances of object categories, the constituent parts will still have consistent underlying 3D geometry. We propose to learn this geometrydriven deformable part-based model (gDPM) from a set of labeled RGBD images. We also demonstrate how the geometric representation of gDPM can help us leverage depth data during training and constrain the latent model learning problem. But most importantly, a joint geometric and appearance based representation not only allows us to achieve state-of-the-art results on object detection but also allows us to tackle the grand challenge of understanding 3D objects from 2D images.
****
Active Visual Recognition with Expertise Estimation in Crowdsourcing
Chengjiang Long, Gang Hua, Ashish Kapoor; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3000-3007
We present a noise resilient probabilistic model for active learning of a Gaussian process classifier from crowds, i.e., a set of noisy labelers. It explicitly models both the overall label noises and the expertise level of each individual labeler in two levels of flip models. Expectation propagation is adopted for efficient approximate Bayesian inference of our probabilistic model for classification, based on which, a generalized EM algorithm is derived to estimate both the global label noise and the expertise of each individual labeler. The probabilistic nature of our model immediately allows the adoption of the prediction entropy and estimated expertise for active selection of data sample to be labeled, and active selection of high quality labelers to label the data, respectively. We apply the proposed model for three visual recognition tasks, i.e, object category recognition, gender recognition, and multi-modal activity recognition, on three datasets with real crowd-sourced labels from Amazon Mechanical Turk. The experiments clearly demonstrated the efficacy of the proposed model.
****
Attribute Pivots for Guiding Relevance Feedback in Image Search
Adriana Kovashka, Kristen Grauman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 297-304
In interactive image search, a user iteratively refines his results by giving feedback on exemplar images. Active selection methods aim to elicit useful feedback, but traditional approaches suffer from expensive selection criteria and cannot predict informativeness reliably due to the imprecision of relevance feedback. To address these drawbacks, we propose to actively select "pivot" exemplars for which feedback in the form of a visual comparison will most reduce the system's uncertainty. For example, the system might ask, "Is your target image more or less crowded than this image?" Our approach relies on a series of binary search trees in relative attribute space, together with a selection function that predicts the information gain were the user to compare his envisioned target to the next node deeper in a given attribute's tree. It makes interactive search more efficient than existing strategies--both in terms of the system's selection time as well as the user's feedback effort.
****
Initialization-Insensitive Visual Tracking through Voting with Salient Local Features
Kwang Moo Yi, Hawook Jeong, Byeongho Heo, Hyung Jin Chang, Jin Young Choi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2912-2919
In this paper we propose an object tracking method in case of inaccurate initializations. To track objects accurately in such situation, the proposed method use

s "motion saliency" and "descriptor saliency" of local features and performs tra cking based on generalized Hough transform (GHT). The proposed motion saliency o f a local feature emphasizes features having distinctive motions, compared to th e motions which are not from the target object. The descriptor saliency emphasiz es features which are likely to be of the object in terms of its feature descrip tors. Through these saliencies, the proposed method tries to "learn and find" th e target object rather than looking for what was given at initialization, giving robust results even with inaccurate initializations. Also, our tracking result is obtained by combining the results of each local feature of the target and the surroundings with GHT voting, thus is robust against severe occlusions as well. The proposed method is compared against nine other methods, with nine image seq uences, and hundred random initializations. The experimental results show that o ur method outperforms all other compared methods.
****

Refractive Structure-from-Motion on Underwater Images
Anne Jordt-Sedlazeck, Reinhard Koch; Proceedings of the IEEE International Confe rence on Computer Vision (ICCV), 2013, pp. 57-64
In underwater environments, cameras need to be confined in an underwater housing , viewing the scene through a piece of glass. In case of flat port underwater ho usings, light rays entering the camera housing are refracted twice, due to diffe rent medium densities of water, glass, and air. This causes the usually linear r ays of light to bend and the commonly used pinhole camera model to be invalid. W hen using the pinhole camera model without explicitly modeling refraction in Str ucture-from-Motion (SfM) methods, a systematic model error occurs. Therefore, in this paper, we propose a system for computing camera path and 3D points with ex plicit incorporation of refraction using new methods for pose estimation. Additi onally, a new error function is introduced for non-linear optimization, especial ly bundle adjustment. The proposed method allows to increase reconstruction accu racy and is evaluated in a set of experiments, where the proposed method's perfo rmance is compared to SfM with the perspective camera model.
****

Semi-dense Visual Odometry for a Monocular Camera
Jakob Engel, Jurgen Sturm, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1449-1456
We propose a fundamentally novel approach to real-time visual odometry for a mon ocular camera. It allows to benefit from the simplicity and accuracy of dense tr acking which does not depend on visual features while running in real-time on a CPU. The key idea is to continuously estimate a semi-dense inverse depth map for the current frame, which in turn is used to track the motion of the camera usin g dense image alignment. More specifically, we estimate the depth of all pixels which have a non-negligible image gradient. Each estimate is represented as a Ga ussian probability distribution over the inverse depth. We propagate this inform ation over time, and update it with new measurements as new images arrive. In te rms of tracking accuracy and computational speed, the proposed method compares f avorably to both state-of-the-art dense and feature-based visual odometry and SL AM algorithms. As our method runs in real-time on a CPU, it is of large practica l value for robotics and augmented reality applications.
****

Characterizing Layouts of Outdoor Scenes Using Spatial Topic Processes
Dahua Lin, Jianxiong Xiao; Proceedings of the IEEE International Conference on C omputer Vision (ICCV), 2013, pp. 841-848
In this paper, we develop a generative model to describe the layouts of outdoor scenes the spatial configuration of regions. Specifically, the layout of an imag e is represented as a composite of regions, each associated with a semantic topi c. At the heart of this model is a novel stochastic process called Spatial Topic Process, which generates a spatial map of topics from a set of coupled Gaussian processes, thus allowing the distributions of topics to vary continuously acros s the image plane. A key aspect that distinguishes this model from previous ones consists in its capability of capturing dependencies across both locations and topics while allowing substantial variations in the layouts. We demonstrate the

practical utility of the proposed model by testing it on scene classification, s
emantic segmentation, and layout hallucination.
****

A Deformable Mixture Parsing Model with Parselets
Jian Dong, Qiang Chen, Wei Xia, Zhongyang Huang, Shuicheng Yan; Proceedings of t
he IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3408-3415
In this work, we address the problem of human parsing, namely partitioning the h
uman body into semantic regions, by using the novel Parselet representation. Pre
vious works often consider solving the problem of human pose estimation as the p
rerequisite of human parsing. We argue that these approaches cannot obtain optim
al pixel level parsing due to the inconsistent targets between these tasks. In t
his paper, we propose to use Parselets as the building blocks of our parsing mod
el. Parselets are a group of parsable segments which can generally be obtained b
y lowlevel over-segmentation algorithms and bear strong semantic meaning. We the
n build a Deformable Mixture Parsing Model (DMPM) for human parsing to simultane
ously handle the deformation and multi-modalities of Parselets. The proposed mod
el has two unique characteristics: (1) the possible numerous modalities of Parse
let ensembles are exhibited as the "And-Or" structure of sub-trees; (2) to furth
er solve the practical problem of Parselet occlusion or absence, we directly mod
el the visibility property at some leaf nodes. The DMPM thus directly solves the
 problem of human parsing by searching for the best graph configuration from a p
ool of Parselet hypotheses without intermediate tasks. Comprehensive evaluations
 demonstrate the encouraging performance of the proposed approach.
****

Dictionary Learning and Sparse Coding on Grassmann Manifolds: An Extrinsic Solut
ion
Mehrtash Harandi, Conrad Sanderson, Chunhua Shen, Brian C. Lovell; Proceedings o
f the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3120-31
27
Recent advances in computer vision and machine learning suggest that a wide rang
e of problems can be addressed more appropriately by considering non-Euclidean g
eometry. In this paper we explore sparse dictionary learning over the space of l
inear subspaces, which form Riemannian structures known as Grassmann manifolds.
To this end, we propose to embed Grassmann manifolds into the space of symmetric
 matrices by an isometric mapping, which enables us to devise a closed-form solu
tion for updating a Grassmann dictionary, atom by atom. Furthermore, to handle n
on-linearity in data, we propose a kernelised version of the dictionary learning
 algorithm. Experiments on several classification tasks (face recognition, actio
n recognition, dynamic texture classification) show that the proposed approach a
chieves considerable improvements in discrimination accuracy, in comparison to s
tate-of-the-art methods such as kernelised Affine Hull Method and graphembedding
 Grassmann discriminant analysis.
****

Real-Time Articulated Hand Pose Estimation Using Semi-supervised Transductive Re
gression Forests
Danhang Tang, Tsz-Ho Yu, Tae-Kyun Kim; Proceedings of the IEEE International Con
ference on Computer Vision (ICCV), 2013, pp. 3224-3231
This paper presents the first semi-supervised transductive algorithm for real-ti
me articulated hand pose estimation. Noisy data and occlusions are the major cha
llenges of articulated hand pose estimation. In addition, the discrepancies amon
g realistic and synthetic pose data undermine the performances of existing appro
aches that use synthetic data extensively in training. We therefore propose the
Semi-supervised Transductive Regression (STR) forest which learns the relationsh
ip between a small, sparsely labelled realistic dataset and a large synthetic da
taset. We also design a novel data-driven, pseudo-kinematic technique to refine
noisy or occluded joints. Our contributions include: (i) capturing the benefits
of both realistic and synthetic data via transductive learning; (ii) showing acc
uracies can be improved by considering unlabelled data; and (iii) introducing a
pseudo-kinematic technique to refine articulations efficiently. Experimental res
ults show not only the promising performance of our method with respect to noise

and occlusions, but also its superiority over state-ofthe-arts in accuracy, robustness and speed.
****

Face Recognition Using Face Patch Networks
Chaochao Lu, Deli Zhao, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3288-3295
When face images are taken in the wild, the large variations in facial pose, illumination, and expression make face recognition challenging. The most fundamental problem for face recognition is to measure the similarity between faces. The traditional measurements such as various mathematical norms, Hausdorff distance, and approximate geodesic distance cannot accurately capture the structural information between faces in such complex circumstances. To address this issue, we develop a novel face patch network, based on which we define a new similarity measure called the random path (RP) measure. The RP measure is derived from the collective similarity of paths by performing random walks in the network. It can globally characterize the contextual and curved structures of the face space. To apply the RP measure, we construct two kinds of networks: the in-face network and the out-face network. The in-face network is drawn from any two face images and captures the local structural information. The out-face network is constructed from all the training face patches, thereby modeling the global structures of face space. The two face networks are structurally complementary and can be combined together to improve the recognition performance. Experiments on the Multi-PIE and LFW benchmarks show that the RP measure outperforms most of the state-of-art algorithms for face recognition.
****

Depth from Combining Defocus and Correspondence Using Light-Field Cameras
Michael W. Tao, Sunil Hadap, Jitendra Malik, Ravi Ramamoorthi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 673-680
Light-field cameras have recently become available to the consumer market. An array of micro-lenses captures enough information that one can refocus images after acquisition, as well as shift one's viewpoint within the subapertures of the main lens, effectively obtaining multiple views. Thus, depth cues from both defocus and correspondence are available simultaneously in a single capture. Previously, defocus could be achieved only through multiple image exposures focused at different depths, while correspondence cues needed multiple exposures at different viewpoints or multiple cameras; moreover, both cues could not easily be obtained together. In this paper, we present a novel simple and principled algorithm that computes dense depth estimation by combining both defocus and correspondence depth cues. We analyze the x-u 2D epipolar image (EPI), where by convention we assume the spatial lrcoordinate is horizontal and the angular umcoordinate is vertical (our final algorithm uses the full 4D EPI). We show that defocus depth cues are obtained by computing the horizontal (spatial) variance after vertical (angular) integration, and correspondence depth cues by computing the vertical (angular) variance. We then show how to combine the two cues into a high quality depth map, suitable for computer vision applications such as matting, full control of depth-of-field, and surface reconstruction.
****

Minimal Basis Facility Location for Subspace Segmentation
Choon-Meng Lee, Loong-Fah Cheong; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1585-1592
In contrast to the current motion segmentation paradigm that assumes independence between the motion subspaces, we approach the motion segmentation problem by seeking the parsimonious basis set that can represent the data. Our formulation explicitly looks for the overlap between subspaces in order to achieve a minimal basis representation. This parsimonious basis set is important for the performance of our model selection scheme because the sharing of basis results in savings of model complexity cost. We propose the use of affinity propagation based method to determine the number of motion. The key lies in the incorporation of a global cost model into the factor graph, serving
****

Unsupervised Random Forest Manifold Alignment for Lipreading
Yuru Pei, Tae-Kyun Kim, Hongbin Zha; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 129-136
Lipreading from visual channels remains a challenging topic considering the various speaking characteristics. In this paper, we address an efficient lipreading approach by investigating the unsupervised random forest manifold alignment (RFMA). The density random forest is employed to estimate affinity of patch trajectories in speaking facial videos. We propose novel criteria for node splitting to avoid the rank-deficiency in learning density forests. By virtue of the hierarchical structure of random forests, the trajectory affinities are measured efficiently, which are used to find embeddings of the speaking video clips by a graph-based algorithm. Lipreading is formulated as matching between manifolds of query and reference video clips. We employ the manifold alignment technique for matching, where the L ? norm-based manifold-to-manifold distance is proposed to find the matching pairs. We apply this random forest manifold alignment technique to various video data sets captured by consumer cameras. The experiments demonstrate that lipreading can be performed effectively, and outperform state-of-the-arts.
****

Visual Reranking through Weakly Supervised Multi-graph Learning
Cheng Deng, Rongrong Ji, Wei Liu, Dacheng Tao, Xinbo Gao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2600-2607
Visual reranking has been widely deployed to refine the quality of conventional content-based image retrieval engines. The current trend lies in employing a crowd of retrieved results stemming from multiple feature modalities to boost the overall performance of visual reranking. However, a major challenge pertaining to current reranking methods is how to take full advantage of the complementary property of distinct feature modalities. Given a query image and one feature modality, a regular visual reranking framework treats the top-ranked images as pseudo positive instances which are inevitably noisy, difficult to reveal this complementary property, and thus lead to inferior ranking performance. This paper proposes a novel image reranking approach by introducing a Co-Regularized Multi-Graph Learning (Co-RMGL) framework, in which the intra-graph and inter-graph constraints are simultaneously imposed to encode affinities in a single graph and consistency across different graphs. Moreover, weakly supervised learning driven by image attributes is performed to denoise the pseudolabeled instances, thereby highlighting the unique strength of individual feature modality. Meanwhile, such learning can yield a few anchors in graphs that vitally enable the alignment and fusion of multiple graphs. As a result, an edge weight matrix learned from the fused graph automatically gives the ordering to the initially retrieved results. We evaluate our approach on four benchmark image retrieval datasets, demonstrating a significant performance gain over the state-of-the-arts.
****

Volumetric Semantic Segmentation Using Pyramid Context Features
Jonathan T. Barron, Mark D. Biggin, Pablo Arbelaez, David W. Knowles, Soile V.E. Keranen, Jitendra Malik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3448-3455
We present an algorithm for the per-voxel semantic segmentation of a three-dimensional volume. At the core of our algorithm is a novel "pyramid context" feature, a descriptive representation designed such that exact per-voxel linear classification can be made extremely efficient. This feature not only allows for efficient semantic segmentation but enables other aspects of our algorithm, such as novel learned features and a stacked architecture that can reason about self-consistency. We demonstrate our technique on 3D fluorescence microscopy data of Drosophila embryos for which we are able to produce extremely accurate semantic segmentations in a matter of minutes, and for which other algorithms fail due to the size and high-dimensionality of the data, or due to the difficulty of the task.
****

Transfer Feature Learning with Joint Distribution Adaptation
Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, Philip S. Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2

200-2207
Transfer learning is established as an effective technology in computer vision for leveraging rich labeled data in the source domain to build an accurate classifier for the target domain. However, most prior methods have not simultaneously reduced the difference in both the marginal distribution and conditional distribution between domains. In this paper, we put forward a novel transfer learning approach, referred to as Joint Distribution Adaptation (JDA). Specifically, JDA aims to jointly adapt both the marginal distribution and conditional distribution in a principled dimensionality reduction procedure, and construct new feature representation that is effective and robust for substantial distribution difference. Extensive experiments verify that JDA can significantly outperform several state-of-the-art methods on four types of cross-domain image classification problems.
****

A Novel Earth Mover's Distance Methodology for Image Matching with Gaussian Mixture Models
Peihua Li, Qilong Wang, Lei Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1689-1696
The similarity or distance measure between Gaussian mixture models (GMMs) plays a crucial role in contentbased image matching. Though the Earth Mover's Distance (EMD) has shown its advantages in matching histogram features, its potentials in matching GMMs remain unclear and are not fully explored. To address this problem, we propose a novel EMD methodology for GMM matching. We srst present a sparse representation based EMD called SR-EMD by exploiting the sparse property of the underlying problem. SR-EMD is more efscient and robust than the conventional EMD. Second, we present two novel ground distances between component Gaussians based on the information geometry. The perspective from the Riemannian geometry distinguishes the proposed ground distances from the classical entropyor divergence-based ones. Furthermore, motivated by the success of distance metric learning of vector data, we make the srst attempt to learn the EMD distance metrics between GMMs by using a simple yet effective supervised pair-wise based method. It can adapt the distance metrics between GMMs to specisc classiscation tasks. The proposed method is evaluated on both simulated data and benchmark real databases and achieves very promising performance.
****

Proportion Priors for Image Sequence Segmentation
Claudia Nieuwenhuis, Evgeny Strekalovskiy, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2328-2335
We propose a convex multilabel framework for image sequence segmentation which allows to impose proportion priors on object parts in order to preserve their size ratios across multiple images. The key idea is that for strongly deformable objects such as a gymnast the size ratio of respective regions (head versus torso, legs versus full body, etc.) is typically preserved. We propose different ways to impose such priors in a Bayesian framework for image segmentation. We show that near-optimal solutions can be computed using convex relaxation techniques. Extensive qualitative and quantitative evaluations demonstrate that the proportion priors allow for highly accurate segmentations, avoiding seeping-out of regions and preserving semantically relevant small-scale structures such as hands or feet. They naturally apply to multiple object instances such as players in sports scenes, and they can relate different objects instead of object parts, e.g. organs in medical imaging. The algorithm is efficient and easily parallelized leading to proportion-consistent segmentations at runtimes around one second.
****

Global Fusion of Relative Motions for Robust, Accurate and Scalable Structure from Motion
Pierre Moulon, Pascal Monasse, Renaud Marlet; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3248-3255
Multi-view structure from motion (SfM) estimates the position and orientation of pictures in a common 3D coordinate frame. When views are treated incrementally, this external calibration can be subject to drift, contrary to global methods t

hat distribute residual errors evenly. We propose a new global calibration appro ach based on the fusion of relative motions between image pairs. We improve an e xisting method for robustly computing global rotations. We present an efficient a contrario trifocal tensor estimation method, from which stable and precise tra nslation directions can be extracted. We also define an efficient translation re gistration method that recovers accurate camera positions. These components are combined into an original SfM pipeline. Our experiments show that, on most datas ets, it outperforms in accuracy other existing incremental and global pipelines. It also achieves strikingly good running times: it is about 20 times faster tha n the other global method we could compare to, and as fast as the best increment al method. More importantly, it features better scalability properties.
****

Complex 3D General Object Reconstruction from Line Drawings
Linjie Yang, Jianzhuang Liu, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1433-1440
An important topic in computer vision is 3D object reconstruction from line draw ings. Previous algorithms either deal with simple general objects or are limited to only manifolds (a subset of solids). In this paper, we propose a novel appro ach to 3D reconstruction of complex general objects, including manifolds, non-ma nifold solids, and non-solids. Through developing some 3D object properties, we use the degree of freedom of objects to decompose a complex line drawing into mu ltiple simpler line drawings that represent meaningful building blocks of a comp lex object. After 3D objects are reconstructed from the decomposed line drawings , they are merged to form a complex object from their touching faces, edges, and vertices. Our experiments show a number of reconstruction examples from both co mplex line drawings and images with line drawings superimposed. Comparisons are also given to indicate that our algorithm can deal with much more complex line d rawings of general objects than previous algorithms.
****

From Large Scale Image Categorization to Entry-Level Categories
Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, Tamara L. Berg; Procee dings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2768-2775
Entry level categories the labels people will use to name an object were origina lly defined and studied by psychologists in the 1980s. In this paper we study en trylevel categories at a large scale and learn the first models for predicting e ntry-level categories for images. Our models combine visual recognition predicti ons with proxies for word "naturalness" mined from the enormous amounts of text on the web. We demonstrate the usefulness of our models for predicting nouns (en try-level words) associated with images by people. We also learn mappings betwee n concepts predicted by existing visual recognition systems and entry-level conc epts that could be useful for improving human-focused applications such as natur al language image description or retrieval.
****

Deterministic Fitting of Multiple Structures Using Iterative MaxFS with Inlier S cale Estimation
Kwang Hee Lee, Sang Wook Lee; Proceedings of the IEEE International Conference o n Computer Vision (ICCV), 2013, pp. 41-48
We present an efficient deterministic hypothesis generation algorithm for robust fitting of multiple structures based on the maximum feasible subsystem (MaxFS) framework. Despite its advantage, a global optimization method such as MaxFS has two main limitations for geometric model fitting. First, its performance is muc h influenced by the user-specified inlier scale. Second, it is computationally i nefficient for large data. The presented algorithm, called iterative MaxFS with inlier scale (IMaxFS-ISE), iteratively estimates model parameters and inlier sca le and also overcomes the second limitation by reducing data for the MaxFS probl em. The IMaxFS-ISE algorithm generates hypotheses only with top-n ranked subsets based on matching scores and data fitting residuals. This reduction of data for the MaxFS problem makes the algorithm computationally realistic. A sequential " fitting and remaving" procedure is repeated until overall energy function does n

ot decrease. Experimental results demonstrate that our method can generate more reliable and consistent hypotheses than random sampling-based methods for estimating multiple structures from data with many outliers.
****

Efficient and Robust Large-Scale Rotation Averaging
Avishek Chatterjee, Venu Madhav Govindu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 521-528
In this paper we address the problem of robust and efficient averaging of relative 3D rotations. Apart from having an interesting geometric structure, robust rotation averaging addresses the need for a good initialization for largescale optimization used in structure-from-motion pipelines. Such pipelines often use unstructured image datasets harvested from the internet thereby requiring an initialization method that is robust to outliers. Our approach works on the Lie group structure of 3D rotations and solves the problem of large-scale robust rotation averaging in two ways. Firstly, we use modern 1 optimizers to carry out robust averaging of relative rotations that is efficient, scalable and robust to outliers. In addition, we also develop a twostep method that uses the 1 solution as an initialisation for an iteratively reweighted least squares (IRLS) approach. These methods achieve excellent results on large-scale, real world datasets and significantly outperform existing methods, i.e. the state-of-the-art discrete-continuous optimization method of [3] as well as the Weiszfeld method of [8]. We demonstrate the efficacy of our method on two largescale real world datasets and also provide the results of the two aforementioned methods for comparison.
****

Automatic Registration of RGB-D Scans via Salient Directions
Bernhard Zeisl, Kevin Koser, Marc Pollefeys; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2808-2815
We address the problem of wide-baseline registration of RGB-D data, such as photo-textured laser scans without any artificial targets or prediction on the relative motion. Our approach allows to fully automatically register scans taken in GPS-denied environments such as urban canyon, industrial facilities or even indoors. We build upon image features which are plenty, localized well and much more discriminative than geometry features; however, they suffer from viewpoint distortions and request for normalization. We utilize the principle of salient directions present in the geometry and propose to extract (several) directions from the distribution of surface normals or other cues such as observable symmetries. Compared to previous work we pose no requirements on the scanned scene (like containing large textured planes) and can handle arbitrary surface shapes. Rendering the whole scene from these repeatable directions using an orthographic camera generates textures which are identical up to 2D similarity transformations. This ambiguity is naturally handled by 2D features and allows to find stable correspondences among scans. For geometric pose estimation from tentative matches we propose a fast and robust 2 point sample consensus scheme integrating an early rejection phase. We evaluate our approach on different challenging real world scenes.
****

Video Co-segmentation for Meaningful Action Extraction
Jiaming Guo, Zhuwen Li, Loong-Fah Cheong, Steven Zhiying Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2232-2239
Given a pair of videos having a common action, our goal is to simultaneously segment this pair of videos to extract this common action. As a preprocessing step, we first remove background trajectories by a motion-based figureground segmentation. To remove the remaining background and those extraneous actions, we propose the trajectory cosaliency measure, which captures the notion that trajectories recurring in all the videos should have their mutual saliency boosted. This requires a trajectory matching process which can compare trajectories with different lengths and not necessarily spatiotemporally aligned, and yet be discriminative enough despite significant intra-class variation in the common action. We further leverage the graph matching to enforce geometric coherence between regions so as to reduce feature ambiguity and matching errors. Finally, to classify the t

rajectories into common action and action outliers, we formulate the problem as a binary labeling of a Markov Random Field, in which the data term is measured by the trajectory co-saliency and the smoothness term is measured by the spatiotemporal consistency between trajectories. To evaluate the performance of our framework, we introduce a dataset containing clips that have animal actions as well as human actions. Experimental results show that the proposed method performs well in common action extraction.
****

Coherent Motion Segmentation in Moving Camera Videos Using Optical Flow Orientations

Manjunath Narayana, Allen Hanson, Erik Learned-Miller; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1577-1584

In moving camera videos, motion segmentation is commonly performed using the image plane motion of pixels, or optical flow. However, objects that are at different depths from the camera can exhibit different optical flows even if they share the same real-world motion. This can cause a depth-dependent segmentation of the scene. Our goal is to develop a segmentation algorithm that clusters pixels that have similar real-world motion irrespective of their depth in the scene. Our solution uses optical flow orientations instead of the complete vectors and exploits the well-known property that under camera translation, optical flow orientations are independent of object depth. We introduce a probabilistic model that automatically estimates the number of observed independent motions and results in a labeling that is consistent with real-world motion in the scene. The result of our system is that static objects are correctly identified as one segment, even if they are at different depths. Color features and information from previous frames in the video sequence are used to correct occasional errors due to the orientation-based segmentation. We present results on more than thirty videos from different benchmarks. The system is particularly robust on complex background scenes containing objects at significantly different depths.
****

Live Metric 3D Reconstruction on Mobile Phones

Petri Tanskanen, Kalin Kolev, Lorenz Meier, Federico Camposeco, Olivier Saurer, Marc Pollefeys; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 65-72

In this paper, we propose a complete on-device 3D reconstruction pipeline for mobile monocular hand-held devices, which generates dense 3D models with absolute scale on-site while simultaneously supplying the user with real-time interactive feedback. The method fills a gap in current cloud-based mobile reconstruction services as it ensures at capture time that the acquired image set fulfills desired quality and completeness criteria. In contrast to existing systems, the developed framework offers multiple innovative solutions. In particular, we investigate the usability of the available on-device inertial sensors to make the tracking and mapping process more resilient to rapid motions and to estimate the metric scale of the captured scene. Moreover, we propose an efficient and accurate scheme for dense stereo matching which allows to reduce the processing time to interaction speed. We demonstrate the performance of the reconstruction pipeline on multiple challenging indoor and outdoor scenes of different size and depth variability.
****

Dynamic Structured Model Selection

David Weiss, Benjamin Sapp, Ben Taskar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2656-2663

In many cases, the predictive power of structured models for for complex vision tasks is limited by a trade-off between the expressiveness and the computational tractability of the model. However, choosing this trade-off statically a priori is suboptimal, as images and videos in different settings vary tremendously in complexity. On the other hand, choosing the trade-off dynamically requires knowledge about the accuracy of different structured models on any given example. In this work, we propose a novel two-tier architecture that provides dynamic speed/accuracy trade-offs through a simple type of introspection. Our approach, which

we call dynamic structured model selection (DMS), leverages typically intractable features in structured learning problems in order to automatically determine' which of several models should be used at test-time in order to maximize accuracy under a fixed budgetary constraint. We demonstrate DMS on two sequential modeling vision tasks, and we establish a new state-of-the-art in human pose estimation in video with an implementation that is roughly 23x faster than the previous standard implementation.
****

Ensemble Projection for Semi-supervised Image Classification
Dengxin Dai, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2072-2079
This paper investigates the problem of semi-supervised classification. Unlike previous methods to regularize classifying boundaries with unlabeled data, our method learns a new image representation from all available data (labeled and unlabeled) and performs plain supervised learning with the new feature. In particular, an ensemble of image prototype sets are sampled automatically from the available data, to represent a rich set of visual categories/attributes. Discriminative functions are then learned on these prototype sets, and image are represented by the concatenation of their projected values onto the prototypes (similarities to them) for further classification. Experiments on four standard datasets show three interesting phenomena: (1) our method consistently outperforms previous methods for semi-supervised image classification; (2) our method lets itself combine well with these methods; and (3) our method works well for self-taught image classification where unlabeled data are not coming from the same distribution as labeled ones, but rather from a random collection of images.
****

Saliency Detection in Large Point Sets
Elizabeth Shtrom, George Leifman, Ayellet Tal; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3591-3598
While saliency in images has been extensively studied in recent years, there is very little work on saliency of point sets. This is despite the fact that point sets and range data are becoming ever more widespread and have myriad applications. In this paper we present an algorithm for detecting the salient points in unorganized 3D point sets. Our algorithm is designed to cope with extremely large sets, which may contain tens of millions of points. Such data is typical of urban scenes, which have recently become commonly available on the web. No previous work has handled such data. For general data sets, we show that our results are competitive with those of saliency detection of surfaces, although we do not have any connectivity information. We demonstrate the utility of our algorithm in two applications: producing a set of the most informative viewpoints and suggesting an informative city tour given a city scan.
****

Segmentation Driven Object Detection with Fisher Vectors
Ramazan Gokberk Cinbis, Jakob Verbeek, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2968-2975
We present an object detection system based on the Fisher vector (FV) image representation computed over SIFT and color descriptors. For computational and storage efficiency, we use a recent segmentation-based method to generate class-independent object detection hypotheses, in combination with data compression techniques. Our main contribution is a method to produce tentative object segmentation masks to suppress background clutter in the features. Re-weighting the local image features based on these masks is shown to improve object detection significantly. We also exploit contextual features in the form of a full-image FV descriptor, and an inter-category rescoring mechanism. Our experiments on the PASCAL VOC 2007 and 2010 datasets show that our detector improves over the current state-of-the-art detection results.
****

Joint Segmentation and Pose Tracking of Human in Natural Videos
Taegyu Lim, Seunghoon Hong, Bohyung Han, Joon Hee Han; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 833-840

We propose an on-line algorithm to extract a human by foreground/background segm
entation and estimate pose of the human from the videos captured by moving camer
as. We claim that a virtuous cycle can be created by appropriate interactions be
tween the two modules to solve individual problems. This joint estimation proble
m is divided into two subproblems, foreground/background segmentation and pose t
racking, which alternate iteratively for optimization; segmentation step generat
es foreground mask for human pose tracking, and human pose tracking step provide
s foreground response map for segmentation. The final solution is obtained when
the iterative procedure converges. We evaluate our algorithm quantitatively and
qualitatively in real videos involving various challenges, and present its outst
anding performance compared to the state-of-the-art techniques for segmentation
and pose estimation.
****

NYC3DCars: A Dataset of 3D Vehicles in Geographic Context
Kevin Matzen, Noah Snavely; Proceedings of the IEEE International Conference on
Computer Vision (ICCV), 2013, pp. 761-768
Geometry and geography can play an important role in recognition tasks in comput
er vision. To aid in studying connections between geometry and recognition, we i
ntroduce NYC3DCars, a rich dataset for vehicle detection in urban scenes built f
rom Internet photos drawn from the wild, focused on densely trafficked areas of
New York City. Our dataset is augmented with detailed geometric and geographic i
nformation, including full camera poses derived from structure from motion, 3D v
ehicle annotations, and geographic information from open resources, including ro
ad segmentations and directions of travel. NYC3DCars can be used to study new qu
estions about using geometric information in detection tasks, and to explore app
lications of Internet photos in understanding cities. To demonstrate the utility
 of our data, we evaluate the use of the geographic information in our dataset t
o enhance a parts-based detection method, and suggest other avenues for future e
xploration.
****

Robust Trajectory Clustering for Motion Segmentation
Feng Shi, Zhong Zhou, Jiangjian Xiao, Wei Wu; Proceedings of the IEEE Internatio
nal Conference on Computer Vision (ICCV), 2013, pp. 3088-3095
Due to occlusions and objects' non-rigid deformation in the scene, the obtained
motion trajectories from common trackers may contain a number of missing or mis-
associated entries. To cluster such corrupted point based trajectories into mult
iple motions is still a hard problem. In this paper, we present an approach that
 exploits temporal and spatial characteristics from tracked points to facilitate
 segmentation of incomplete and corrupted trajectories, thereby obtain highly ro
bust results against severe data missing and noises. Our method first uses the D
iscrete Cosine Transform (DCT) bases as a temporal smoothness constraint on traj
ectory projection to ensure the validity of resulting components to repair patho
logical trajectories. Then, based on an observation that the trajectories of for
eground and background in a scene may have different spatial distributions, we p
ropose a two-stage clustering strategy that first performs foreground-background
 separation then segments remaining foreground trajectories. We show that, with
this new clustering strategy, sequences with complex motions can be accurately s
egmented by even using a simple translational model. Finally, a series of experi
ments on Hopkins 155 dataset and Berkeley motion segmentation dataset show the a
dvantage of our method over other state-of-the-art motion segmentation algorithm
s in terms of both effectiveness and robustness.
****

Active Learning of an Action Detector from Untrimmed Videos
Sunil Bandla, Kristen Grauman; Proceedings of the IEEE International Conference
on Computer Vision (ICCV), 2013, pp. 1833-1840
Collecting and annotating videos of realistic human actions is tedious, yet crit
ical for training action recognition systems. We propose a method to actively re
quest the most useful video annotations among a large set of unlabeled videos. P
redicting the utility of annotating unlabeled video is not trivial, since any gi
ven clip may contain multiple actions of interest, and it need not be trimmed to

temporal regions of interest. To deal with this problem, we propose a detection-based active learner to train action category models. We develop a voting-based framework to localize likely intervals of interest in an unlabeled clip, and use them to estimate the total reduction in uncertainty that annotating that clip would yield. On three datasets, we show our approach can learn accurate action detectors more efficiently than alternative active learning strategies that fail to accommodate the "untrimmed" nature of real video data.
****

YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-Shot Recognition
Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, Kate Saenko; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2712-2719
Despite a recent push towards large-scale object recognition, activity recognition remains limited to narrow domains and small vocabularies of actions. In this paper, we tackle the challenge of recognizing and describing activities "in-the-wild". We present a solution that takes a short video clip and outputs a brief sentence that sums up the main activity in the video, such as the actor, the action and its object. Unlike previous work, our approach works on out-of-domain actions: it does not require training videos of the exact activity. If it cannot find an accurate prediction for a pre-trained model, it finds a less specific answer that is also plausible from a pragmatic standpoint. We use semantic hierarchies learned from the data to help to choose an appropriate level of generalization, and priors learned from web-scale natural language corpora to penalize unlikely combinations of actors/actions/objects; we also use a web-scale language model to "fill in" novel verbs, i.e. when the verb does not appear in the training set. We evaluate our method on a large YouTube corpus and demonstrate it is able to generate short sentence descriptions of video clips better than baseline approaches.
****

Manifold Based Face Synthesis from Sparse Samples
Hongteng Xu, Hongyuan Zha; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2208-2215
Data sparsity has been a thorny issue for manifold-based image synthesis, and in this paper we address this critical problem by leveraging ideas from transfer learning. Specifically, we propose methods based on generating auxiliary data in the form of synthetic samples using transformations of the original sparse samples. To incorporate the auxiliary data, we propose a weighted data synthesis method, which adaptively selects from the generated samples for inclusion during the manifold learning process via a weighted iterative algorithm. To demonstrate the feasibility of the proposed method, we apply it to the problem of face image synthesis from sparse samples. Compared with existing methods, the proposed method shows encouraging results with good performance improvements.
****

Like Father, Like Son: Facial Expression Dynamics for Kinship Verification
Hamdi Dibeklioglu, Albert Ali Salah, Theo Gevers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1497-1504
Kinship verification from facial appearance is a difficult problem. This paper explores the possibility of employing facial expression dynamics in this problem. By using features that describe facial dynamics and spatio-temporal appearance over smile expressions, we show that it is possible to improve the state of the art in this problem, and verify that it is indeed possible to recognize kinship by resemblance of facial expressions. The proposed method is tested on different kin relationships. On the average, 72.89% verification accuracy is achieved on spontaneous smiles.
****

Toward Guaranteed Illumination Models for Non-convex Objects
Yuqian Zhang, Cun Mu, Han-Wen Kuo, John Wright; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 937-944
Illumination variation remains a central challenge in object detection and recog

nition. Existing analyses of illumination variation typically pertain to convex, Lambertian objects, and guarantee quality of approximation in an average case sense. We show that it is possible to build models for the set of images across illumination variation with worstcase performance guarantees, for nonconvex Lambertian objects. Namely, a natural verification test based on the distance to the model guarantees to accept any image which can be sufficiently well-approximated by an image of the object under some admissible lighting condition, and guarantees to reject any image that does not have a sufficiently good approximation. These models are generated by sampling illumination directions with sufficient density, which follows from a new perturbation bound for directional illuminated images in the Lambertian model. As the number of such images required for guaranteed verification may be large, we introduce a new formulation for cone preserving dimensionality reduction, which leverages tools from sparse and low-rank decomposition to reduce the complexity, while controlling the approximation error with respect to the original model. 1
****

Nonparametric Blind Super-resolution
Tomer Michaeli, Michal Irani; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 945-952
Super resolution (SR) algorithms typically assume that the blur kernel is known (either the Point Spread Function 'PSF' of the camera, or some default low-pass filter, e.g. a Gaussian). However, the performance of SR methods significantly deteriorates when the assumed blur kernel deviates from the true one. We propose a general framework for "blind" super resolution. In particular, we show that: (i) Unlike the common belief, the PSF of the camera is the wrong blur kernel to use in SR algorithms. (ii) We show how the correct SR blur kernel can be recovered directly from the low-resolution image. This is done by exploiting the inherent recurrence property of small natural image patches (either internally within the same image, or externally in a collection of other natural images). In particular, we show that recurrence of small patches across scales of the low-res image (which forms the basis for single-image SR), can also be used for estimating the optimal blur kernel. This leads to significant improvement in SR results.
****

Pedestrian Parsing via Deep Decompositional Network
Ping Luo, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2648-2655
We propose a new Deep Decompositional Network (DDN) for parsing pedestrian images into semantic regions, such as hair, head, body, arms, and legs, where the pedestrians can be heavily occluded. Unlike existing methods based on template matching or Bayesian inference, our approach directly maps low-level visual features to the label maps of body parts with DDN, which is able to accurately estimate complex pose variations with good robustness to occlusions and background clutters. DDN jointly estimates occluded regions and segments body parts by stacking three types of hidden layers: occlusion estimation layers, completion layers, and decomposition layers. The occlusion estimation layers estimate a binary mask, indicating which part of a pedestrian is invisible. The completion layers synthesize low-level features of the invisible part from the original features and the occlusion mask. The decomposition layers directly transform the synthesized visual features to label maps. We devise a new strategy to pre-train these hidden layers, and then fine-tune the entire network using the stochastic gradient descent. Experimental results show that our approach achieves better segmentation accuracy than the state-of-the-art methods on pedestrian images with or without occlusions. Another important contribution of this paper is that it provides a large scale benchmark human parsing dataset 1 that includes 3, 673 annotated samples collected from 171 surveillance videos. It is 20 times larger than existing public datasets.
****

Large-Scale Video Hashing via Structure Learning
Guangnan Ye, Dong Liu, Jun Wang, Shih-Fu Chang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2272-2279

Recently, learning based hashing methods have become popular for indexing large-scale media data. Hashing methods map high-dimensional features to compact binary codes that are efficient to match and robust in preserving original similarity. However, most of the existing hashing methods treat videos as a simple aggregation of independent frames and index each video through combining the indexes of frames. The structure information of videos, e.g., discriminative local visual commonality and temporal consistency, is often neglected in the design of hash functions. In this paper, we propose a supervised method that explores the structure learning techniques to design efficient hash functions. The proposed video hashing method formulates a minimization problem over a structure-regularized empirical loss. In particular, the structure regularization exploits the common local visual patterns occurring in video frames that are associated with the same semantic class, and simultaneously preserves the temporal consistency over successive frames from the same video. We show that the minimization objective can be efficiently solved by an Accelerated Proximal Gradient (APG) method. Extensive experiments on two large video benchmark datasets (up to around 150K video clips with over 12 million frames) show that the proposed method significantly outperforms the state-ofthe-art hashing methods.

****

Salient Region Detection by UFO: Uniqueness, Focusness and Objectness

Peng Jiang, Haibin Ling, Jingyi Yu, Jingliang Peng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1976-1983

The goal of saliency detection is to locate important pixels or regions in an image which attract humans' visual attention the most. This is a fundamental task whose output may serve as the basis for further computer vision tasks like segmentation, resizing, tracking and so forth. In this paper we propose a novel salient region detection algorithm by integrating three important visual cues namely uniqueness, focusness and objectness (UFO). In particular, uniqueness captures the appearance-derived visual contrast; focusness reflects the fact that salient regions are often photographed in focus; and objectness helps keep completeness of detected salient regions. While uniqueness has been used for saliency detection for long, it is new to integrate focusness and objectness for this purpose. In fact, focusness and objectness both provide important saliency information complementary of uniqueness. In our experiments using public benchmark datasets, we show that, even with a simple pixel level combination of the three components, the proposed approach yields significant improvement compared with previously reported methods.

****

Revisiting the PnP Problem: A Fast, General and Optimal Solution

Yinqiang Zheng, Yubin Kuang, Shigeki Sugimoto, Kalle Astrom, Masatoshi Okutomi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2344-2351

In this paper, we revisit the classical perspective-n-point (PnP) problem, and propose the first non-iterative O(n) solution that is fast, generally applicable and globally optimal. Our basic idea is to formulate the PnP problem into a functional minimization problem and retrieve all its stationary points by using the Gr??bner basis technique. The novelty lies in a non-unit quaternion representation to parameterize the rotation and a simple but elegant formulation of the PnP problem into an unconstrained optimization problem. Interestingly, the polynomial system arising from its first-order optimality condition assumes two-fold symmetry, a nice property that can be utilized to improve speed and numerical stability of a Gr??bner basis solver. Experiment results have demonstrated that, in terms of accuracy, our proposed solution is definitely better than the state-ofthe-art O(n) methods, and even comparable with the reprojection error minimization method.

****

Rectangling Stereographic Projection for Wide-Angle Image Visualization

Che-Han Chang, Min-Chun Hu, Wen-Huang Cheng, Yung-Yu Chuang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2824-2831

This paper proposes a new projection model for mapping a hemisphere to a plane.

Such a model can be useful for viewing wide-angle images. Our model consists of two steps. In the first step, the hemisphere is projected onto a swung surface constructed by a circular profile and a rounded rectangular trajectory. The second step maps the projected image on the swung surface onto the image plane through the perspective projection. We also propose a method for automatically determining proper parameters for the projection model based on image content. The proposed model has several advantages. It is simple, efficient and easy to control. Most importantly, it makes a better compromise between distortion minimization and line preserving than popular projection models, such as stereographic and Pannini projections. Experiments and analysis demonstrate the effectiveness of our model.

****

Hidden Factor Analysis for Age Invariant Face Recognition

Dihong Gong, Zhifeng Li, Dahua Lin, Jianzhuang Liu, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2872-2879

Age invariant face recognition has received increasing attention due to its great potential in real world applications. In spite of the great progress in face recognition techniques, reliably recognizing faces across ages remains a difficult task. The facial appearance of a person changes substantially over time, resulting in significant intra-class variations. Hence, the key to tackle this problem is to separate the variation caused by aging from the person-specific features that are stable. Specifically, we propose a new method, called Hidden Factor Analysis (HFA). This method captures the intuition above through a probabilistic model with two latent factors: an identity factor that is age-invariant and an age factor affected by the aging process. Then, the observed appearance can be modeled as a combination of the components generated based on these factors. We also develop a learning algorithm that jointly estimates the latent factors and the model parameters using an EM procedure. Extensive experiments on two well-known public domain face aging datasets: MORPH (the largest public face aging database) and FGNET, clearly show that the proposed method achieves notable improvement over state-of-the-art algorithms.

****

Randomized Ensemble Tracking

Qinxun Bai, Zheng Wu, Stan Sclaroff, Margrit Betke, Camille Monnier; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2040-2047

We propose a randomized ensemble algorithm to model the time-varying appearance of an object for visual tracking. In contrast with previous online methods for updating classifier ensembles in tracking-by-detection, the weight vector that combines weak classifiers is treated as a random variable and the posterior distribution for the weight vector is estimated in a Bayesian manner. In essence, the weight vector is treated as a distribution that reflects the confidence among the weak classifiers used to construct and adapt the classifier ensemble. The resulting formulation models the time-varying discriminative ability among weak classifiers so that the ensembled strong classifier can adapt to the varying appearance, backgrounds, and occlusions. The formulation is tested in a tracking-by-detection implementation. Experiments on 28 challenging benchmark videos demonstrate that the proposed method can achieve results comparable to and often better than those of stateof-the-art approaches.

****

Motion-Aware KNN Laplacian for Video Matting

Dingzeyu Li, Qifeng Chen, Chi-Keung Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3599-3606

This paper demonstrates how the nonlocal principle benefits video matting via the KNN Laplacian, which comes with a straightforward implementation using motionaware K nearest neighbors. In hindsight, the fundamental problem to solve in video matting is to produce spatiotemporally coherent clusters of moving foreground pixels. When used as described, the motion-aware KNN Laplacian is effective in addressing this fundamental problem, as demonstrated by sparse user markups typically on only one frame in a variety of challenging examples featuring ambiguous

foreground and background colors, changing topologies with disocclusion, signifi
cant illumination changes, fast motion, and motion blur. When working with exist
ing Laplacian-based systems, our Laplacian is expected to benefit them immediate
ly with improved clustering of moving foreground pixels.
****

Pose-Configurable Generic Tracking of Elongated Objects
Daniel Wesierski, Patrick Horain; Proceedings of the IEEE International Conferen
ce on Computer Vision (ICCV), 2013, pp. 2920-2927
Elongated objects have various shapes and can shift, rotate, change scale, and b
e rigid or deform by flexing, articulating, and vibrating, with examples as vari
ed as a glass bottle, a robotic arm, a surgical suture, a finger pair, a tram, a
nd a guitar string. This generally makes tracking of poses of elongated objects
very challenging. We describe a unified, configurable framework for tracking the
 pose of elongated objects, which move in the image plane and extend over the im
age region. Our method strives for simplicity, versatility, and efficiency. The
object is decomposed into a chained assembly of segments of multiple parts that
are arranged under a hierarchy of tailored spatio-temporal constraints. In this
hierarchy, segments can rescale independently while their elasticity is controll
ed with global orientations and local distances. While the trend in tracking is
to design complex, structure-free algorithms that update object appearance onlin
e, we show that our tracker, with the novel but remarkably simple, structured or
ganization of parts with constant appearance, reaches or improves state-of-the-a
rt performance. Most importantly, our model can be easily configured to track ex
act pose of arbitrary, elongated objects in the image plane. The tracker can run
 up to 100 fps on a desktop PC, yet the computation time scales linearly with th
e number of object parts. To our knowledge, this is the first approach to generi
c tracking of elongated objects.
****

Heterogeneous Auto-similarities of Characteristics (HASC): Exploiting Relational
 Information for Classification
Marco San Biagio, Marco Crocco, Marco Cristani, Samuele Martelli, Vittorio Murin
o; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2
013, pp. 809-816
Capturing the essential characteristics of visual objects by considering how the
ir features are inter-related is a recent philosophy of object classification. I
n this paper, we embed this principle in a novel image descriptor, dubbed Hetero
geneous Auto-Similarities of Characteristics (HASC). HASC is applied to heteroge
neous dense features maps, encoding linear relations by covariances and nonlinea
r associations through information-theoretic measures such as mutual information
 and entropy. In this way, highly complex structural information can be expresse
d in a compact, scale invariant and robust manner. The effectiveness of HASC is
tested on many diverse detection and classification scenarios, considering objec
ts, textures and pedestrians, on widely known benchmarks (Caltech-101, Brodatz,
Daimler Multi-Cue). In all the cases, the results obtained with standard classif
iers demonstrate the superiority of HASC with respect to the most adopted local
feature descriptors nowadays, such as SIFT, HOG, LBP and feature covariances. In
 addition, HASC sets the state-of-the-art on the Brodatz texture dataset and the
 Daimler Multi-Cue pedestrian dataset, without exploiting ad-hoc sophisticated c
lassifiers.
****

A Learning-Based Approach to Reduce JPEG Artifacts in Image Matting
Inchang Choi, Sunyeong Kim, Michael S. Brown, Yu-Wing Tai; Proceedings of the IE
EE International Conference on Computer Vision (ICCV), 2013, pp. 2880-2887
Single image matting techniques assume high-quality input images. The vast major
ity of images on the web and in personal photo collections are encoded using JPE
G compression. JPEG images exhibit quantization artifacts that adversely affect
the performance of matting algorithms. To address this situation, we propose a l
earning-based post-processing method to improve the alpha mattes extracted from
JPEG images. Our approach learns a set of sparse dictionaries from training exam
ples that are used to transfer details from high-quality alpha mattes to alpha m

attes corrupted by JPEG compression. Three different dictionaries are defined to accommodate different object structure (long hair, short hair, and sharp boundaries). A back-projection criteria combined within an MRF framework is used to automatically select the best dictionary to apply on the object's local boundary. We demonstrate that our method can produces superior results over existing state-of-the-art matting algorithms on a variety of inputs and compression levels.
****

Content-Aware Rotation
Kaiming He, Huiwen Chang, Jian Sun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 553-560
We present an image editing tool called Content-Aware Rotation. Casually shot photos can appear tilted, and are often corrected by rotation and cropping. This trivial solution may remove desired content and hurt image integrity. Instead of doing rigid rotation, we propose a warping method that creates the perception of rotation and avoids cropping. Human vision studies suggest that the perception of rotation is mainly due to horizontal/vertical lines. We design an optimization-based method that preserves the rotation of horizontal/vertical lines, maintains the completeness of the image content, and reduces the warping distortion. An efficient algorithm is developed to address the challenging optimization. We demonstrate our content-aware rotation method on a variety of practical cases.
****

Perspective Motion Segmentation via Collaborative Clustering
Zhuwen Li, Jiaming Guo, Loong-Fah Cheong, Steven Zhiying Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1369-1376
This paper addresses real-world challenges in the motion segmentation problem, including perspective effects, missing data, and unknown number of motions. It first formulates the 3-D motion segmentation from two perspective views as a subspace clustering problem, utilizing the epipolar constraint of an image pair. It then combines the point correspondence information across multiple image frames via a collaborative clustering step, in which tight integration is achieved via a mixed norm optimization scheme. For model selection, we propose an over-segment and merge approach, where the merging step is based on the property of the 1-norm of the mutual sparse representation of two oversegmented groups. The resulting algorithm can deal with incomplete trajectories and perspective effects substantially better than state-of-the-art two-frame and multi-frame methods. Experiments on a 62-clip dataset show the significant superiority of the proposed idea in both segmentation accuracy and model selection.
****

Progressive Multigrid Eigensolvers for Multiscale Spectral Segmentation
Michael Maire, Stella X. Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2184-2191
We reexamine the role of multiscale cues in image segmentation using an architecture that constructs a globally coherent scale-space output representation. This characteristic is in contrast to many existing works on bottom-up segmentation, which prematurely compress information into a single scale. The architecture is a standard extension of Normalized Cuts from an image plane to an image pyramid, with cross-scale constraints enforcing consistency in the solution while allowing emergence of coarse-to-fine detail. We observe that multiscale processing, in addition to improving segmentation quality, offers a route by which to speed computation. We make a significant algorithmic advance in the form of a custom multigrid eigensolver for constrained Angular Embedding problems possessing coarse-to-fine structure. Multiscale Normalized Cuts is a special case. Our solver builds atop recent results on randomized matrix approximation, using a novel interpolation operation to mold its computational strategy according to crossscale constraints in the problem definition. Applying our solver to multiscale segmentation problems demonstrates speedup by more than an order of magnitude. This speedup is at the algorithmic level and carries over to any implementation target.
****

Shape Index Descriptors Applied to Texture-Based Galaxy Analysis
Kim Steenstrup Pedersen, Kristoffer Stensbo-Smidt, Andrew Zirm, Christian Igel;

Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2440-2447
A texture descriptor based on the shape index and the accompanying curvedness measure is proposed, and it is evaluated for the automated analysis of astronomical image data. A representative sample of images of low-redshift galaxies from the Sloan Digital Sky Survey (SDSS) serves as a testbed. The goal of applying texture descriptors to these data is to extract novel information about galaxies; information which is often lost in more traditional analysis. In this study, we build a regression model for predicting a spectroscopic quantity, the specific star-formation rate (sSFR). As texture features we consider multi-scale gradient orientation histograms as well as multi-scale shape index histograms, which lead to a new descriptor. Our results show that we can successfully predict spectroscopic quantities from the texture in optical multi-band images. We successfully recover the observed bi-modal distribution of galaxies into quiescent and star-forming. The state-ofthe-art for predicting the sSFR is a color-based physical model. We significantly improve its accuracy by augmenting the model with texture information. This study is the first step towards enabling the quantification of physical galaxy properties from imaging data alone.
****

Markov Network-Based Unified Classifier for Face Identification
Wonjun Hwang, Kyungshik Roh, Junmo Kim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1952-1959
We propose a novel unifying framework using a Markov network to learn the relationship between multiple classifiers in face recognition. We assume that we have several complementary classifiers and assign observation nodes to the features of a query image and hidden nodes to the features of gallery images. We connect each hidden node to its corresponding observation node and to the hidden nodes of other neighboring classifiers. For each observation-hidden node pair, we collect a set of gallery candidates that are most similar to the observation instance, and the relationship between the hidden nodes is captured in terms of the similarity matrix between the collected gallery images. Posterior probabilities in the hidden nodes are computed by the belief-propagation algorithm. The novelty of the proposed framework is the method that takes into account the classifier dependency using the results of each neighboring classifier. We present extensive results on two different evaluation protocols, known and unknown image variation tests, using three different databases, which shows that the proposed framework always leads to good accuracy in face recognition.
****

Learning Hash Codes with Listwise Supervision
Jun Wang, Wei Liu, Andy X. Sun, Yu-Gang Jiang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3032-3039
Hashing techniques have been intensively investigated in the design of highly efficient search engines for largescale computer vision applications. Compared with prior approximate nearest neighbor search approaches like treebased indexing, hashing-based search schemes have prominent advantages in terms of both storage and computational efficiencies. Moreover, the procedure of devising hash functions can be easily incorporated into sophisticated machine learning tools, leading to data-dependent and task-specific compact hash codes. Therefore, a number of learning paradigms, ranging from unsupervised to supervised, have been applied to compose appropriate hash functions. However, most of the existing hash function learning methods either treat hash function design as a classification problem or generate binary codes to satisfy pairwise supervision, and have not yet directly optimized the search accuracy. In this paper, we propose to leverage listwise supervision into a principled hash function learning framework. In particular, the ranking information is represented by a set of rank triplets that can be used to assess the quality of ranking. Simple linear projection-based hash functions are solved efficiently through maximizing the ranking quality over the training data. We carry out experiments on large image datasets with size up to one million and compare with the state-of-the-art hashing techniques. The extensive results corroborate that our learned hash codes via listwise supervision can prov

ide superior search accuracy without incurring heavy computational overhead.
****
A Robust Analytical Solution to Isometric Shape-from-Template with Focal Length Calibration
Adrien Bartoli, Daniel Pizarro, Toby Collins; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 961-968
We study the uncalibrated isometric Shape-fromTemplate problem, that consists in estimating an isometric deformation from a template shape to an input image whose focal length is unknown. Our method is the first that combines the following features: solving for both the 3D deformation and the camera's focal length, involving only local analytical solutions (there is no numerical optimization), being robust to mismatches, handling general surfaces and running extremely fast. This was achieved through two key steps. First, an 'uncalibrated' 3D deformation is computed thanks to a novel piecewise weak-perspective projection model. Second, the camera's focal length is estimated and enables upgrading the 3D deformation to metric. We use a variational framework, implemented using a smooth function basis and sampled local deformation models. The only degeneracy which we easily detectfor focal length estimation is a flat and fronto-parallel surface. Experimental results on simulated and real datasets show that our method achieves a 3D shape accuracy slightly below state of the art methods using a precalibrated or the true focal length, and a focal length accuracy slightly below static calibration methods.
****
Real-World Normal Map Capture for Nearly Flat Reflective Surfaces
Bastien Jacquet, Christian Hane, Kevin Koser, Marc Pollefeys; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 713-720
Although specular objects have gained interest in recent years, virtually no approaches exist for markerless reconstruction of reflective scenes in the wild. In this work, we present a practical approach to capturing normal maps in real-world scenes using video only. We focus on nearly planar surfaces such as windows, facades from glass or metal, or frames, screens and other indoor objects and show how normal maps of these can be obtained without the use of an artificial calibration object. Rather, we track the reflections of real-world straight lines, while moving with a hand-held or vehicle-mounted camera in front of the object. In contrast to error-prone local edge tracking, we obtain the reflections by a robust, global segmentation technique of an ortho-rectified 3D video cube that also naturally allows efficient user interaction. Then, at each point of the reflective surface, the resulting 2D-curve to 3D-line correspondence provides a novel quadratic constraint on the local surface normal. This allows to globally solve for the shape by integrability and smoothness constraints and easily supports the usage of multiple lines. We demonstrate the technique on several objects and facades.
****
Perceptual Fidelity Aware Mean Squared Error
Wufeng Xue, Xuanqin Mou, Lei Zhang, Xiangchu Feng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 705-712
How to measure the perceptual quality of natural images is an important problem in low level vision. It is known that the Mean Squared Error (MSE) is not an effective index to describe the perceptual fidelity of images. Numerous perceptual fidelity indices have been developed, while the representatives include the Structural SIMilarity (SSIM) index and its variants. However, most of those perceptual measures are nonlinear, and they cannot be easily adopted as an objective function to minimize in various low level vision tasks. Can MSE be perceptual fidelity aware after some minor adaptation? In this paper we propose a simple framework to enhance the perceptual fidelity awareness of MSE by introducing an l 2 -norm structural error term to it. Such a Structural MSE (SMSE) can lead to very competitive image quality assessment (IQA) results. More surprisingly, we show that by using certain structure extractors, SMSE can be further turned into a Gaussian smoothed MSE (i.e., the Euclidean distance between the original and distorted images after Gaussian smooth filtering), which is much simpler to calculate bu

t achieves rather better IQA performance than SSIM. The socalled Perceptual-fide lity Aware MSE (PAMSE) can have great potentials in applications such as percept ual image coding and perceptual image restoration.
****

Temporally Consistent Superpixels
Matthias Reso, Jorn Jachalsky, Bodo Rosenhahn, Jorn Ostermann; Proceedings of th e IEEE International Conference on Computer Vision (ICCV), 2013, pp. 385-392
Superpixel algorithms represent a very useful and increasingly popular preproces sing step for a wide range of computer vision applications, as they offer the po tential to boost efficiency and effectiveness. In this regards, this paper prese nts a highly competitive approach for temporally consistent superpixels for vide o content. The approach is based on energy-minimizing clustering utilizing a nov el hybrid clustering strategy for a multi-dimensional feature space working in a global color subspace and local spatial subspaces. Moreover, a new contour evol ution based strategy is introduced to ensure spatial coherency of the generated superpixels. For a thorough evaluation the proposed approach is compared to stat e of the art supervoxel algorithms using established benchmarks and shows a supe rior performance.
****

Efficient Pedestrian Detection by Directly Optimizing the Partial Area under the ROC Curve
Sakrapee Paisitkriangkrai, Chunhua Shen, Anton Van Den Hengel; Proceedings of th e IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1057-1064
Many typical applications of object detection operate within a prescribed false- positive range. In this situation the performance of a detector should be assess ed on the basis of the area under the ROC curve over that range, rather than ove r the full curve, as the performance outside the range is irrelevant. This measu re is labelled as the partial area under the ROC curve (pAUC). Effective cascade -based classification, for example, depends on training node classifiers that ac hieve the maximal detection rate at a moderate false positive rate, e.g., around 40% to 50%. We propose a novel ensemble learning method which achieves a maxima l detection rate at a user-defined range of false positive rates by directly opt imizing the partial AUC using structured learning. By optimizing for different r anges of false positive rates, the proposed method can be used to train either a single strong classifier or a node classifier forming part of a cascade classif ier. Experimental results on both synthetic and real-world data sets demonstrate the effectiveness of our approach, and we show that it is possible to train sta te-of-the-art pedestrian detectors using the proposed structured ensemble learni ng method.
****

Rank Minimization across Appearance and Shape for AAM Ensemble Fitting
Xin Cheng, Sridha Sridharan, Jason Saragih, Simon Lucey; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 577-584
Active Appearance Models (AAMs) employ a paradigm of inverting a synthesis model of how an object can vary in terms of shape and appearance. As a result, the ab ility of AAMs to register an unseen object image is intrinsically linked to two factors. First, how well the synthesis model can reconstruct the object image. S econd, the degrees of freedom in the model. Fewer degrees of freedom yield a hig her likelihood of good fitting performance. In this paper we look at how these s eemingly contrasting factors can complement one another for the problem of AAM f itting of an ensemble of images stemming from a constrained set (e.g. an ensembl e of face images of the same person).
****

Random Forests of Local Experts for Pedestrian Detection
Javier Marin, David Vazquez, Antonio M. Lopez, Jaume Amores, Bastian Leibe; Proc eedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp . 2592-2599
Pedestrian detection is one of the most challenging tasks in computer vision, an d has received a lot of attention in the last years. Recently, some authors have shown the advantages of using combinations of part/patch-based detectors in ord

er to cope with the large variability of poses and the existence of partial occl
usions. In this paper, we propose a pedestrian detection method that efficiently
 combines multiple local experts by means of a Random Forest ensemble. The propo
sed method works with rich block-based representations such as HOG and LBP, in s
uch a way that the same features are reused by the multiple local experts, so th
at no extra computational cost is needed with respect to a holistic method. Furt
hermore, we demonstrate how to integrate the proposed approach with a cascaded a
rchitecture in order to achieve not only high accuracy but also an acceptable ef
ficiency. In particular, the resulting detector operates at five frames per seco
nd using a laptop machine. We tested the proposed method with well-known challen
ging datasets such as Caltech, ETH, Daimler, and INRIA. The method proposed in t
his work consistently ranks among the top performers in all the datasets, being
 either the best method or having a small difference with the best one.
****

Monocular Image 3D Human Pose Estimation under Self-Occlusion
Ibrahim Radwan, Abhinav Dhall, Roland Goecke; Proceedings of the IEEE Internatio
nal Conference on Computer Vision (ICCV), 2013, pp. 1888-1895
In this paper, an automatic approach for 3D pose reconstruction from a single im
age is proposed. The presence of human body articulation, hallucinated parts and
 cluttered background leads to ambiguity during the pose inference, which makes
the problem non-trivial. Researchers have explored various methods based on moti
on and shading in order to reduce the ambiguity and reconstruct the 3D pose. The
 key idea of our algorithm is to impose both kinematic and orientation constrain
ts. The former is imposed by projecting a 3D model onto the input image and prun
ing the parts, which are incompatible with the anthropomorphism. The latter is a
pplied by creating synthetic views via regressing the input view to multiple ori
ented views. After applying the constraints, the 3D model is projected onto the
initial and synthetic views, which further reduces the ambiguity. Finally, we bo
rrow the direction of the unambiguous parts from the synthetic views to the init
ial one, which results in the 3D pose. Quantitative experiments are performed on
 the HumanEva-I dataset and qualitatively on unconstrained images from the Image
 Parse dataset. The results show the robustness of the proposed approach to accu
rately reconstruct the 3D pose form a single image.
****

Constructing Adaptive Complex Cells for Robust Visual Tracking
Dapeng Chen, Zejian Yuan, Yang Wu, Geng Zhang, Nanning Zheng; Proceedings of the
 IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1113-1120
Representation is a fundamental problem in object tracking. Conventional methods
 track the target by describing its local or global appearance. In this paper we
 present that, besides the two paradigms, the composition of local region histog
rams can also provide diverse and important object cues. We use cells to extract
 local appearance, and construct complex cells to integrate the information from
 cells. With different spatial arrangements of cells, complex cells can explore
various contextual information at multiple scales, which is important to improve
 the tracking performance. We also develop a novel template-matching algorithm f
or object tracking, where the template is composed of temporal varying cells and
 has two layers to capture the target and background appearance respectively. An
 adaptive weight is associated with each complex cell to cope with occlusion as
well as appearance variation. A fusion weight is associated with each complex ce
ll type to preserve the global distinctiveness. Our algorithm is evaluated on 25
 challenging sequences, and the results not only confirm the contribution of eac
h component in our tracking system, but also outperform other competing trackers
.
****

Mining Motion Atoms and Phrases for Complex Action Recognition
Limin Wang, Yu Qiao, Xiaoou Tang; Proceedings of the IEEE International Conferen
ce on Computer Vision (ICCV), 2013, pp. 2680-2687
This paper proposes motion atom and phrase as a midlevel temporal "part" for rep
resenting and classifying complex action. Motion atom is defined as an atomic pa
rt of action, and captures the motion information of action video in a short tem

poral scale. Motion phrase is a temporal composite of multiple motion atoms with an AND/OR structure, which further enhances the discriminative ability of motion atoms by incorporating temporal constraints in a longer scale. Specifically, given a set of weakly labeled action videos, we firstly design a discriminative clustering method to automatically discover a set of representative motion atoms. Then, based on these motion atoms, we mine effective motion phrases with high discriminative and representative power. We introduce a bottom-up phrase construction algorithm and a greedy selection method for this mining task. We examine the classification performance of the motion atom and phrase based representation on two complex action datasets: Olympic Sports and UCF50. Experimental results show that our method achieves superior performance over recent published methods on both datasets.

****

Video Event Understanding Using Natural Language Descriptions

Vignesh Ramanathan, Percy Liang, Li Fei-Fei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 905-912

Human action and role recognition play an important part in complex event understanding. State-of-the-art methods learn action and role models from detailed spatio temporal annotations, which requires extensive human effort. In this work, we propose a method to learn such models based on natural language descriptions of the training videos, which are easier to collect and scale with the number of actions and roles. There are two challenges with using this form of weak supervision: First, these descriptions only provide a high-level summary and often do not directly mention the actions and roles occurring in a video. Second, natural language descriptions do not provide spatio temporal annotations of actions and roles. To tackle these challenges, we introduce a topic-based semantic relatedness (SR) measure between a video description and an action and role label, and incorporate it into a posterior regularization objective. Our event recognition system based on these action and role models matches the state-ofthe-art method on the TRECVID-MED11 event kit, despite weaker supervision.

****

Human Re-identification by Matching Compositional Template with Cluster Sampling

Yuanlu Xu, Liang Lin, Wei-Shi Zheng, Xiaobai Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3152-3159

This paper aims at a newly raising task in visual surveillance: re-identifying people at a distance by matching body information, given several reference examples. Most of existing works solve this task by matching a reference template with the target individual, but often suffer from large human appearance variability (e.g. different poses/views, illumination) and high false positives in matching caused by conjunctions, occlusions or surrounding clutters. Addressing these problems, we construct a simple yet expressive template from a few reference images of a certain individual, which represents the body as an articulated assembly of compositional and alternative parts, and propose an effective matching algorithm with cluster sampling. This algorithm is designed within a candidacy graph whose vertices are matching candidates (i.e. a pair of source and target body parts), and iterates in two steps for convergence. (i) It generates possible partial matches based on compatible and competitive relations among body parts. (ii) It confirms the partial matches to generate a new matching solution, which is accepted by the Markov Chain Monte Carlo (MCMC) mechanism. In the experiments, we demonstrate the superior performance of our approach on three public databases compared to existing methods.

****

Estimating the Material Properties of Fabric from Video

Katherine L. Bouman, Bei Xiao, Peter Battaglia, William T. Freeman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1984-1991

Passively estimating the intrinsic material properties of deformable objects moving in a natural environment is essential for scene understanding. We present a framework to automatically analyze videos of fabrics moving under various unknown wind forces, and recover two key material properties of the fabric: stiffness

and area weight. We extend features previously developed to compactly represent static image textures to describe video textures, such as fabric motion. A discriminatively trained regression model is then used to predict the physical properties of fabric from these features. The success of our model is demonstrated on a new, publicly available database of fabric videos with corresponding measured ground truth material properties. We show that our predictions are well correlated with ground truth measurements of stiffness and density for the fabrics. Our contributions include: (a) a database that can be used for training and testing algorithms for passively predicting fabric properties from video, (b) an algorithm for predicting the material properties of fabric from a video, and (c) a perceptual study of humansâ■■ ability to estimate the material properties of fabric from videos and images.
****

An Adaptive Descriptor Design for Object Recognition in the Wild
Zhenyu Guo, Z. Jane Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2568-2575
Digital images nowadays show large appearance variabilities on picture styles, in terms of color tone, contrast, vignetting, and etc. These 'picture styles' are directly related to the scene radiance, image pipeline of the camera, and post processing functions (e.g., photography effect filters). Due to the complexity and nonlinearity of these factors, popular gradient-based image descriptors generally are not invariant to different picture styles, which could degrade the performance for object recognition. Given that images shared online or created by individual users are taken with a wide range of devices and may be processed by various post processing functions, to find a robust object recognition system is useful and challenging. In this paper, we investigate the influence of picture styles on object recognition by making a connection between image descriptors and a pixel mapping function g, and accordingly propose an adaptive approach based on a g-incorporated kernel descriptor and multiple kernel learning, without estimating or specifying the image styles used in training and testing. We conduct experiments on the Domain Adaptation data set, the Oxford Flower data set, and several variants of the Flower data set by introducing popular photography effects through post-processing. The results demonstrate that the proposed method consistently yields recognition improvements over standard descriptors in all studied cases.
****

Partial Sum Minimization of Singular Values in RPCA for Low-Level Vision
Tae-Hyun Oh, Hyeongwoo Kim, Yu-Wing Tai, Jean-Charles Bazin, In So Kweon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 145-152
Robust Principal Component Analysis (RPCA) via rank minimization is a powerful tool for recovering underlying low-rank structure of clean data corrupted with sparse noise/outliers. In many low-level vision problems, not only it is known that the underlying structure of clean data is low-rank, but the exact rank of clean data is also known. Yet, when applying conventional rank minimization for those problems, the objective function is formulated in a way that does not fully utilize a priori target rank information about the problems. This observation motivates us to investigate whether there is a better alternative solution when using rank minimization. In this paper, instead of minimizing the nuclear norm, we propose to minimize the partial sum of singular values. The proposed objective function implicitly encourages the target rank constraint in rank minimization. Our experimental analyses show that our approach performs better than conventional rank minimization when the number of samples is deficient, while the solutions obtained by the two approaches are almost identical when the number of samples is more than sufficient. We apply our approach to various low-level vision problems, e.g. high dynamic range imaging, photometric stereo and image alignment, and show that our results outperform those obtained by the conventional nuclear norm rank minimization method.
****

Semantically-Based Human Scanpath Estimation with HMMs

Huiying Liu, Dong Xu, Qingming Huang, Wen Li, Min Xu, Stephen Lin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3232-3239

We present a method for estimating human scanpaths, which are sequences of gaze shifts that follow visual attention over an image. In this work, scanpaths are modeled based on three principal factors that influence human attention, namely low-level feature saliency, spatial position, and semantic content. Low-level feature saliency is formulated as transition probabilities between different image regions based on feature differences. The effect of spatial position on gaze shifts is modeled as a Levy flight with the shifts following a 2D Cauchy distribution. To account for semantic content, we propose to use a Hidden Markov Model (HMM) with a Bag-of-Visual-Words descriptor of image regions. An HMM is well-suited for this purpose in that 1) the hidden states, obtained by unsupervised learning, can represent latent semantic concepts, 2) the prior distribution of the hidden states describes visual attraction to the semantic concepts, and 3) the transition probabilities represent human gaze shift patterns. The proposed method is applied to task-driven viewing processes. Experiments and analysis performed on human eye gaze data verify the effectiveness of this method.

****

From Semi-supervised to Transfer Counting of Crowds

Chen Change Loy, Shaogang Gong, Tao Xiang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2256-2263

Regression-based techniques have shown promising results for people counting in crowded scenes. However, most existing techniques require expensive and laborious data annotation for model training. In this study, we propose to address this problem from three perspectives: (1) Instead of exhaustively annotating every single frame, the most informative frames are selected for annotation automatically and actively. (2) Rather than learning from only labelled data, the abundant unlabelled data are exploited. (3) Labelled data from other scenes are employed to further alleviate the burden for data annotation. All three ideas are implemented in a unified active and semi-supervised regression framework with ability to perform transfer learning, by exploiting the underlying geometric structure of crowd patterns via manifold analysis. Extensive experiments validate the effectiveness of our approach.

****

Enhanced Continuous Tabu Search for Parameter Estimation in Multiview Geometry

Guoqing Zhou, Qing Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3240-3247

Optimization using the L ? norm has been becoming an effective way to solve parameter estimation problems in multiview geometry. But the computational cost increases rapidly with the size of measurement data. Although some strategies have been presented to improve the efficiency of L ? optimization, it is still an open issue. In the paper, we propose a novel approach under the framework of enhanced continuous tabu search (ECTS) for generic parameter estimation in multiview geometry. ECTS is an optimization method in the domain of artificial intelligence, which has an interesting ability of covering a wide solution space by promoting the search far away from current solution and consecutively decreasing the possibility of trapping in the local minima. Taking the triangulation as an example, we propose the corresponding ways in the key steps of ECTS, diversification and intensification. We also present theoretical proof to guarantee the global convergence of search with probability one. Experimental results have validated that the ECTS based approach can obtain global optimum efficiently, especially for large scale dimension of parameter. Potentially, the novel ECTS based algorithm can be applied in many applications of multiview geometry.

****

3D Sub-query Expansion for Improving Sketch-Based Multi-view Image Retrieval

Yen-Liang Lin, Cheng-Yu Huang, Hao-Jeng Wang, Winston Hsu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3495-3502

We propose a 3D sub-query expansion approach for boosting sketch-based multi-view image retrieval. The core idea of our method is to automatically convert two (

guided) 2D sketches into an approximated 3D sketch model, and then generate multi-view sketches as expanded sub-queries to improve the retrieval performance. To learn the weights among synthesized views (sub-queries), we present a new multi-query feature to model the similarity between subqueries and dataset images, and formulate it into a convex optimization problem. Our approach shows superior performance compared with the state-of-the-art approach on a public multi-view image dataset. Moreover, we also conduct sensitivity tests to analyze the parameters of our approach based on the gathered user sketches.
****

Tracking Revisited Using RGBD Camera: Unified Benchmark and Baselines
Shuran Song, Jianxiong Xiao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 233-240
Despite significant progress, tracking is still considered to be a very challenging task. Recently, the increasing popularity of depth sensors has made it possible to obtain reliable depth easily. This may be a game changer for tracking, since depth can be used to prevent model drift and handle occlusion. We also observe that current tracking algorithms are mostly evaluated on a very small number of videos collected and annotated by different groups. The lack of a reasonable size and consistently constructed benchmark has prevented a persuasive comparison among different algorithms. In this paper, we construct a unified benchmark dataset of 100 RGBD videos with high diversity, propose different kinds of RGBD tracking algorithms using 2D or 3D model, and present a quantitative comparison of various algorithms with RGB or RGBD input. We aim to lay the foundation for further research in both RGB and RGBD tracking, and our benchmark is available at http://tracking.cs.princeton.edu.
****

Modeling Self-Occlusions in Dynamic Shape and Appearance Tracking
Yanchao Yang, Ganesh Sundaramoorthi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 201-208
We present a method to track the precise shape of a dynamic object in video. Joint dynamic shape and appearance models, in which a template of the object is propagated to match the object shape and radiance in the next frame, are advantageous over methods employing global image statistics in cases of complex object radiance and cluttered background. In cases of complex 3D object motion and relative viewpoint change, self-occlusions and disocclusions of the object are prominent, and current methods employing joint shape and appearance models are unable to accurately adapt to new shape and appearance information, leading to inaccurate shape detection. In this work, we model self-occlusions and dis-occlusions in a joint shape and appearance tracking framework. Experiments on video exhibiting occlusion/dis-occlusion, complex radiance and background show that occlusion/dis-occlusion modeling leads to superior shape accuracy compared to recent methods employing joint shape/appearance models or employing global statistics.
****

Fingerspelling Recognition with Semi-Markov Conditional Random Fields
Taehwan Kim, Greg Shakhnarovich, Karen Livescu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1521-1528
Recognition of gesture sequences is in general a very difficult problem, but in certain domains the difficulty may be mitigated by exploiting the domain's "grammar". One such grammatically constrained gesture sequence domain is sign language. In this paper we investigate the case of fingerspelling recognition, which can be very challenging due to the quick, small motions of the fingers. Most prior work on this task has assumed a closed vocabulary of fingerspelled words; here we study the more natural open-vocabulary case, where the only domain knowledge is the possible fingerspelled letters and statistics of their sequences. We develop a semi-Markov conditional model approach, where feature functions are defined over segments of video and their corresponding letter labels. We use classifiers of letters and linguistic handshape features, along with expected motion profiles, to define segmental feature functions. This approach improves letter error rate (Levenshtein distance between hypothesized and correct letter sequences) from 16.3% using a hidden Markov model baseline to 11.6% using the proposed semi-

Markov model.
****
Attribute Dominance: What Pops Out?
Naman Turakhia, Devi Parikh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1225-1232
When we look at an image, some properties or attributes of the image stand out more than others. When describing an image, people are likely to describe these dominant attributes first. Attribute dominance is a result of a complex interplay between the various properties present or absent in the image. Which attributes in an image are more dominant than others reveals rich information about the content of the image. In this paper we tap into this information by modeling attribute dominance. We show that this helps improve the performance of vision systems on a variety of human-centric applications such as zero-shot learning, image search and generating textual descriptions of images.
****
Modifying the Memorability of Face Photographs
Aditya Khosla, Wilma A. Bainbridge, Antonio Torralba, Aude Oliva; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3200-3207
Contemporary life bombards us with many new images of faces every day, which poses non-trivial constraints on human memory. The vast majority of face photographs are intended to be remembered, either because of personal relevance, commercial interests or because the pictures were deliberately designed to be memorable. Can we make a portrait more memorable or more forgettable automatically? Here, we provide a method to modify the memorability of individual face photographs, while keeping the identity and other facial traits (e.g. age, attractiveness, and emotional magnitude) of the individual fixed. We show that face photographs manipulated to be more memorable (or more forgettable) are indeed more often remembered (or forgotten) in a crowd-sourcing experiment with an accuracy of 74%. Quantifying and modifying the 'memorability' of a face lends itself to many useful applications in computer vision and graphics, such as mnemonic aids for learning, photo editing applications for social networks and tools for designing memorable advertisements.
****
A Fully Hierarchical Approach for Finding Correspondences in Non-rigid Shapes
Ivan Sipiran, Benjamin Bustos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 817-824
This paper presents a hierarchical method for finding correspondences in non-rigid shapes. We propose a new representation for 3D meshes: the decomposition tree. This structure characterizes the recursive decomposition process of a mesh into regions of interest and keypoints. The internal nodes contain regions of interest (which may be recursively decomposed) and the leaf nodes contain the keypoints to be matched. We also propose a hierarchical matching algorithm that performs in a level-wise manner. The matching process is guided by the similarity between regions in high levels of the tree, until reaching the keypoints stored in the leaves. This allows us to reduce the search space of correspondences, making also the matching process efficient. We evaluate the effectiveness of our approach using the SHREC'2010 robust correspondence benchmark. In addition, we show that our results outperform the state of the art.
****
Fast Direct Super-Resolution by Simple Functions
Chih-Yuan Yang, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 561-568
The goal of single-image super-resolution is to generate a high-quality high-resolution image based on a given low-resolution input. It is an ill-posed problem which requires exemplars or priors to better reconstruct the missing high-resolution image details. In this paper, we propose to split the feature space into numerous subspaces and collect exemplars to learn priors for each subspace, thereby creating effective mapping functions. The use of split input space facilitates both feasibility of using simple functions for super-resolution, and efficiency

of generating highresolution results. High-quality high-resolution images are reconstructed based on the effective learned priors. Experimental results demonstrate that the proposed algorithm performs efficiently and effectively over state-of-the-art methods.
****

Tree Shape Priors with Connectivity Constraints Using Convex Relaxation on General Graphs
Jan Stuhmer, Peter Schroder, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2336-2343
We propose a novel method to include a connectivity prior into image segmentation that is based on a binary labeling of a directed graph, in this case a geodesic shortest path tree. Specifically we make two contributions: First, we construct a geodesic shortest path tree with a distance measure that is related to the image data and the bending energy of each path in the tree. Second, we include a connectivity prior in our segmentation model, that allows to segment not only a single elongated structure, but instead a whole connected branching tree. Because both our segmentation model and the connectivity constraint are convex, a global optimal solution can be found. To this end, we generalize a recent primal-dual algorithm for continuous convex optimization to an arbitrary graph structure. To validate our method we present results on data from medical imaging in angiography and retinal blood vessel segmentation.
****

Learning the Visual Interpretation of Sentences
C. L. Zitnick, Devi Parikh, Lucy Vanderwende; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1681-1688
Sentences that describe visual scenes contain a wide variety of information pertaining to the presence of objects, their attributes and their spatial relations. In this paper we learn the visual features that correspond to semantic phrases derived from sentences. Specifically, we extract predicate tuples that contain two nouns and a relation. The relation may take several forms, such as a verb, preposition, adjective or their combination. We model a scene using a Conditional Random Field (CRF) formulation where each node corresponds to an object, and the edges to their relations. We determine the potentials of the CRF using the tuples extracted from the sentences. We generate novel scenes depicting the sentences' visual meaning by sampling from the CRF. The CRF is also used to score a set of scenes for a text-based image retrieval task. Our results show we can generate (retrieve) scenes that convey the desired semantic meaning, even when scenes (queries) are described by multiple sentences. Significant improvement is found over several baseline approaches.
****

A Unified Probabilistic Approach Modeling Relationships between Attributes and Objects
Xiaoyang Wang, Qiang Ji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2120-2127
This paper proposes a unified probabilistic model to model the relationships between attributes and objects for attribute prediction and object recognition. As a list of semantically meaningful properties of objects, attributes generally relate to each other statistically. In this paper, we propose a unified probabilistic model to automatically discover and capture both the object-dependent and objectindependent attribute relationships. The model utilizes the captured relationships to benefit both attribute prediction and object recognition. Experiments on four benchmark attribute datasets demonstrate the effectiveness of the proposed unified model for improving attribute prediction as well as object recognition in both standard and zero-shot learning cases.
****

Domain Transfer Support Vector Ranking for Person Re-identification without Target Camera Label Information
Andy J. Ma, Pong C. Yuen, Jiawei Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3567-3574
This paper addresses a new person re-identification problem without the label in

formation of persons under non-overlapping target cameras. Given the matched (positive) and unmatched (negative) image pairs from source domain cameras, as well as unmatched (negative) image pairs which can be easily generated from target domain cameras, we propose a Domain Transfer Ranked Support Vector Machines (DTRSVM) method for re-identification under target domain cameras. To overcome the problems introduced due to the absence of matched (positive) image pairs in target domain, we relax the discriminative constraint to a necessary condition only relying on the positive mean in target domain. By estimating the target positive mean using source and target domain data, a new discriminative model with high confidence in target positive mean and low confidence in target negative image pairs is developed. Since the necessary condition may not truly preserve the discriminability, multi-task support vector ranking is proposed to incorporate the training data from source domain with label information. Experimental results show that the proposed DTRSVM outperforms existing methods without using label information in target cameras. And the top 30 rank accuracy can be improved by the proposed method upto 9.40% on publicly available person re-identification datasets.
****

Sparse Variation Dictionary Learning for Face Recognition with a Single Training Sample per Person
Meng Yang, Luc Van Gool, Lei Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 689-696
Face recognition (FR) with a single training sample per person (STSPP) is a very challenging problem due to the lack of information to predict the variations in the query sample. Sparse representation based classification has shown interesting results in robust FR; however, its performance will deteriorate much for FR with STSPP. To address this issue, in this paper we learn a sparse variation dictionary from a generic training set to improve the query sample representation by STSPP. Instead of learning from the generic training set independently w.r.t. the gallery set, the proposed sparse variation dictionary learning (SVDL) method is adaptive to the gallery set by jointly learning a projection to connect the generic training set with the gallery set. The learnt sparse variation dictionary can be easily integrated into the framework of sparse representation based classification so that various variations in face images, including illumination, expression, occlusion, pose, etc., can be better handled. Experiments on the large-scale CMU Multi-PIE, FRGC and LFW databases demonstrate the promising performance of SVDL on FR with STSPP.
****

Estimating the 3D Layout of Indoor Scenes and Its Clutter from Depth Sensors
Jian Zhang, Chen Kan, Alexander G. Schwing, Raquel Urtasun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1273-1280
In this paper we propose an approach to jointly estimate the layout of rooms as well as the clutter present in the scene using RGB-D data. Towards this goal, we propose an effective model that is able to exploit both depth and appearance features, which are complementary. Furthermore, our approach is efficient as we exploit the inherent decomposition of additive potentials. We demonstrate the effectiveness of our approach on the challenging NYU v2 dataset and show that employing depth reduces the layout error by 6% and the clutter estimation by 13%.
****

Learning to Share Latent Tasks for Action Recognition
Qiang Zhou, Gang Wang, Kui Jia, Qi Zhao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2264-2271
Sharing knowledge for multiple related machine learning tasks is an effective strategy to improve the generalization performance. In this paper, we investigate knowledge sharing across categories for action recognition in videos. The motivation is that many action categories are related, where common motion pattern are shared among them (e.g. diving and high jump share the jump motion). We propose a new multi-task learning method to learn latent tasks shared across categories, and reconstruct a classifier for each category from these latent tasks. Compared to previous methods, our approach has two advantages: (1) The learned latent tasks correspond to basic motion patterns instead of full actions, thus enhancin

g discrimination power of the classifiers. (2) Categories are selected to share information with a sparsity regularizer, avoiding falsely forcing all categories to share knowledge. Experimental results on multiple public data sets show that the proposed approach can effectively transfer knowledge between different action categories to improve the performance of conventional single task learning methods.
****

Space-Time Robust Representation for Action Recognition
Nicolas Ballas, Yi Yang, Zhen-Zhong Lan, Bertrand Delezoide, Francoise Preteux, Alexander Hauptmann; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2704-2711
We address the problem of action recognition in unconstrained videos. We propose a novel content driven pooling that leverages space-time context while being robust toward global space-time transformations. Being robust to such transformations is of primary importance in unconstrained videos where the action localizations can drastically shift between frames. Our pooling identifies regions of interest using video structural cues estimated by different saliency functions. To combine the different structural information, we introduce an iterative structure learning algorithm, WSVM (weighted SVM), that determines the optimal saliency layout of an action model through a sparse regularizer. A new optimization method is proposed to solve the WSVM' highly non-smooth objective function. We evaluate our approach on standard action datasets (KTH, UCF50 and HMDB). Most noticeably, the accuracy of our algorithm reaches 51.8% on the challenging HMDB dataset which outperforms the state-of-the-art of HHMMDrelatively.
****

Street View Motion-from-Structure-from-Motion
Bryan Klingner, David Martin, James Roseborough; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 953-960
We describe a structure-from-motion framework that handles "generalized" cameras, such as moving rollingshutter cameras, and works at an unprecedented scale-billions of images covering millions of linear kilometers of roads--by exploiting a good relative pose prior along vehicle paths. We exhibit a planet-scale, appearanceaugmented point cloud constructed with our framework and demonstrate its practical use in correcting the pose of a street-level image collection.
****

Quantize and Conquer: A Dimensionality-Recursive Solution to Clustering, Vector Quantization, and Image Retrieval
Yannis Avrithis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3024-3031
Inspired by the close relation between nearest neighbor search and clustering in high-dimensional spaces as well as the success of one helping to solve the other, we introduce a new paradigm where both problems are solved simultaneously. Our solution is recursive, not in the size of input data but in the number of dimensions. One result is a clustering algorithm that is tuned to small codebooks but does not need all data in memory at the same time and is practically constant in the data size. As a by-product, a tree structure performs either exact or approximate quantization on trained centroids, the latter being not very precise but extremely fast. A lesser contribution is a new indexing scheme for image retrieval that exploits multiple small codebooks to provide an arbitrarily fine partition of the descriptor space. Large scale experiments on public datasets exhibit state of the art performance and remarkable generalization.
****

Dynamic Pooling for Complex Event Recognition
Weixin Li, Qian Yu, Ajay Divakaran, Nuno Vasconcelos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2728-2735
The problem of adaptively selecting pooling regions for the classification of complex video events is considered. Complex events are defined as events composed of several characteristic behaviors, whose temporal configuration can change from sequence to sequence. A dynamic pooling operator is defined so as to enable a unified solution to the problems of event specific video segmentation, temporal

structure modeling, and event detection. Video is decomposed into segments, and the segments most informative for detecting a given event are identified, so as to dynamically determine the pooling operator most suited for each sequence. Thi s dynamic pooling is implemented by treating the locations of characteristic seg ments as hidden information, which is inferred, on a sequence-by-sequence basis, via a large-margin classification rule with latent variables. Although the feas ible set of segment selections is combinatorial, it is shown that a globally opt imal solution to the inference problem can be obtained efficiently, through the solution of a series of linear programs. Besides the coarselevel location of seg ments, a finer model of video structure is implemented by jointly pooling featur es of segmenttuples. Experimental evaluation demonstrates that the resulting eve nt detector has state-of-the-art performance on challenging video datasets.
****

A Practical Transfer Learning Algorithm for Face Verification
Xudong Cao, David Wipf, Fang Wen, Genquan Duan, Jian Sun; Proceedings of the IEE E International Conference on Computer Vision (ICCV), 2013, pp. 3208-3215
Face verification involves determining whether a pair of facial images belongs t o the same or different subjects. This problem can prove to be quite challenging in many important applications where labeled training data is scarce, e.g., fam ily album photo organization software. Herein we propose a principled transfer l earning approach for merging plentiful source-domain data with limited samples f rom some target domain of interest to create a classifier that ideally performs nearly as well as if rich target-domain data were present. Based upon a surprisi ngly simple generative Bayesian model, our approach combines a KL-divergencebase d regularizer/prior with a robust likelihood function leading to a scalable impl ementation via the EM algorithm. As justification for our design choices, we lat er use principles from convex analysis to recast our algorithm as an equivalent structured rank minimization problem leading to a number of interesting insights related to solution structure and feature-transform invariance. These insights help to both explain the effectiveness of our algorithm as well as elucidate a w ide variety of related Bayesian approaches. Experimental testing with challengin g datasets validate the utility of the proposed algorithm.
****

Incorporating Cloud Distribution in Sky Representation
Kuan-Chuan Peng, Tsuhan Chen; Proceedings of the IEEE International Conference o n Computer Vision (ICCV), 2013, pp. 2152-2159
Most sky models only describe the cloudiness of the overall sky by a single cate gory or parameter such as sky index, which does not account for the distribution of the clouds across the sky. To capture variable cloudiness, we extend the con cept of sky index to a random field indicating the level of cloudiness of each s ky pixel in our proposed sky representation based on the Igawa sky model. We for mulate the problem of solving the sky index of every sky pixel as a labeling pro blem, where an approximate solution can be efficiently found. Experimental resul ts show that our proposed sky model has better expressiveness, stability with re spect to variation in camera parameters, and geo-location estimation in outdoor images compared to the uniform sky index model. Potential applications of our pr oposed sky model include sky image rendering, where sky images can be generated with an arbitrary cloud distribution at any time and any location, previously im possible with traditional sky models.
****

From Actemes to Action: A Strongly-Supervised Representation for Detailed Action Understanding
Weiyu Zhang, Menglong Zhu, Konstantinos G. Derpanis; Proceedings of the IEEE Int ernational Conference on Computer Vision (ICCV), 2013, pp. 2248-2255
This paper presents a novel approach for analyzing human actions in non-scripted , unconstrained video settings based on volumetric, x-y-t, patch classifiers, te rmed actemes. Unlike previous action-related work, the discovery of patch classi fiers is posed as a strongly-supervised process. Specifically, keypoint labels (e.g., position) across spacetime are used in a data-driven training process to d iscover patches that are highly clustered in the spacetime keypoint configuratio

n space. To support this process, a new human action dataset consisting of chall
enging consumer videos is introduced, where notably the action label, the 2D pos
ition of a set of keypoints and their visibilities are provided for each video f
rame. On a novel input video, each acteme is used in a sliding volume scheme to
yield a set of sparse, non-overlapping detections. These detecsseddeetecctions p
rovide the intermediate substrate for segmeegmenatot the action. For action clas
sification, the proposed representation shows significant improvement over state
-of-the-art low-level features, while providing spatiotemporal localization as a
dditional output. This output sheds further light into detailed action understan
ding.
****

Distributed Low-Rank Subspace Segmentation
Ameet Talwalkar, Lester Mackey, Yadong Mu, Shih-Fu Chang, Michael I. Jordan; Pro
ceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, p
p. 3543-3550
Vision problems ranging from image clustering to motion segmentation to semi-sup
ervised learning can naturally be framed as subspace segmentation problems, in w
hich one aims to recover multiple low-dimensional subspaces from noisy and corru
pted input data. Low-Rank Representation (LRR), a convex formulation of the subs
pace segmentation problem, is provably and empirically accurate on small problem
s but does not scale to the massive sizes of modern vision datasets. Moreover, p
ast work aimed at scaling up low-rank matrix factorization is not applicable to
LRR given its non-decomposable constraints. In this work, we propose a novel div
ide-and-conquer algorithm for large-scale subspace segmentation that can cope wi
th LRR's non-decomposable constraints and maintains LRR's strong recovery guaran
tees. This has immediate implications for the scalability of subspace segmentati
on, which we demonstrate on a benchmark face recognition dataset and in simulati
ons. We then introduce novel applications of LRR-based subspace segmentation to
large-scale semisupervised learning for multimedia event detection, concept dete
ction, and image tagging. In each case, we obtain stateof-the-art results and or
der-of-magnitude speed ups.
****

Modeling 4D Human-Object Interactions for Event and Object Recognition
Ping Wei, Yibiao Zhao, Nanning Zheng, Song-Chun Zhu; Proceedings of the IEEE Int
ernational Conference on Computer Vision (ICCV), 2013, pp. 3272-3279
Recognizing the events and objects in the video sequence are two challenging tas
ks due to the complex temporal structures and the large appearance variations. I
n this paper, we propose a 4D human-object interaction model, where the two task
s jointly boost each other. Our human-object interaction is defined in 4D space:
 i) the cooccurrence and geometric constraints of human pose and object in 3D sp
ace; ii) the sub-events transition and objects coherence in 1D temporal dimensio
n. We represent the structure of events, sub-events and objects in a hierarchica
l graph. For an input RGB-depth video, we design a dynamic programming beam sear
ch algorithm to: i) segment the video, ii) recognize the events, and iii) detect
 the objects simultaneously. For evaluation, we built a large-scale multiview 3D
 event dataset which contains 3815 video sequences and 383,036 RGBD frames captu
red by the Kinect cameras. The experiment results on this dataset show the effec
tiveness of our method.
****

Codemaps - Segment, Classify and Search Objects Locally
Zhenyang Li, Efstratios Gavves, Koen E.A. van de Sande, Cees G.M. Snoek, Arnold
W.M. Smeulders; Proceedings of the IEEE International Conference on Computer Vis
ion (ICCV), 2013, pp. 2136-2143
In this paper we aim for segmentation and classification of objects. We propose
codemaps that are a joint formulation of the classification score and the local
neighborhood it belongs to in the image. We obtain the codemap by reordering the
 encoding, pooling and classification steps over lattice elements. Other than ex
isting linear decompositions who emphasize only the efficiency benefits for loca
lized search, we make three novel contributions. As a preliminary, we provide a
theoretical generalization of the sufficient mathematical conditions under which

image encodings and classification becomes locally decomposable. As first novelty we introduce l 2 normalization for arbitrarily shaped image regions, which is fast enough for semantic segmentation using our Fisher codemaps. Second, using the same lattice across images, we propose kernel pooling which embeds nonlinearities into codemaps for object classification by explicit or approximate feature mappings. Results demonstrate that l 2 normalized Fisher codemaps improve the state-of-the-art in semantic segmentation for PASCAL VOC. For object classification the addition of nonlinearities brings us on par with the state-of-the-art, but is 3x faster. Because of the codemaps' inherent efficiency, we can reach significant speed-ups for localized search as well. We exploit the efficiency gain for our third novelty: object segment retrieval using a single query image only.

****

Bounded Labeling Function for Global Segmentation of Multi-part Objects with Geometric Constraints

Masoud S. Nosrati, Shawn Andrews, Ghassan Hamarneh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2032-2039

The inclusion of shape and appearance priors have proven useful for obtaining more accurate and plausible segmentations, especially for complex objects with multiple parts. In this paper, we augment the popular MumfordShah model to incorporate two important geometrical constraints, termed containment and detachment, between different regions with a specified minimum distance between their boundaries. Our method is able to handle multiple instances of multi-part objects defined by these geometrical constraints using a single labeling function while maintaining global optimality. We demonstrate the utility and advantages of these two constraints and show that the proposed convex continuous method is superior to other state-of-theart methods, including its discrete counterpart, in terms of memory usage, and metrication errors.

****

Recognizing Text with Perspective Distortion in Natural Scenes

Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, Chew Lim Tan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 569-576

This paper presents an approach to text recognition in natural scene images. Unlike most existing works which assume that texts are horizontal and frontal parallel to the image plane, our method is able to recognize perspective texts of arbitrary orientations. For individual character recognition, we adopt a bag-of-keypoints approach, in which Scale Invariant Feature Transform (SIFT) descriptors are extracted densely and quantized using a pre-trained vocabulary. Following [1, 2], the context information is utilized through lexicons. We formulate word recognition as finding the optimal alignment between the set of characters and the list of lexicon words. Furthermore, we introduce a new dataset called StreetView Text-Perspective, which contains texts in street images with a great variety of viewpoints. Experimental results on public datasets and the proposed dataset show that our method significantly outperforms the state-of-the-art on perspective texts of arbitrary orientations.

****

Unifying Nuclear Norm and Bilinear Factorization Approaches for Low-Rank Matrix Decomposition

Ricardo Cabral, Fernando De La Torre, Joao P. Costeira, Alexandre Bernardino; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2488-2495

Low rank models have been widely used for the representation of shape, appearance or motion in computer vision problems. Traditional approaches to fit low rank models make use of an explicit bilinear factorization. These approaches benefit from fast numerical methods for optimization and easy kernelization. However, they suffer from serious local minima problems depending on the loss function and the amount/type of missing data. Recently, these lowrank models have alternatively been formulated as convex problems using the nuclear norm regularizer; unlike factorization methods, their numerical solvers are slow and it is unclear how to kernelize them or to impose a rank a priori. This paper proposes a unified app

roach to bilinear factorization and nuclear norm regularization, that inherits t
he benefits of both. We analyze the conditions under which these approaches are
equivalent. Moreover, based on this analysis, we propose a new optimization algo
rithm and a "rank continuation" strategy that outperform state-of-theart approac
hes for Robust PCA, Structure from Motion and Photometric Stereo with outliers a
nd missing data.
****

Efficient 3D Scene Labeling Using Fields of Trees
Olaf Kahler, Ian Reid; Proceedings of the IEEE International Conference on Compu
ter Vision (ICCV), 2013, pp. 3064-3071
We address the problem of 3D scene labeling in a structured learning framework.
Unlike previous work which uses structured Support Vector Machines, we employ th
e recently described Decision Tree Field and Regression Tree Field frameworks, w
hich learn the unary and binary terms of a Conditional Random Field from trainin
g data. We show this has significant advantages in terms of inference speed, whi
le maintaining similar accuracy. We also demonstrate empirically the importance
for overall labeling accuracy of features that make use of prior knowledge about
 the coarse scene layout such as the location of the ground plane. We show how t
his coarse layout can be estimated by our framework automatically, and that this
 information can be used to bootstrap improved accuracy in the detailed labeling
.
****

Efficient Hand Pose Estimation from a Single Depth Image
Chi Xu, Li Cheng; Proceedings of the IEEE International Conference on Computer V
ision (ICCV), 2013, pp. 3456-3462
We tackle the practical problem of hand pose estimation from a single noisy dept
h image. A dedicated three-step pipeline is proposed: Initial estimation step pr
ovides an initial estimation of the hand in-plane orientation and 3D location; C
andidate generation step produces a set of 3D pose candidate from the Hough voti
ng space with the help of the rotational invariant depth features; Verification
step delivers the final 3D hand pose as the solution to an optimization problem.
 We analyze the depth noises, and suggest tips to minimize their negative impact
s on the overall performance. Our approach is able to work with Kinecttype noisy
 depth images, and reliably produces pose estimations of general motions efficie
ntly (12 frames per second). Extensive experiments are conducted to qualitativel
y and quantitatively evaluate the performance with respect to the state-of-the-a
rt methods that have access to additional RGB images. Our approach is shown to d
eliver on par or even better results.
****

Supervised Binary Hash Code Learning with Jensen Shannon Divergence
Lixin Fan; Proceedings of the IEEE International Conference on Computer Vision (
ICCV), 2013, pp. 2616-2623
This paper proposes to learn binary hash codes within a statistical learning fra
mework, in which an upper bound of the probability of Bayes decision errors is d
erived for different forms of hash functions and a rigorous proof of the converg
ence of the upper bound is presented. Consequently, minimizing such an upper bou
nd leads to consistent performance improvements of existing hash code learning a
lgorithms, regardless of whether original algorithms are unsupervised or supervi
sed. This paper also illustrates a fast hash coding method that exploits simple
binary tests to achieve orders of magnitude improvement in coding speed as compa
red to projection based methods.
****

Internet Based Morphable Model
Ira Kemelmacher-Shlizerman; Proceedings of the IEEE International Conference on
Computer Vision (ICCV), 2013, pp. 3256-3263
In this paper we present a new concept of building a morphable model directly fr
om photos on the Internet. Morphable models have shown very impressive results m
ore than a decade ago, and could potentially have a huge impact on all aspects o
f face modeling and recognition. One of the challenges, however, is to capture a
nd register 3D laser scans of large number of people and facial expressions. Now

adays, there are enormous amounts of face photos on the Internet, large portion of which has semantic labels. We propose a framework to build a morphable model directly from photos, the framework includes dense registration of Internet photos, as well as, new single view shape reconstruction and modification algorithms.

****

Curvature-Aware Regularization on Riemannian Submanifolds

Kwang In Kim, James Tompkin, Christian Theobalt; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 881-888

One fundamental assumption in object recognition as well as in other computer vision and pattern recognition problems is that the data generation process lies on a manifold and that it respects the intrinsic geometry of the manifold. This assumption is held in several successful algorithms for diffusion and regularization, in particular, in graph-Laplacian-based algorithms. We claim that the performance of existing algorithms can be improved if we additionally account for how the manifold is embedded within the ambient space, i.e., if we consider the extrinsic geometry of the manifold. We present a procedure for characterizing the extrinsic (as well as intrinsic) curvature of a manifold M which is described by a sampled point cloud in a high-dimensional Euclidean space. Once estimated, we use this characterization in general diffusion and regularization on M , and form a new regularizer on a point cloud. The resulting re-weighted graph Laplacian demonstrates superior performance over classical graph Laplacian in semisupervised learning and spectral clustering.

****

From Subcategories to Visual Composites: A Multi-level Framework for Object Detection

Tian Lan, Michalis Raptis, Leonid Sigal, Greg Mori; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 369-376

The appearance of an object changes profoundly with pose, camera view and interactions of the object with other objects in the scene. This makes it challenging to learn detectors based on an object-level label (e.g., "car"). We postulate that having a richer set of labelings (at different levels of granularity) for an object, including finer-grained subcategories, consistent in appearance and view, and higherorder composites contextual groupings of objects consistent in their spatial layout and appearance, can significantly alleviate these problems. However, obtaining such a rich set of annotations, including annotation of an exponentially growing set of object groupings, is simply not feasible. We propose a weakly-supervised framework for object detection where we discover subcategories and the composites automatically with only traditional object-level category labels as input. To this end, we first propose an exemplar-SVM-based clustering approach, with latent SVM refinement, that discovers a variable length set of discriminative subcategories for each object class. We then develop a structured model for object detection that captures interactions among object subcategories and automatically discovers semantically meaningful and discriminatively relevant visual composites. We show that this model produces state-of-the-art performance on UIUC phrase object detection benchmark.

****

A Generalized Low-Rank Appearance Model for Spatio-temporally Correlated Rain Streaks

Yi-Lei Chen, Chiou-Ting Hsu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1968-1975

In this paper, we propose a novel low-rank appearance model for removing rain streaks. Different from previous work, our method needs neither rain pixel detection nor time-consuming dictionary learning stage. Instead, as rain streaks usually reveal similar and repeated patterns on imaging scene, we propose and generalize a low-rank model from matrix to tensor structure in order to capture the spatio-temporally correlated rain streaks. With the appearance model, we thus remove rain streaks from image/video (and also other high-order image structure) in a unified way. Our experimental results demonstrate competitive (or even better) visual quality and efficient run-time in comparison with state of the art.

****
Parallel Transport of Deformations in Shape Space of Elastic Surfaces
Qian Xie, Sebastian Kurtek, Huiling Le, Anuj Srivastava; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 865-872
Statistical shape analysis develops methods for comparisons, deformations, summarizations, and modeling of shapes in given data sets. These tasks require a fundamental tool called parallel transport of tangent vectors along arbitrary paths. This tool is essential for: (1) computation of geodesic paths using either shooting or path-straightening method, (2) transferring deformations across objects, and (3) modeling of statistical variability in shapes. Using the square-root normal field (SRNF) representation of parameterized surfaces, we present a method for transporting deformations along paths in the shape space. This is difficult despite the underlying space being a vector space because the chosen (elastic) Riemannian metric is non-standard. Using a finite-basis for representing SRNFs of shapes, we derive expressions for Christoffel symbols that enable parallel transports. We demonstrate this framework using examples from shape analysis of parameterized spherical surfaces, in the three contexts mentioned above.
****
Human Attribute Recognition by Rich Appearance Dictionary
Jungseock Joo, Shuo Wang, Song-Chun Zhu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 721-728
We present a part-based approach to the problem of human attribute recognition from a single image of a human body. To recognize the attributes of human from the body parts, it is important to reliably detect the parts. This is a challenging task due to the geometric variation such as articulation and view-point changes as well as the appearance variation of the parts arisen from versatile clothing types. The prior works have primarily focused on handling geometric variation by relying on pre-trained part detectors or pose estimators, which require manual part annotation, but the appearance variation has been relatively neglected in these works. This paper explores the importance of the appearance variation, which is directly related to the main task, attribute recognition. To this end, we propose to learn a rich appearance part dictionary of human with significantly less supervision by decomposing image lattice into overlapping windows at multiscale and iteratively refining local appearance templates. We also present quantitative results in which our proposed method outperforms the existing approaches.
****
Bayesian Joint Topic Modelling for Weakly Supervised Object Localisation
Zhiyuan Shi, Timothy M. Hospedales, Tao Xiang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2984-2991
We address the problem of localisation of objects as bounding boxes in images with weak labels. This weakly supervised object localisation problem has been tackled in the past using discriminative models where each object class is localised independently from other classes. We propose a novel framework based on Bayesian joint topic modelling. Our framework has three distinctive advantages over previous works: (1) All object classes and image backgrounds are modelled jointly together in a single generative model so that "explaining away" inference can resolve ambiguity and lead to better learning and localisation. (2) The Bayesian formulation of the model enables easy integration of prior knowledge about object appearance to compensate for limited supervision. (3) Our model can be learned with a mixture of weakly labelled and unlabelled data, allowing the large volume of unlabelled images on the Internet to be exploited for learning. Extensive experiments on the challenging VOC dataset demonstrate that our approach outperforms the state-of-the-art competitors.
****
Frustratingly Easy NBNN Domain Adaptation
Tatiana Tommasi, Barbara Caputo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 897-904
Over the last years, several authors have signaled that state of the art categorization methods fail to perform well when trained and tested on data from different databases. The general consensus in the literature is that this issue, known

as domain adaptation and/or dataset bias, is due to a distribution mismatch between data collections. Methods addressing it go from max-margin classifiers to learning how to modify the features and obtain a more robust representation. The large majority of these works use BOW feature descriptors, and learning methods based on imageto-image distance functions. Following the seminal work of [6], in this paper we challenge these two assumptions. We experimentally show that using the NBNN classifier over existing domain adaptation databases achieves always very strong performances. We build on this result, and present an NBNN-based domain adaptation algorithm that learns iteratively a class metric while inducing, for each sample, a large margin separation among classes. To the best of our knowledge, this is the first work casting the domain adaptation problem within the NBNN framework. Experiments show that our method achieves the state of the art, both in the unsupervised and semi-supervised settings.
****

Beyond Hard Negative Mining: Efficient Detector Learning via Block-Circulant Decomposition
Joao F. Henriques, Joao Carreira, Rui Caseiro, Jorge Batista; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2760-2767
Competitive sliding window detectors require vast training sets. Since a pool of natural images provides a nearly endless supply of negative samples, in the form of patches at different scales and locations, training with all the available data is considered impractical. A staple of current approaches is hard negative mining, a method of selecting relevant samples, which is nevertheless expensive. Given that samples at slightly different locations have overlapping support, there seems to be an enormous amount of duplicated work. It is natural, then, to ask whether these redundancies can be eliminated. In this paper, we show that the Gram matrix describing such data is block-circulant. We derive a transformation based on the Fourier transform that block-diagonalizes the Gram matrix, at once eliminating redundancies and partitioning the learning problem. This decomposition is valid for any dense features and several learning algorithms, and takes full advantage of modern parallel architectures. Surprisingly, it allows training with all the potential samples in sets of thousands of images. By considering the full set, we generate in a single shot the optimal solution, which is usually obtained only after several rounds of hard negative mining. We report speed gains on Caltech Pedestrians and INRIA Pedestrians of over an order of magnitude, allowing training on a desktop computer in a couple of minutes.
****

Cross-Field Joint Image Restoration via Scale Map
Qiong Yan, Xiaoyong Shen, Li Xu, Shaojie Zhuo, Xiaopeng Zhang, Liang Shen, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1537-1544
Color, infrared, and flash images captured in different fields can be employed to effectively eliminate noise and other visual artifacts. We propose a two-image restoration framework considering input images in different fields, for example, one noisy color image and one dark-flashed nearinfrared image. The major issue in such a framework is to handle structure divergence and find commonly usable edges and smooth transition for visually compelling image reconstruction. We introduce a scale map as a competent representation to explicitly model derivative-level confidence and propose new functions and a numerical solver to effectively infer it following new structural observations. Our method is general and shows a principled way for cross-field restoration.
****

STAR3D: Simultaneous Tracking and Reconstruction of 3D Objects Using RGB-D Data
Carl Yuheng Ren, Victor Prisacariu, David Murray, Ian Reid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1561-1568
We introduce a probabilistic framework for simultaneous tracking and reconstruction of 3D rigid objects using an RGB-D camera. The tracking problem is handled using a bag-of-pixels representation and a back-projection scheme. Surface and background appearance models are learned online, leading to robust tracking in the presence of heavy occlusion and outliers. In both our tracking and reconstructi

on modules, the 3D object is implicitly embedded using a 3D level-set function. The framework is initialized with a simple shape primitive model (e.g. a sphere or a cube), and the real 3D object shape is tracked and reconstructed online. Unlike existing depth-based 3D reconstruction works, which either rely on calibrated/fixed camera set up or use the observed world map to track the depth camera, our framework can simultaneously track and reconstruct small moving objects. We use both qualitative and quantitative results to demonstrate the superior performance of both tracking and reconstruction of our method.
****

Detecting Irregular Curvilinear Structures in Gray Scale and Color Imagery Using Multi-directional Oriented Flux

Engin Turetken, Carlos Becker, Przemyslaw Glowacki, Fethallah Benmansour, Pascal Fua; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1553-1560

We propose a new approach to detecting irregular curvilinear structures in noisy image stacks. In contrast to earlier approaches that rely on circular models of the crosssections, ours allows for the arbitrarily-shaped ones that are prevalent in biological imagery. This is achieved by maximizing the image gradient flux along multiple directions and radii, instead of only two with a unique radius as is usually done. This yields a more complex optimization problem for which we propose a computationally efficient solution. We demonstrate the effectiveness of our approach on a wide range of challenging gray scale and color datasets and show that it outperforms existing techniques, especially on very irregular structures.
****

Learning Slow Features for Behaviour Analysis

Lazaros Zafeiriou, Mihalis A. Nicolaou, Stefanos Zafeiriou, Symeon Nikitidis, Maja Pantic; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2840-2847

A recently introduced latent feature learning technique for time varying dynamic phenomena analysis is the socalled Slow Feature Analysis (SFA). SFA is a deterministic component analysis technique for multi-dimensional sequences that by minimizing the variance of the first order time derivative approximation of the input signal finds uncorrelated projections that extract slowly-varying features ordered by their temporal consistency and constancy. In this paper, we propose a number of extensions in both the deterministic and the probabilistic SFA optimization frameworks. In particular, we derive a novel deterministic SFA algorithm that is able to identify linear projections that extract the common slowest varying features of two or more sequences. In addition, we propose an Expectation Maximization (EM) algorithm to perform inference in a probabilistic formulation of SFA and similarly extend it in order to handle two and more time varying data sequences. Moreover, we demonstrate that the probabilistic SFA (EMSFA) algorithm that discovers the common slowest varying latent space of multiple sequences can be combined with dynamic time warping techniques for robust sequence timealignment. The proposed SFA algorithms were applied for facial behavior analysis demonstrating their usefulness and appropriateness for this task.
****

Multi-view Object Segmentation in Space and Time

Abdelaziz Djelouah, Jean-Sebastien Franco, Edmond Boyer, Francois Le Clerc, Patrick Perez; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2640-2647

In this paper, we address the problem of object segmentation in multiple views or videos when two or more viewpoints of the same scene are available. We propose a new approach that propagates segmentation coherence information in both space and time, hence allowing evidences in one image to be shared over the complete set. To this aim the segmentation is cast as a single efficient labeling problem over space and time with graph cuts. In contrast to most existing multi-view segmentation methods that rely on some form of dense reconstruction, ours only requires a sparse 3D sampling to propagate information between viewpoints. The approach is thoroughly evaluated on standard multiview datasets, as well as on video

s. With static views, results compete with state of the art methods but they are achieved with significantly fewer viewpoints. With multiple videos, we report results that demonstrate the benefit of segmentation propagation through temporal cues.

****

Non-convex P-Norm Projection for Robust Sparsity
Mithun Das Gupta, Sanjeev Kumar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1593-1600
In this paper, we investigate the properties of L p norm (p Ittiswithin a projection framework. We start with the KKT equations of the non-linear optimization problem and then use its key properties to arrive at an algorithm for L p norm projection on the non-negative simplex. We compare with L 1 projection which needs prior knowledge of the true norm, as well as hard thresholding based sparsification proposed in recent compressed sensing literature. We show performance improvements compared to these techniques across different vision applications.

****

Robust Matrix Factorization with Unknown Noise
Deyu Meng, Fernando De La Torre; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1337-1344
Many problems in computer vision can be posed as recovering a low-dimensional subspace from highdimensional visual data. Factorization approaches to lowrank subspace estimation minimize a loss function between an observed measurement matrix and a bilinear factorization. Most popular loss functions include the L 2 and L 1 losses. L 2 is optimal for Gaussian noise, while L 1 is for Laplacian distributed noise. However, real data is often corrupted by an unknown noise distribution, which is unlikely to be purely Gaussian or Laplacian. To address this problem, this paper proposes a low-rank matrix factorization problem with a Mixture of Gaussians (MoG) noise model. The MoG model is a universal approximator for any continuous distribution, and hence is able to model a wider range of noise distributions. The parameters of the MoG model can be estimated with a maximum likelihood method, while the subspace is computed with standard approaches. We illustrate the benefits of our approach in extensive synthetic and real-world experiments including structure from motion, face modeling and background subtraction.

****

A General Two-Step Approach to Learning-Based Hashing
Guosheng Lin, Chunhua Shen, David Suter, Anton van den Hengel; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2552-2559
Most existing approaches to hashing apply a single form of hash function, and an optimization process which is typically deeply coupled to this specific form. This tight coupling restricts the flexibility of the method to respond to the data, and can result in complex optimization problems that are difficult to solve. Here we propose a flexible yet simple framework that is able to accommodate different types of loss functions and hash functions. This framework allows a number of existing approaches to hashing to be placed in context, and simplifies the development of new problemspecific hashing methods. Our framework decomposes the hashing learning problem into two steps: hash bit learning and hash function learning based on the learned bits. The first step can typically be formulated as binary quadratic problems, and the second step can be accomplished by training standard binary classifiers. Both problems have been extensively studied in the literature. Our extensive experiments demonstrate that the proposed framework is effective, flexible and outperforms the state-of-the-art.

****

Dynamic Scene Deblurring
Tae Hyun Kim, Byeongjoo Ahn, Kyoung Mu Lee; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3160-3167
Most conventional single image deblurring methods assume that the underlying scene is static and the blur is caused by only camera shake. In this paper, in contrast to this restrictive assumption, we address the deblurring problem of general dynamic scenes which contain multiple moving objects as well as camera shake. In case of dynamic scenes, moving objects and background have different blur mot

ions, so the segmentation of the motion blur is required for deblurring each dis
tinct blur motion accurately. Thus, we propose a novel energy model designed wit
h the weighted sum of multiple blur data models, which estimates different motio
n blurs and their associated pixelwise weights, and resulting sharp image. In th
is framework, the local weights are determined adaptively and get high values wh
en the corresponding data models have high data fidelity. And, the weight inform
ation is used for the segmentation of the motion blur. Non-local regularization
of weights are also incorporated to produce more reliable segmentation results.
A convex optimization-based method is used for the solution of the proposed ener
gy model. Experimental results demonstrate that our method outperforms conventio
nal approaches in deblurring both dynamic scenes and static scenes.
****

Paper Doll Parsing: Retrieving Similar Styles to Parse Clothing Items
Kota Yamaguchi, M. Hadi Kiapour, Tamara L. Berg; Proceedings of the IEEE Interna
tional Conference on Computer Vision (ICCV), 2013, pp. 3519-3526
Clothing recognition is an extremely challenging problem due to wide variation i
n clothing item appearance, layering, and style. In this paper, we tackle the cl
othing parsing problem using a retrieval based approach. For a query image, we f
ind similar styles from a large database of tagged fashion images and use these
examples to parse the query. Our approach combines parsing from: pre-trained glo
bal clothing models, local clothing models learned on the fly from retrieved exa
mples, and transferred parse masks (paper doll item transfer) from retrieved exa
mples. Experimental evaluation shows that our approach significantly outperforms
state of the art in parsing accuracy.
****

Deep Learning Identity-Preserving Face Space
Zhenyao Zhu, Ping Luo, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE Inter
national Conference on Computer Vision (ICCV), 2013, pp. 113-120
Face recognition with large pose and illumination variations is a challenging pr
oblem in computer vision. This paper addresses this challenge by proposing a new
learningbased face representation: the face identity-preserving (FIP) features.
Unlike conventional face descriptors, the FIP features can significantly reduce
intra-identity variances, while maintaining discriminativeness between identiti
es. Moreover, the FIP features extracted from an image under any pose and illumi
nation can be used to reconstruct its face image in the canonical view. This pro
perty makes it possible to improve the performance of traditional descriptors, s
uch as LBP [2] and Gabor [31], which can be extracted from our reconstructed ima
ges in the canonical view to eliminate variations. In order to learn the FIP fea
tures, we carefully design a deep network that combines the feature extraction l
ayers and the reconstruction layer. The former encodes a face image into the FIP
features, while the latter transforms them to an image in the canonical view. E
xtensive experiments on the large MultiPIE face database [7] demonstrate that it
significantly outperforms the state-of-the-art face recognition methods.
****

Predicting Primary Gaze Behavior Using Social Saliency Fields
Hyun Soo Park, Eakta Jain, Yaser Sheikh; Proceedings of the IEEE International C
onference on Computer Vision (ICCV), 2013, pp. 3503-3510
We present a method to predict primary gaze behavior in a social scene. Inspired
by the study of electric fields, we posit "social charges"--latent quantities t
hat drive the primary gaze behavior of members of a social group. These charges
induce a gradient field that defines the relationship between the social charges
and the primary gaze direction of members in the scene. This field model is use
d to predict primary gaze behavior at any location or time in the scene. We pres
ent an algorithm to estimate the time-varying behavior of these charges from the
primary gaze behavior of measured observers in the scene. We validate the model
by evaluating its predictive precision via cross-validation in a variety of soc
ial scenes.
****

A Flexible Scene Representation for 3D Reconstruction Using an RGB-D Camera
Diego Thomas, Akihiro Sugimoto; Proceedings of the IEEE International Conference

on Computer Vision (ICCV), 2013, pp. 2800-2807
Updating a global 3D model with live RGB-D measurements has proven to be successful for 3D reconstruction of indoor scenes. Recently, a Truncated Signed Distance Function (TSDF) volumetric model and a fusion algorithm have been introduced (KinectFusion), showing significant advantages such as computational speed and accuracy of the reconstructed scene. This algorithm, however, is expensive in memory when constructing and updating the global model. As a consequence, the method is not well scalable to large scenes. We propose a new flexible 3D scene representation using a set of planes that is cheap in memory use and, nevertheless, achieves accurate reconstruction of indoor scenes from RGB-D image sequences. Projecting the scene onto different planes reduces significantly the size of the scene representation and thus it allows us to generate a global textured 3D model with lower memory requirement while keeping accuracy and easiness to update with live RGB-D measurements. Experimental results demonstrate that our proposed flexible 3D scene representation achieves accurate reconstruction, while keeping the scalability for large indoor scenes.
****

Multi-view Normal Field Integration for 3D Reconstruction of Mirroring Objects
Michael Weinmann, Aljosa Osep, Roland Ruiters, Reinhard Klein; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2504-2511
In this paper, we present a novel, robust multi-view normal field integration technique for reconstructing the full 3D shape of mirroring objects. We employ a turntablebased setup with several cameras and displays. These are used to display illumination patterns which are reflected by the object surface. The pattern information observed in the cameras enables the calculation of individual volumetric normal fields for each combination of camera, display and turntable angle. As the pattern information might be blurred depending on the surface curvature or due to nonperfect mirroring surface characteristics, we locally adapt the decoding to the finest still resolvable pattern resolution. In complex real-world scenarios, the normal fields contain regions without observations due to occlusions and outliers due to interreflections and noise. Therefore, a robust reconstruction using only normal information is challenging. Via a non-parametric clustering of normal hypotheses derived for each point in the scene, we obtain both the most likely local surface normal and a local surface consistency estimate. This information is utilized in an iterative mincut based variational approach to reconstruct the surface geometry.
****

Exemplar-Based Graph Matching for Robust Facial Landmark Localization
Feng Zhou, Jonathan Brandt, Zhe Lin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1025-1032
Localizing facial landmarks is a fundamental step in facial image analysis. However, the problem is still challenging due to the large variability in pose and appearance, and the existence of occlusions in real-world face images. In this paper, we present exemplar-based graph matching (EGM), a robust framework for facial landmark localization. Compared to conventional algorithms, EGM has three advantages: (1) an affine-invariant shape constraint is learned online from similar exemplars to better adapt to the test face; (2) the optimal landmark configuration can be directly obtained by solving a graph matching problem with the learned shape constraint; (3) the graph matching problem can be optimized efficiently by linear programming. To our best knowledge, this is the first attempt to apply a graph matching technique for facial landmark localization. Experiments on several challenging datasets demonstrate the advantages of EGM over state-of-the-art methods.
****

Space-Time Tradeoffs in Photo Sequencing
Tali Dekel (Basha), Yael Moses, Shai Avidan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 977-984
Photo-sequencing is the problem of recovering the temporal order of a set of still images of a dynamic event, taken asynchronously by a set of uncalibrated cameras. Solving this problem is a first, crucial step for analyzing (or visualizing

) the dynamic content of the scene captured by a large number of freely moving s
pectators. We propose a geometric based solution, followed by rank aggregation t
o the photo-sequencing problem. Our algorithm trades spatial certainty for tempo
ral certainty. Whereas the previous solution proposed by [4] relies on two image
s taken from the same static camera to eliminate uncertainty in space, we drop t
he static-camera assumption and replace it with temporal information available f
rom images taken from the same (moving) camera. Our method thus overcomes the li
mitation of the static-camera assumption, and scales much better with the durati
on of the event and the spread of cameras in space. We present successful result
s on challenging real data sets and large scale synthetic data (250 images).
****

Estimating Human Pose with Flowing Puppets
Silvia Zuffi, Javier Romero, Cordelia Schmid, Michael J. Black; Proceedings of t
he IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3312-3319
We address the problem of upper-body human pose estimation in uncontrolled monoc
ular video sequences, without manual initialization. Most current methods focus
on isolated video frames and often fail to correctly localize arms and hands. In
ferring pose over a video sequence is advantageous because poses of people in ad
jacent frames exhibit properties of smooth variation due to the nature of human
and camera motion. To exploit this, previous methods have used prior knowledge a
bout distinctive actions or generic temporal priors combined with static image l
ikelihoods to track people in motion. Here we take a different approach based on
 a simple observation: Information about how a person moves from frame to frame
is present in the optical flow field. We develop an approach for tracking articu
lated motions that "links" articulated shape models of people in adjacent frames
 through the dense optical flow. Key to this approach is a 2D shape model of the
 body that we use to compute how the body moves over time. The resulting "flowin
g puppets" provide a way of integrating image evidence across frames to improve
pose inference. We apply our method on a challenging dataset of TV video sequenc
es and show state-of-the-art performance.
****

Action and Event Recognition with Fisher Vectors on a Compact Feature Set
Dan Oneata, Jakob Verbeek, Cordelia Schmid; Proceedings of the IEEE Internationa
l Conference on Computer Vision (ICCV), 2013, pp. 1817-1824
Action recognition in uncontrolled video is an important and challenging compute
r vision problem. Recent progress in this area is due to new local features and
models that capture spatio-temporal structure between local features, or human-o
bject interactions. Instead of working towards more complex models, we focus on
the low-level features and their encoding. We evaluate the use of Fisher vectors
 as an alternative to bag-of-word histograms to aggregate a small set of state-o
f-the-art low-level descriptors, in combination with linear classifiers. We pres
ent a large and varied set of evaluations, considering (i) classification of sho
rt actions in five datasets, (ii) localization of such actions in feature-length
 movies, and (iii) large-scale recognition of complex events. We find that for b
asic action recognition and localization MBH features alone are enough for state
of-the-art performance. For complex events we find that SIFT and MFCC features p
rovide complementary cues. On all three problems we obtain state-of-the-art resu
lts, while using fewer features and less complex models.
****

Hierarchical Joint Max-Margin Learning of Mid and Top Level Representations for
Visual Recognition
Hans Lobel, Rene Vidal, Alvaro Soto; Proceedings of the IEEE International Confe
rence on Computer Vision (ICCV), 2013, pp. 1697-1704
Currently, Bag-of-Visual-Words (BoVW) and part-based methods are the most popula
r approaches for visual recognition. In both cases, a mid-level representation i
s built on top of low-level image descriptors and top-level classifiers use this
 mid-level representation to achieve visual recognition. While in current part-b
ased approaches, midand top-level representations are usually jointly trained, t
his is not the usual case for BoVW schemes. A main reason for this is the comple
x data association problem related to the usual large dictionary size needed by

BoVW approaches. As a further observation, typical solutions based on BoVW and p
art-based representations are usually limited to extensions of binary classifica
tion schemes, a strategy that ignores relevant correlations among classes. In th
is work we propose a novel hierarchical approach to visual recognition based on
a BoVW scheme that jointly learns suitable midand top-level representations. Fur
thermore, using a maxmargin learning framework, the proposed approach directly h
andles the multiclass case at both levels of abstraction. We test our proposed m
ethod using several popular benchmark datasets. As our main result, we demonstra
te that, by coupling learning of midand top-level representations, the proposed
approach fosters sharing of discriminative visual words among target classes, be
ing able to achieve state-ofthe-art recognition performance using far less visua
l words than previous approaches.
****

Saliency and Human Fixations: State-of-the-Art and Study of Comparison Metrics
Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, Thierry Dutoit
; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 20
13, pp. 1153-1160
Visual saliency has been an increasingly active research area in the last ten ye
ars with dozens of saliency models recently published. Nowadays, one of the big
challenges in the field is to find a way to fairly evaluate all of these models.
 In this paper, on human eye fixations ,we compare the ranking of 12 state-of-th
e art saliency models using 12 similarity metrics. The comparison is done on Jia
n Li's database containing several hundreds of natural images. Based on Kendall
concordance coefficient, it is shown that some of the metrics are strongly corre
lated leading to a redundancy in the performance metrics reported in the availab
le benchmarks. On the other hand, other metrics provide a more diverse picture o
f models' overall performance. As a recommendation, three similarity metrics sho
uld be used to obtain a complete point of view of saliency model performance.
****

Dynamic Label Propagation for Semi-supervised Multi-class Multi-label Classifica
tion
Bo Wang, Zhuowen Tu, John K. Tsotsos; Proceedings of the IEEE International Conf
erence on Computer Vision (ICCV), 2013, pp. 425-432
In graph-based semi-supervised learning approaches, the classification rate is h
ighly dependent on the size of the availabel labeled data, as well as the accura
cy of the similarity measures. Here, we propose a semi-supervised multi-class/mu
lti-label classification scheme, dynamic label propagation (DLP), which performs
 transductive learning through propagation in a dynamic process. Existing semi-s
upervised classification methods often have difficulty in dealing with multi-cla
ss/multi-label problems due to the lack in consideration of label correlation; o
ur algorithm instead emphasizes dynamic metric fusion with label information. Si
gnificant improvement over the state-of-the-art methods is observed on benchmark
 datasets for both multiclass and multi-label tasks.
****

Learning Discriminative Part Detectors for Image Classification and Cosegmentati
on
Jian Sun, Jean Ponce; Proceedings of the IEEE International Conference on Comput
er Vision (ICCV), 2013, pp. 3400-3407
In this paper, we address the problem of learning discriminative part detectors
from image sets with category labels. We propose a novel latent SVM model regula
rized by group sparsity to learn these part detectors. Starting from a large set
 of initial parts, the group sparsity regularizer forces the model to jointly se
lect and optimize a set of discriminative part detectors in a max-margin framewo
rk. We propose a stochastic version of a proximal algorithm to solve the corresp
onding optimization problem. We apply the proposed method to image classificatio
n and cosegmentation, and quantitative experiments with standard benchmarks show
 that it matches or improves upon the state of the art.
****

Co-segmentation by Composition
Alon Faktor, Michal Irani; Proceedings of the IEEE International Conference on C

omputer Vision (ICCV), 2013, pp. 1297-1304
Given a set of images which share an object from the same semantic category, we would like to co-segment the shared object. We define 'good' co-segments to be ones which can be easily composed (like a puzzle) from large pieces of other co-segments, yet are difficult to compose from remaining image parts. These pieces must not only match well but also be statistically significant (hard to compose at random). This gives rise to co-segmentation of objects in very challenging scenarios with large variations in appearance, shape and large amounts of clutter. We further show how multiple images can collaborate and "score" each others' co-segments to improve the overall fidelity and accuracy of the co-segmentation. Our co-segmentation can be applied both to large image collections, as well as to very few images (where there is too little data for unsupervised learning). At the extreme, it can be applied even to a single image, to extract its co-occurring objects. Our approach obtains state-of-the-art results on benchmark datasets. We further show very encouraging co-segmentation results on the challenging PASCAL-VOC dataset.
****

A New Image Quality Metric for Image Auto-denoising
Xiangfei Kong, Kuan Li, Qingxiong Yang, Liu Wenyin, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2888-2895
This paper proposes a new non-reference image quality metric that can be adopted by the state-of-the-art image/video denoising algorithms for auto-denoising. The proposed metric is extremely simple and can be implemented in four lines of Matlab code 1 . The basic assumption employed by the proposed metric is that the noise should be independent of the original image. A direct measurement of this dependence is, however, impractical due to the relatively low accuracy of existing denoising method. The proposed metric thus aims at maximizing the structure similarity between the input noisy image and the estimated image noise around homogeneous regions and the structure similarity between the input noisy image and the denoised image around highly-structured regions, and is computed as the linear correlation coefficient of the two corresponding structure similarity maps. Numerous experimental results demonstrate that the proposed metric not only outperforms the current state-of-the-art non-reference quality metric quantitatively and qualitatively, but also better maintains temporal coherence when used for video denoising.
****

Random Faces Guided Sparse Many-to-One Encoder for Pose-Invariant Face Recognition
Yizhe Zhang, Ming Shao, Edward K. Wong, Yun Fu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2416-2423
One of the most challenging task in face recognition is to identify people with varied poses. Namely, the test faces have significantly different poses compared with the registered faces. In this paper, we propose a high-level feature learning scheme to extract pose-invariant identity feature for face recognition. First, we build a single-hiddenlayer neural network with sparse constraint, to extract poseinvariant feature in a supervised fashion. Second, we further enhance the discriminative capability of the proposed feature by using multiple random faces as the target values for multiple encoders. By enforcing the target values to be unique for input faces over different poses, the learned highlevel feature that is represented by the neurons in the hidden layer is pose free and only relevant to the identity information. Finally, we conduct face identification on CMU MultiPIE, and verification on Labeled Faces in the Wild (LFW) databases, where identification rank-1 accuracy and face verification accuracy with ROC curve are reported. These experiments demonstrate that our model is superior to other state-of-the-art approaches on handling pose variations.
****

Text Localization in Natural Images Using Stroke Feature Transform and Text Covariance Descriptors
Weilin Huang, Zhe Lin, Jianchao Yang, Jue Wang; Proceedings of the IEEE Internat

ional Conference on Computer Vision (ICCV), 2013, pp. 1241-1248
In this paper, we present a new approach for text localization in natural images , by discriminating text and non-text regions at three levels: pixel, component and textline levels. Firstly, a powerful low-level filter called the Stroke Feat ure Transform (SFT) is proposed, which extends the widely-used Stroke Width Tran sform (SWT) by incorporating color cues of text pixels, leading to significantly enhanced performance on inter-component separation and intra-component connecti on. Secondly, based on the output of SFT, we apply two classifiers, a text compo nent classifier and a text-line classifier, sequentially to extract text regions , eliminating the heuristic procedures that are commonly used in previous approa ches. The two classifiers are built upon two novel Text Covariance Descriptors (TCDs) that encode both the heuristic properties and the statistical characterist ics of text stokes. Finally, text regions are located by simply thresholding the text-line confident map. Our method was evaluated on two benchmark datasets: IC DAR 2005 and ICDAR 2011, and the corresponding Fmeasure values are 0.72 and 0.73 , respectively, surpassing previous methods in accuracy by a large margin.
****

Modeling the Calibration Pipeline of the Lytro Camera for High Quality Light-Fie ld Image Reconstruction
Donghyeon Cho, Minhaeng Lee, Sunyeong Kim, Yu-Wing Tai; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3280-3287
Light-field imaging systems have got much attention recently as the next generat ion camera model. A light-field imaging system consists of three parts: data acq uisition, manipulation, and application. Given an acquisition system, it is impo rtant to understand how a light-field camera converts from its raw image to its resulting refocused image. In this paper, using the Lytro camera as an example, we describe step-by-step procedures to calibrate a raw light-field image. In par ticular, we are interested in knowing the spatial and angular coordinates of the micro lens array and the resampling process for image reconstruction. Since Lyt ro uses a hexagonal arrangement of a micro lens image, additional treatments in calibration are required. After calibration, we analyze and compare the performa nces of several resampling methods for image reconstruction with and without cal ibration. Finally, a learning based interpolation method is proposed which demon strates a higher quality image reconstruction than previous interpolation method s including a method used in Lytro software.
****

Learning Graph Matching: Oriented to Category Modeling from Cluttered Scenes
Quanshi Zhang, Xuan Song, Xiaowei Shao, Huijing Zhao, Ryosuke Shibasaki; Proceed ings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1 329-1336
Although graph matching is a fundamental problem in pattern recognition, and has drawn broad interest from many fields, the problem of learning graph matching h as not received much attention. In this paper, we redefine the learning of graph matching as a model learning problem. In addition to conventional training of m atching parameters, our approach modifies the graph structure and attributes to generate a graphical model. In this way, the model learning is oriented toward b oth matching and recognition performance, and can proceed in an unsupervised gnf ashion. Experiments demonstrate that our approach outperforms conventional metho ds for learning graph matching.
****

Action Recognition with Improved Trajectories
Heng Wang, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3551-3558
Recently dense trajectories were shown to be an efficient video representation f or action recognition and achieved state-of-the-art results on a variety of data sets. This paper improves their performance by taking into account camera motion to correct them. To estimate camera motion, we match feature points between fra mes using SURF descriptors and dense optical flow, which are shown to be complem entary. These matches are, then, used to robustly estimate a homography with RAN SAC. Human motion is in general different from camera motion and generates incon

sistent matches. To improve the estimation, a human detector is employed to remo ve these matches. Given the estimated camera motion, we remove trajectories cons istent with it. We also use this estimation to cancel out camera motion from the optical flow. This significantly improves motion-based descriptors, such as HOF and MBH. Experimental results on four challenging action datasets (i.e., Hollyw ood2, HMDB51, Olympic Sports and UCF50) significantly outperform the current sta te of the art.

****

A Generic Deformation Model for Dense Non-rigid Surface Registration: A Higher-O rder MRF-Based Approach

Yun Zeng, Chaohui Wang, Xianfeng Gu, Dimitris Samaras, Nikos Paragios; Proceedin gs of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 336 0-3367

We propose a novel approach for dense non-rigid 3D surface registration, which b rings together Riemannian geometry and graphical models. To this end, we first i ntroduce a generic deformation model, called Canonical Distortion Coefficients (CDCs), by characterizing the deformation of every point on a surface using the d istortions along its two principle directions. This model subsumes the deformati on groups commonly used in surface registration such as isometry and conformalit y, and is able to handle more complex deformations. We also derive its discrete counterpart which can be computed very efficiently in a closed form. Based on th ese, we introduce a higher-order Markov Random Field (MRF) model which seamlessl y integrates our deformation model and a geometry/texture similarity metric. The n we jointly establish the optimal correspondences for all the points via maximu m a posteriori (MAP) inference. Moreover, we develop a parallel optimization alg orithm to efficiently perform the inference for the proposed higher-order MRF mo del. The resulting registration algorithm outperforms state-of-the-art methods i n both dense non-rigid 3D surface registration and tracking.

****

Joint Inverted Indexing

Yan Xia, Kaiming He, Fang Wen, Jian Sun; Proceedings of the IEEE International C onference on Computer Vision (ICCV), 2013, pp. 3416-3423

Inverted indexing is a popular non-exhaustive solution to large scale search. An inverted file is built by a quantizer such as k-means or a tree structure. It h as been found that multiple inverted files, obtained by multiple independent ran dom quantizers, are able to achieve practically good recall and speed. Instead o f computing the multiple quantizers independently, we present a method that crea tes them jointly. Our method jointly optimizes all codewords in all quantizers. Then it assigns these codewords to the quantizers. In experiments this method sh ows significant improvement over various existing methods that use multiple inde pendent quantizers. On the one-billion set of SIFT vectors, our method is faster and more accurate than a recent state-of-the-art inverted indexing method.

****

From Point to Set: Extend the Learning of Distance Metrics

Pengfei Zhu, Lei Zhang, Wangmeng Zuo, David Zhang; Proceedings of the IEEE Inter national Conference on Computer Vision (ICCV), 2013, pp. 2664-2671

Most of the current metric learning methods are proposed for point-to-point dist ance (PPD) based classification. In many computer vision tasks, however, we need to measure the point-to-set distance (PSD) and even set-to-set distance (SSD) f or classification. In this paper, we extend the PPD based Mahalanobis distance m etric learning to PSD and SSD based ones, namely point-to-set distance metric le arning (PSDML) and set-to-set distance metric learning (SSDML), and solve them u nder a unified optimization framework. First, we generate positive and negative sample pairs by computing the PSD and SSD between training samples. Then, we cha racterize each sample pair by its covariance matrix, and propose a covariance ke rnel based discriminative function. Finally, we tackle the PSDML and SSDML probl ems by using standard support vector machine solvers, making the metric learning very efficient for multiclass visual classification tasks. Experiments on gende r classification, digit recognition, object categorization and face recognition show that the proposed metric learning methods can effectively enhance the perfo

rmance of PSD and SSD based classification.
****
Learning View-Invariant Sparse Representations for Cross-View Action Recognition
Jingjing Zheng, Zhuolin Jiang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3176-3183
We present an approach to jointly learn a set of viewspecific dictionaries and a common dictionary for crossview action recognition. The set of view-specific dictionaries is learned for specific views while the common dictionary is shared across different views. Our approach represents videos in each view using both the corresponding view-specific dictionary and the common dictionary. More importantly, it encourages the set of videos taken from different views of the same action to have similar sparse representations. In this way, we can align view-specific features in the sparse feature spaces spanned by the viewspecific dictionary set and transfer the view-shared features in the sparse feature space spanned by the common dictionary. Meanwhile, the incoherence between the common dictionary and the view-specific dictionary set enables us to exploit the discrimination information encoded in viewspecific features and view-shared features separately. In addition, the learned common dictionary not only has the capability to represent actions from unseen views, but also makes our approach effective in a semi-supervised setting where no correspondence videos exist and only a few labels exist in the target view. Extensive experiments using the multi-view IXMAS dataset demonstrate that our approach outperforms many recent approaches for cross-view action recognition.
****
Line Assisted Light Field Triangulation and Stereo Matching
Zhan Yu, Xinqing Guo, Haibing Lin, Andrew Lumsdaine, Jingyi Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2792-2799
Light fields are image-based representations that use densely sampled rays as a scene description. In this paper, we explore geometric structures of 3D lines in ray space for improving light field triangulation and stereo matching. The triangulation problem aims to fill in the ray space with continuous and non-overlapping simplices anchored at sampled points (rays). Such a triangulation provides a piecewise-linear interpolant useful for light field superresolution. We show that the light field space is largely bilinear due to 3D line segments in the scene, and direct triangulation of these bilinear subspaces leads to large errors. We instead present a simple but effective algorithm to first map bilinear subspaces to line constraints and then apply Constrained Delaunay Triangulation (CDT). Based on our analysis, we further develop a novel line-assisted graphcut (LAGC) algorithm that effectively encodes 3D line constraints into light field stereo matching. Experiments on synthetic and real data show that both our triangulation and LAGC algorithms outperform state-of-the-art solutions in accuracy and visual quality.
****
Viewing Real-World Faces in 3D
Tal Hassner; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3607-3614
We present a data-driven method for estimating the 3D shapes of faces viewed in single, unconstrained photos (aka "in-the-wild"). Our method was designed with an emphasis on robustness and efficiency with the explicit goal of deployment in real-world applications which reconstruct and display faces in 3D. Our key observation is that for many practical applications, warping the shape of a reference face to match the appearance of a query, is enough to produce realistic impressions of the query's 3D shape. Doing so, however, requires matching visual features between the (possibly very different) query and reference images, while ensuring that a plausible face shape is produced. To this end, we describe an optimization process which seeks to maximize the similarity of appearances and depths, jointly, to those of a reference model. We describe our system for monocular face shape reconstruction and present both qualitative and quantitative experiments, comparing our method against alternative systems, and demonstrating its capabilities. Finally, as a testament to its suitability for real-world applications,

we offer an open, online implementation of our system, providing unique means of instant 3D viewing of faces appearing in web photos.
****

Towards Motion Aware Light Field Video for Dynamic Scenes
Salil Tambe, Ashok Veeraraghavan, Amit Agrawal; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1009-1016
Current Light Field (LF) cameras offer fixed resolution in space, time and angle which is decided a-priori and is independent of the scene. These cameras either trade-off spatial resolution to capture single-shot LF [20, 27, 12] or tradeoff temporal resolution by assuming a static scene to capture high spatial resolution LF [18, 3]. Thus, capturing high spatial resolution LF video for dynamic scenes remains an open and challenging problem. We present the concept, design and implementation of a LF video camera that allows capturing high resolution LF video. The spatial, angular and temporal resolution are not fixed a-priori and we exploit the scene-specific redundancy in space, time and angle. Our reconstruction is motion-aware and offers a continuum of resolution tradeoff with increasing motion in the scene. The key idea is (a) to design efficient multiplexing matrices that allow resolution tradeoffs, (b) use dictionary learning and sparse representations for robust reconstruction, and (c) perform local motion-aware adaptive reconstruction. We perform extensive analysis and characterize the performance of our motion-aware reconstruction algorithm. We show realistic simulations using a graphics simulator as well as real results using a LCoS based programmable camera. We demonstrate novel results such as high resolution digital refocusing for dynamic moving objects.
****

Abnormal Event Detection at 150 FPS in MATLAB
Cewu Lu, Jianping Shi, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2720-2727
Speedy abnormal event detection meets the growing demand to process an enormous number of surveillance videos. Based on inherent redundancy of video structures, we propose an efficient sparse combination learning framework. It achieves decent performance in the detection phase without compromising result quality. The short running time is guaranteed because the new method effectively turns the original complicated problem to one in which only a few costless small-scale least square optimization steps are involved. Our method reaches high detection rates on benchmark datasets at a speed of 140~150 frames per second on average when computing on an ordinary desktop PC using MATLAB.
****

Elastic Fragments for Dense Scene Reconstruction
Qian-Yi Zhou, Stephen Miller, Vladlen Koltun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 473-480
We present an approach to reconstruction of detailed scene geometry from range video. Range data produced by commodity handheld cameras suffers from high-frequency errors and low-frequency distortion. Our approach deals with both sources of error by reconstructing locally smooth scene fragments and letting these fragments deform in order to align to each other. We develop a volumetric registration formulation that leverages the smoothness of the deformation to make optimization practical for large scenes. Experimental results demonstrate that our approach substantially increases the fidelity of complex scene geometry reconstructed with commodity handheld cameras.
****

Shape Anchors for Data-Driven Multi-view Reconstruction
Andrew Owens, Jianxiong Xiao, Antonio Torralba, William Freeman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 33-40
We present a data-driven method for building dense 3D reconstructions using a combination of recognition and multi-view cues. Our approach is based on the idea that there are image patches that are so distinctive that we can accurately estimate their latent 3D shapes solely using recognition. We call these patches shape anchors, and we recoggnioon.WWWeccallteeseepatchhesshhapeaanchorrs,anndwwe that transfers dense, complex geometry between scenes. We "anchor" our 3D interpre

tation from these patches, using them to predict geometry for parts of the scene that are relatively ambiguous. The resulting algorithm produces dense reconstructions from stereo point clouds that are sparse and noisy, and we demonstrate it on a challenging dataset of real-world, indoor scenes.
****

Piecewise Rigid Scene Flow
Christoph Vogel, Konrad Schindler, Stefan Roth; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1377-1384
Estimating dense 3D scene flow from stereo sequences remains a challenging task, despite much progress in both classical disparity and 2D optical flow estimation. To overcome the limitations of existing techniques, we introduce a novel model that represents the dynamic 3D scene by a collection of planar, rigidly moving, local segments. Scene flow estimation then amounts to jointly estimating the pixelto-segment assignment, and the 3D position, normal vector, and rigid motion parameters of a plane for each segment. The proposed energy combines an occlusion-sensitive data term with appropriate shape, motion, and segmentation regularizers. Optimization proceeds in two stages: Starting from an initial superpixelization, we estimate the shape and motion parameters of all segments by assigning a proposal from a set of moving planes. Then the pixel-to-segment assignment is updated, while holding the shape and motion parameters of the moving planes fixed. We demonstrate the benefits of our model on different real-world image sets, including the challenging KITTI benchmark. We achieve leading performance levels, exceeding competing 3D scene flow methods, and even yielding better 2D motion estimates than all tested dedicated optical flow techniques.
****

Spoken Attributes: Mixing Binary and Relative Attributes to Say the Right Thing
Amir Sadovnik, Andrew Gallagher, Devi Parikh, Tsuhan Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2160-2167
In recent years, there has been a great deal of progress in describing objects with attributes. Attributes have proven useful for object recognition, image search, face verification, image description, and zero-shot learning. Typically, attributes are either binary or relative: they describe either the presence or absence of a descriptive characteristic, or the relative magnitude of the characteristic when comparing two exemplars. However, prior work fails to model the actual way in which humans use these attributes in descriptive statements of images. Specifically, it does not address the important interactions between the binary and relative aspects of an attribute. In this work we propose a spoken attribute classifier which models a more natural way of using an attribute in a description. For each attribute we train a classifier which captures the specific way this attribute should be used. We show that as a result of using this model, we produce descriptions about images of people that are more natural and specific than past systems.
****

Coupled Dictionary and Feature Space Learning with Applications to Cross-Domain Image Synthesis and Recognition
De-An Huang, Yu-Chiang Frank Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2496-2503
Cross-domain image synthesis and recognition are typically considered as two distinct tasks in the areas of computer vision and pattern recognition. Therefore, it is not clear whether approaches addressing one task can be easily generalized or extended for solving the other. In this paper, we propose a unified model for coupled dictionary and feature space learning. The proposed learning model not only observes a common feature space for associating cross-domain image data for recognition purposes, the derived feature space is able to jointly update the dictionaries in each image domain for improved representation. This is why our method can be applied to both cross-domain image synthesis and recognition problems. Experiments on a variety of synthesis and recognition tasks such as single image super-resolution, cross-view action recognition, and sketchto-photo face recognition would verify the effectiveness of our proposed learning model.
****

Hierarchical Part Matching for Fine-Grained Visual Categorization
Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan, Bo Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1641-1648
As a special topic in computer vision, fine-grained visual categorization (FGVC) has been attracting growing attention these years. Different with traditional image classification tasks in which objects have large inter-class variation, the visual concepts in the fine-grained datasets, such as hundreds of bird species, often have very similar semantics. Due to the large inter-class similarity, it is very difficult to classify the objects without locating really discriminative features, therefore it becomes more important for the algorithm to make full use of the part information in order to train a robust model. In this paper, we propose a powerful flowchart named Hierarchical Part Matching (HPM) to cope with finegrained classification tasks. We extend the Bag-of-Features (BoF) model by introducing several novel modules to integrate into image representation, including foreground inference and segmentation, Hierarchical Structure Learning (HSL), and Geometric Phrase Pooling (GPP). We verify in experiments that our algorithm achieves the state-ofthe-art classification accuracy in the Caltech-UCSD-Birds200-2011 dataset by making full use of the ground-truth part annotations.
****

Locally Affine Sparse-to-Dense Matching for Motion and Occlusion Estimation
Marius Leordeanu, Andrei Zanfir, Cristian Sminchisescu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1721-1728
Estimating a dense correspondence field between successive video frames, under large displacement, is important in many visual learning and recognition tasks. We propose a novel sparse-to-dense matching method for motion field estimation and occlusion detection. As an alternative to the current coarse-to-fine approaches from the optical flow literature, we start from the higher level of sparse matching with rich appearance and geometric constraints collected over extended neighborhoods, using an occlusion aware, locally affine model. Then, we move towards the simpler, but denser classic flow field model, with an interpolation procedure that offers a natural transition between the sparse and the dense correspondence fields. We experimentally demonstrate that our appearance features and our complex geometric constraints permit the correct motion estimation even in difficult cases of large displacements and significant appearance changes. We also propose a novel classification method for occlusion detection that works in conjunction with the sparse-to-dense matching model. We validate our approach on the newly released Sintel dataset and obtain state-of-the-art results.
****

SUN3D: A Database of Big Spaces Reconstructed Using SfM and Object Labels
Jianxiong Xiao, Andrew Owens, Antonio Torralba; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1625-1632
Existing scene understanding datasets contain only a limited set of views of a place, and they lack representations of complete 3D spaces. In this paper, we introduce SUN3D, a large-scale RGB-D video database with camera pose and object labels, capturing the full 3D extent of many places. The tasks that go into constructing such a dataset are difficult in isolation hand-labeling videos is painstaking, and structure from motion (SfM) is unreliable for large spaces. But if we combine them together, we make the dataset construction task much easier. First, we introduce an intuitive labeling tool that uses a partial reconstruction to propagate labels from one frame to another. Then we use the object labels to fix errors in the reconstruction. For this, we introduce a generalization of bundle adjustment that incorporates object-to-object correspondences. This algorithm works by constraining points for the same object from different frames to lie inside a fixed-size bounding box, parameterized by its rotation and translation. The SUN3D database, the source code for the generalized bundle adjustment, and the web-based 3D annotation tool are all available at http://sun3d.cs.princeton.edu.
****

A Deep Sum-Product Architecture for Robust Facial Attributes Analysis
Ping Luo, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2864-2871

Recent works have shown that facial attributes are useful in a number of applica tions such as face recognition and retrieval. However, estimating attributes in images with large variations remains a big challenge. This challenge is addresse d in this paper. Unlike existing methods that assume the independence of attribu tes during their estimation, our approach captures the interdependencies of loca l regions for each attribute, as well as the high-order correlations between dif ferent attributes, which makes it more robust to occlusions and misdetection of face regions. First, we have modeled region interdependencies with a discriminat ive decision tree, where each node consists of a detector and a classifier train ed on a local region. The detector allows us to locate the region, while the cla ssifier determines the presence or absence of an attribute. Second, correlations of attributes and attribute predictors are modeled by organizing all of the dec ision trees into a large sum-product network (SPN), which is learned by the EM a lgorithm and yields the most probable explanation (MPE) of the facial attributes in terms of the region's localization and classification. Experimental results on a large data set with 22, 400 images show the effectiveness of the proposed a pproach.
****

Point-Based 3D Reconstruction of Thin Objects
Benjamin Ummenhofer, Thomas Brox; Proceedings of the IEEE International Conferen ce on Computer Vision (ICCV), 2013, pp. 969-976
3D reconstruction deals with the problem of finding the shape of an object from a set of images. Thin objects that have virtually no volume pose a special chall enge for reconstruction with respect to shape representation and fusion of depth information. In this paper we present a dense pointbased reconstruction method that can deal with this special class of objects. We seek to jointly optimize a set of depth maps by treating each pixel as a point in space. Points are pulled towards a common surface by pairwise forces in an iterative scheme. The method a lso handles the problem of opposed surfaces by means of penalty forces. Efficien t optimization is achieved by grouping points to superpixels and a spatial hashi ng approach for fast neighborhood queries. We show that the approach is on a par with state-of-the-art methods for standard multi view stereo settings and gives superior results for thin objects.
****

Structured Learning of Sum-of-Submodular Higher Order Energy Functions
Alexander Fix, Thorsten Joachims, Sung Min Park, Ramin Zabih; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3104-3111
Submodular functions can be exactly minimized in polynomial time, and the specia l case that graph cuts solve with max flow [19] has had significant impact in co mputer vision [5, 21, 28]. In this paper we address the important class of sum-o f-submodular (SoS) functions [2, 18], which can be efficiently minimized via a v ariant of max flow called submodular flow [6]. SoS functions can naturally expre ss higher order priors involving, e.g., local image patches; however, it is diff icult to fully exploit their expressive power because they have so many paramete rs. Rather than trying to formulate existing higher order priors as an SoS funct ion, we take a discriminative learning approach, effectively searching the space of SoS functions for a higher order prior that performs well on our training se t. We adopt a structural SVM approach [15, 34] and formulate the training proble m in terms of quadratic programming; as a result we can efficiently search the s pace of SoS priors via an extended cutting-plane algorithm. We also show how the state-of-the-art max flow method for vision problems [11] can be modified to ef ficiently solve the submodular flow problem. Experimental comparisons are made a gainst the OpenCV implementation of the GrabCut interactive segmentation techniq ue [28], which uses hand-tuned parameters instead of machine learning. On a stan dard dataset [12] our method learns higher order priors with hundreds of paramet er values, and produces significantly better segmentations. While our focus is o n binary labeling problems, we show that our techniques can be naturally general ized to handle more than two labels.
****

Affine-Constrained Group Sparse Coding and Its Application to Image-Based Classi

fications
Yu-Tseh Chi, Mohsen Ali, Muhammad Rushdi, Jeffrey Ho; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 681-688
This paper proposes a novel approach for sparse coding that further improves upon the sparse representation-based classification (SRC) framework. The proposed framework, Affine-Constrained Group Sparse Coding (ACGSC), extends the current SRC framework to classification problems with multiple input samples. Geometrically, the affineconstrained group sparse coding essentially searches for the vector in the convex hull spanned by the input vectors that can best be sparse coded using the given dictionary. The resulting objective function is still convex and can be efficiently optimized using iterative block-coordinate descent scheme that is guaranteed to converge. Furthermore, we provide a form of sparse recovery result that guarantees, at least theoretically, that the classification performance of the constrained group sparse coding should be at least as good as the group sparse coding. We have evaluated the proposed approach using three different recognition experiments that involve illumination variation of faces and textures, and face recognition under occlusions. Preliminary experiments have demonstrated the effectiveness of the proposed approach, and in particular, the results from the recognition/occlusion experiment are surprisingly accurate and robust.
****

Latent Space Sparse Subspace Clustering
Vishal M. Patel, Hien Van Nguyen, Rene Vidal; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 225-232
We propose a novel algorithm called Latent Space Sparse Subspace Clustering for simultaneous dimensionality reduction and clustering of data lying in a union of subspaces. Specifically, we describe a method that learns the projection of data and finds the sparse coefficients in the low-dimensional latent space. Cluster labels are then assigned by applying spectral clustering to a similarity matrix built from these sparse coefficients. An efficient optimization method is proposed and its non-linear extensions based on the kernel methods are presented. One of the main advantages of our method is that it is computationally efficient as the sparse coefficients are found in the low-dimensional latent space. Various experiments show that the proposed method performs better than the competitive state-of-theart subspace clustering methods.
****

To Aggregate or Not to aggregate: Selective Match Kernels for Image Search
Giorgos Tolias, Yannis Avrithis, Herve Jegou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1401-1408
This paper considers a family of metrics to compare images based on their local descriptors. It encompasses the VLAD descriptor and matching techniques such as Hamming Embedding. Making the bridge between these approaches leads us to propose a match kernel that takes the best of existing techniques by combining an aggregation procedure with a selective match kernel. Finally, the representation underpinning this kernel is approximated, providing a large scale image search both precise and scalable, as shown by our experiments on several benchmarks.
****

Person Re-identification by Salience Matching
Rui Zhao, Wanli Ouyang, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2528-2535
Human salience is distinctive and reliable information in matching pedestrians across disjoint camera views. In this paper, we exploit the pairwise salience distribution relationship between pedestrian images, and solve the person re-identification problem by proposing a salience matching strategy. To handle the misalignment problem in pedestrian images, patch matching is adopted and patch salience is estimated. Matching patches with inconsistent salience brings penalty. Images of the same person are recognized by minimizing the salience matching cost. Furthermore, our salience matching is tightly integrated with patch matching in a unified structural RankSVM learning framework. The effectiveness of our approach is validated on the VIPeR dataset and the CUHK Campus dataset. It outperforms the state-of-the-art methods on both datasets.

****
Pose Estimation with Unknown Focal Length Using Points, Directions and Lines
Yubin Kuang, Kalle Astrom; Proceedings of the IEEE International Conference on C omputer Vision (ICCV), 2013, pp. 529-536
In this paper, we study the geometry problems of estimating camera pose with unk nown focal length using combination of geometric primitives. We consider points, lines and also rich features such as quivers, i.e. points with one or more dire ctions. We formulate the problems as polynomial systems where the constraints fo r different primitives are handled in a unified way. We develop efficient polyno mial solvers for each of the derived cases with different combinations of primit ives. The availability of these solvers enables robust pose estimation with unkn own focal length for wider classes of features. Such rich features allow for few er feature correspondences and generate larger inlier sets with higher probabili ty. We demonstrate in synthetic experiments that our solvers are fast and numeri cally stable. For real images, we show that our solvers can be used in RANSAC lo ops to provide good initial solutions.
****
Action Recognition and Localization by Hierarchical Space-Time Segments
Shugao Ma, Jianming Zhang, Nazli Ikizler-Cinbis, Stan Sclaroff; Proceedings of t he IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2744-2751
We propose Hierarchical Space-Time Segments as a new representation for action r ecognition and localization. This representation has a two-level hierarchy. The first level comprises the root space-time segments that may contain a human body . The second level comprises multi-grained space-time segments that contain part s of the root. We present an unsupervised method to generate this representation from video, which extracts both static and non-static relevant space-time segme nts, and also preserves their hierarchical and temporal relationships. Using sim ple linear SVM on the resultant bag of hierarchical space-time segments represen tation, we attain better than, or comparable to, state-of-the-art action recogni tion performance on two challenging benchmark datasets and at the same time prod uce good action localization results.
****
A General Dense Image Matching Framework Combining Direct and Feature-Based Cost s
Jim Braux-Zin, Romain Dupont, Adrien Bartoli; Proceedings of the IEEE Internatio nal Conference on Computer Vision (ICCV), 2013, pp. 185-192
Dense motion field estimation (typically optical flow, stereo disparity and surf ace registration) is a key computer vision problem. Many solutions have been pro posed to compute small or large displacements, narrow or wide baseline stereo di sparity, but a unified methodology is still lacking. We here introduce a general framework that robustly combines direct and feature-based matching. The feature -based cost is built around a novel robust distance function that handles keypoi nts and "weak" features such as segments. It allows us to use putative feature m atches which may contain mismatches to guide dense motion estimation out of loca l minima. Our framework uses a robust direct data term (AD-Census). It is implem ented with a powerful second order Total Generalized Variation regularization wi th external and self-occlusion reasoning. Our framework achieves state of the ar t performance in several cases (standard optical flow benchmarks, wide-baseline stereo and non-rigid surface registration). Our framework has a modular design t hat customizes to specific application needs.
****
Fast Neighborhood Graph Search Using Cartesian Concatenation
Jing Wang, Jingdong Wang, Gang Zeng, Rui Gan, Shipeng Li, Baining Guo; Proceedin gs of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 212 8-2135
In this paper, we propose a new data structure for approximate nearest neighbor search. This structure augments the neighborhood graph with a bridge graph. We p ropose to exploit Cartesian concatenation to produce a large set of vectors, cal led bridge vectors, from several small sets of subvectors. Each bridge vector is connected with a few reference vectors near to it, forming a bridge graph. Our

approach finds nearest neighbors by simultaneously traversing the neighborhood g
raph and the bridge graph in the best-first strategy. The success of our approac
h stems from two factors: the exact nearest neighbor search over a large number
of bridge vectors can be done quickly, and the reference vectors connected to a
bridge (reference) vector near the query are also likely to be near the query. E
xperimental results on searching over large scale datasets (SIFT, GIST and HOG)
show that our approach outperforms stateof-the-art ANN search algorithms in term
s of efficiency and accuracy. The combination of our approach with the IVFADC sy
stem [18] also shows superior performance over the BIGANN dataset of 1 billion S
IFT features compared with the best previously published result.
****

Uncertainty-Driven Efficiently-Sampled Sparse Graphical Models for Concurrent Tu
mor Segmentation and Atlas Registration
Sarah Parisot, William Wells III, Stephane Chemouny, Hugues Duffau, Nikos Paragi
os; Proceedings of the IEEE International Conference on Computer Vision (ICCV),
2013, pp. 641-648
Graph-based methods have become popular in recent years and have successfully ad
dressed tasks like segmentation and deformable registration. Their main strength
 is optimality of the obtained solution while their main limitation is the lack
of precision due to the grid-like representations and the discrete nature of the
 quantized search space. In this paper we introduce a novel approach for combine
d segmentation/registration of brain tumors that adapts graph and sampling resol
ution according to the image content. To this end we estimate the segmentation a
nd registration marginals towards adaptive graph resolution and intelligent defi
nition of the search space. This information is considered in a hierarchical fra
mework where uncertainties are propagated in a natural manner. State of the art
results in the joint segmentation/registration of brain images with low-grade gl
iomas demonstrate the potential of our approach.
****

Learning Near-Optimal Cost-Sensitive Decision Policy for Object Detection
Tianfu Wu, Song-Chun Zhu; Proceedings of the IEEE International Conference on Co
mputer Vision (ICCV), 2013, pp. 753-760
Many object detectors, such as AdaBoost, SVM and deformable part-based models (D
PM), compute additive scoring functions at a large number of windows scanned ove
r image pyramid, thus computational efficiency is an important consideration bes
ide accuracy performance. In this paper, we present a framework of learning cost
-sensitive decision policy which is a sequence of two-sided thresholds to execut
e early rejection or early acceptance based on the accumulative scores at each s
tep. A decision policy is said to be optimal if it minimizes an empirical global
 risk function that sums over the loss of false negatives (FN) and false positiv
es (FP), and the cost of computation. While the risk function is very complex du
e to high-order connections among the two-sided thresholds, we find its upper bo
und can be optimized by dynamic programming (DP) efficiently and thus say the le
arned policy is near-optimal. Given the loss of FN and FP and the cost in three
numbers, our method can produce a policy on-the-fly for Adaboost, SVM and DPM. I
n experiments, we show that our decision policy outperforms state-of-the-art cas
cade methods significantly in terms of speed with similar accuracy performance.
****

Coherent Object Detection with 3D Geometric Context from a Single Image
Jiyan Pan, Takeo Kanade; Proceedings of the IEEE International Conference on Com
puter Vision (ICCV), 2013, pp. 2576-2583
Objects in a real world image cannot have arbitrary appearance, sizes and locati
ons due to geometric constraints in 3D space. Such a 3D geometric context plays
an important role in resolving visual ambiguities and achieving coherent object
detection. In this paper, we develop a RANSAC-CRF framework to detect objects th
at are geometrically coherent in the 3D world. Different from existing methods,
we propose a novel generalized RANSAC algorithm to generate global 3D geometry h
ypotheses from local entities such that outlier suppression and noise reduction
is achieved simultaneously. In addition, we evaluate those hypotheses using a CR
F which considers both the compatibility of individual objects under global 3D g

eometric context and the compatibility between adjacent objects under local 3D g
eometric context. Experiment results show that our approach compares favorably w
ith the state of the art.
****

Unsupervised Visual Domain Adaptation Using Subspace Alignment
Basura Fernando, Amaury Habrard, Marc Sebban, Tinne Tuytelaars; Proceedings of t
he IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2960-2967
In this paper, we introduce a new domain adaptation (DA) algorithm where the sou
rce and target domains are represented by subspaces described by eigenvectors. I
n this context, our method seeks a domain adaptation solution by learning a mapp
ing function which aligns the source subspace with the target one. We show that
the solution of the corresponding optimization problem can be obtained in a simp
le closed form, leading to an extremely fast algorithm. We use a theoretical res
ult to tune the unique hyperparameter corresponding to the size of the subspaces
. We run our method on various datasets and show that, despite its intrinsic sim
plicity, it outperforms state of the art DA methods.
****

Semi-supervised Learning for Large Scale Image Cosegmentation
Zhengxiang Wang, Rujie Liu; Proceedings of the IEEE International Conference on
Computer Vision (ICCV), 2013, pp. 393-400
This paper introduces to use semi-supervised learning for large scale image cose
gmentation. Different from traditional unsupervised cosegmentation that does not
 use any segmentation groundtruth, semi-supervised cosegmentation exploits the s
imilarity from both the very limited training image foregrounds, as well as the
common object shared between the large number of unsegmented images. This would
be a much practical way to effectively cosegment a large number of related image
s simultaneously, where previous unsupervised cosegmentation work poorly due to
the large variances in appearance between different images and the lack of segme
ntation groundtruth for guidance in cosegmentation. For semi-supervised cosegmen
tation in large scale, we propose an effective method by minimizing an energy fu
nction, which consists of the inter-image distance, the intraimage distance and
the balance term. We also propose an iterative updating algorithm to efficiently
 solve this energy function, which decomposes the original energy minimization p
roblem into sub-problems, and updates each image alternatively to reduce the num
ber of variables in each subproblem for computation efficiency. Experiment resul
ts on iCoseg and Pascal VOC datasets show that the proposed cosegmentation metho
d can effectively cosegment hundreds of images in less than one minute. And our
semi-supervised cosegmentation is able to outperform both unsupervised cosegment
ation as well as fully supervised single image segmentation, especially when the
 training data is limited.
****

Mining Multiple Queries for Image Retrieval: On-the-Fly Learning of an Object-Sp
ecific Mid-level Representation
Basura Fernando, Tinne Tuytelaars; Proceedings of the IEEE International Confere
nce on Computer Vision (ICCV), 2013, pp. 2544-2551
In this paper we present a new method for object retrieval starting from multipl
e query images. The use of multiple queries allows for a more expressive formula
tion of the query object including, e.g., different viewpoints and/or viewing co
nditions. This, in turn, leads to more diverse and more accurate retrieval resul
ts. When no query images are available to the user, they can easily be retrieved
 from the internet using a standard image search engine. In particular, we propo
se a new method based on pattern mining. Using the minimal description length pr
inciple, we derive the most suitable set of patterns to describe the query objec
t, with patterns corresponding to local feature configurations. This results in
a powerful object-specific mid-level image representation. The archive can then
be searched efficiently for similar images based on this representation, using a
 combination of two inverted file systems. Since the patterns already encode loc
al spatial information, good results on several standard image retrieval dataset
s are obtained even without costly re-ranking based on geometric verification.
****

Event Detection in Complex Scenes Using Interval Temporal Constraints
Yifan Zhang, Qiang Ji, Hanqing Lu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3184-3191
In complex scenes with multiple atomic events happening sequentially or in parallel, detecting each individual event separately may not always obtain robust and reliable result. It is essential to detect them in a holistic way which incorporates the causality and temporal dependency among them to compensate the limitation of current computer vision techniques. In this paper, we propose an interval temporal constrained dynamic Bayesian network to extend Allen's interval algebra network (IAN) [2] from a deterministic static model to a probabilistic dynamic system, which can not only capture the complex interval temporal relationships, but also model the evolution dynamics and handle the uncertainty from the noisy visual observation. In the model, the topology of the IAN on each time slice and the interlinks between the time slices are discovered by an advanced structure learning method. The duration of the event and the unsynchronized time lags between two correlated event intervals are captured by a duration model, so that we can better determine the temporal boundary of the event. Empirical results on two real world datasets show the power of the proposed interval temporal constrained model.
****
Orderless Tracking through Model-Averaged Posterior Estimation
Seunghoon Hong, Suha Kwak, Bohyung Han; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2296-2303
We propose a novel offline tracking algorithm based on model-averaged posterior estimation through patch matching across frames. Contrary to existing online and offline tracking methods, our algorithm is not based on temporallyordered estimates of target state but attempts to select easyto-track frames first out of the remaining ones without exploiting temporal coherency of target. The posterior of the selected frame is estimated by propagating densities from the already tracked frames in a recursive manner. The density propagation across frames is implemented by an efficient patch matching technique, which is useful for our algorithm since it does not require motion smoothness assumption. Also, we present a hierarchical approach, where a small set of key frames are tracked first and non-key frames are handled by local key frames. Our tracking algorithm is conceptually well-suited for the sequences with abrupt motion, shot changes, and occlusion. We compare our tracking algorithm with existing techniques in real videos with such challenges and illustrate its superior performance qualitatively and quantitatively.
****
Log-Euclidean Kernels for Sparse Representation and Dictionary Learning
Peihua Li, Qilong Wang, Wangmeng Zuo, Lei Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1601-1608
The symmetric positive desnite (SPD) matrices have been widely used in image and vision problems. Recently there are growing interests in studying sparse representation (SR) of SPD matrices, motivated by the great success of SR for vector data. Though the space of SPD matrices is well-known to form a Lie group that is a Riemannian manifold, existing work fails to take full advantage of its geometric structure. This paper attempts to tackle this problem by proposing a kernel based method for SR and dictionary learning (DL) of SPD matrices. We disclose that the space of SPD matrices, with the operations of logarithmic multiplication and scalar logarithmic multiplication desned in the Log-Euclidean framework, is a complete inner product space. We can thus develop a broad family of kernels that satisses Mercer's condition. These kernels characterize the geodesic distance and can be computed efsciently. We also consider the geometric structure in the DL process by updating atom matrices in the Riemannian space instead of in the Euclidean space. The proposed method is evaluated with various vision problems and shows notable performance gains over state-of-the-arts.
****
A Rotational Stereo Model Based on XSlit Imaging
Jinwei Ye, Yu Ji, Jingyi Yu; Proceedings of the IEEE International Conference on

Computer Vision (ICCV), 2013, pp. 489-496
Traditional stereo matching assumes perspective viewing cameras under a translational motion: the second camera is translated away from the first one to create parallax. In this paper, we investigate a different, rotational stereo model on a special multi-perspective camera, the XSlit camera [9, 24]. We show that rotational XSlit (R-XSlit) stereo can be effectively created by fixing the sensor and slit locations but switching the two slits' directions. We first derive the epipolar geometry of R-XSlit in the 4D light field ray space. Our derivation leads to a simple but effective scheme for locating corresponding epipolar "curves". To conduct stereo matching, we further derive a new disparity term in our model and develop a patch-based graph-cut solution. To validate our theory, we assemble an XSlit lens by using a pair of cylindrical lenses coupled with slit-shaped apertures. The XSlit lens can be mounted on commodity cameras where the slit directions are adjustable to form desirable R-XSlit pairs. We show through experiments that R-XSlit provides a potentially advantageous imaging system for conducting fixed-location, dynamic baseline stereo.
****

Total Variation Regularization for Functions with Values in a Manifold
Jan Lellmann, Evgeny Strekalovskiy, Sabrina Koetter, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2944-2951
While total variation is among the most popular regularizers for variational problems, its extension to functions with values in a manifold is an open problem. In this paper, we propose the first algorithm to solve such problems which applies to arbitrary Riemannian manifolds. The key idea is to reformulate the variational problem as a multilabel optimization problem with an infinite number of labels. This leads to a hard optimization problem which can be approximately solved using convex relaxation techniques. The framework can be easily adapted to different manifolds including spheres and three-dimensional rotations, and allows to obtain accurate solutions even with a relatively coarse discretization. With numerous examples we demonstrate that the proposed framework can be applied to variational models that incorporate chromaticity values, normal fields, or camera trajectories.
****

Capturing Global Semantic Relationships for Facial Action Unit Recognition
Ziheng Wang, Yongqiang Li, Shangfei Wang, Qiang Ji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3304-3311
In this paper we tackle the problem of facial action unit (AU) recognition by exploiting the complex semantic relationships among AUs, which carry crucial top-down information yet have not been thoroughly exploited. Towards this goal, we build a hierarchical model that combines the bottom-level image features and the top-level AU relationships to jointly recognize AUs in a principled manner. The proposed model has two major advantages over existing methods. 1) Unlike methods that can only capture local pair-wise AU dependencies, our model is developed upon the restricted Boltzmann machine and therefore can exploit the global relationships among AUs. 2) Although AU relationships are influenced by many related factors such as facial expressions, these factors are generally ignored by the current methods. Our model, however, can successfully capture them to more accurately characterize the AU relationships. Efficient learning and inference algorithms of the proposed model are also developed. Experimental results on benchmark databases demonstrate the effectiveness of the proposed approach in modelling complex AU relationships as well as its superior AU recognition performance over existing approaches.
****

POP: Person Re-identification Post-rank Optimisation
Chunxiao Liu, Chen Change Loy, Shaogang Gong, Guijin Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 441-448
Owing to visual ambiguities and disparities, person reidentification methods inevitably produce suboptimal ranklist, which still requires exhaustive human eyeballing to identify the correct target from hundreds of different likelycandidates

. Existing re-identification studies focus on improving the ranking performance, but rarely look into the critical problem of optimising the time-consuming and error-prone post-rank visual search at the user end. In this study, we present a novel one-shot Post-rank OPtimisation (POP) method, which allows a user to quickly refine their search by either "one-shot" or a couple of sparse negative selections during a re-identification process. We conduct systematic behavioural studies to understand user's searching behaviour and show that the proposed method allows correct re-identification to converge 2.6 times faster than the conventional exhaustive search. Importantly, through extensive evaluations we demonstrate that the method is capable of achieving significant improvement over the stateof-the-art distance metric learning based ranking models, even with just "one shot" feedback optimisation, by as much as over 30% performance improvement for rank 1 reidentification on the VIPeR and i-LIDS datasets.

****

Joint Deep Learning for Pedestrian Detection

Wanli Ouyang, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2056-2063

Feature extraction, deformation handling, occlusion handling, and classi?cation are four important components in pedestrian detection. Existing methods learn or design these components either individually or sequentially. The interaction among these components is not yet well explored. This paper proposes that they should be jointly learned in order to maximize their strengths through cooperation. We formulate these four components into a joint deep learning framework and propose a new deep network architecture 1 . By establishing automatic, mutual interaction among components, the deep model achieves a 9% reduction in the average miss rate compared with the current best-performing pedestrian detection approaches on the largest Caltech benchmark dataset.

****

Visual Semantic Complex Network for Web Images

Shi Qiu, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3623-3630

This paper proposes modeling the complex web image collections with an automatically generated graph structure called visual semantic complex network (VSCN). The nodes on this complex network are clusters of images with both visual and semantic consistency, called semantic concepts. These nodes are connected based on the visual and semantic correlations. Our VSCN with 33, 240 concepts is generated from a collection of Nwwmillion web images. iiA great deal of valuable information on the structures of the web image collections can be revealed by exploring the VSCN, such as the small-world behavior, concept community, indegree distribution, hubs, and isolated concepts. It not only helps us better understand the web image collections at a macroscopic level, but also has many important practical applications. This paper presents two application examples: content-based image retrieval and image browsing. Experimental results show that the VSCN leads to significant improvement on both the precision of image retrieval (over 200%) and user experience for image browsing.

****

Multi-scale Topological Features for Hand Posture Representation and Analysis

Kaoning Hu, Lijun Yin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1928-1935

In this paper, we propose a multi-scale topological feature representation for automatic analysis of hand posture. Such topological features have the advantage of being posture-dependent while being preserved under certain variations of illumination, rotation, personal dependency, etc. Our method studies the topology of the holes between the hand region and its convex hull. Inspired by the principle of Persistent Homology, which is the theory of computational topology for topological feature analysis over multiple scales, we construct the multi-scale Betti Numbers matrix (MSBNM) for the topological feature representation. In our experiments, we used 12 different hand postures and compared our features with three popular features (HOG, MCT, and Shape Context) on different data sets. In addition to hand postures, we also extend the feature representations to arm posture

s. The results demonstrate the feasibility and reliability of the proposed metho
d.
****

An Enhanced Structure-from-Motion Paradigm Based on the Absolute Dual Quadric an
d Images of Circular Points
Lilian Calvet, Pierre Gurdjos; Proceedings of the IEEE International Conference
on Computer Vision (ICCV), 2013, pp. 985-992
This work aims at introducing a new unified Structurefrom-Motion (SfM) paradigm
in which images of circular point-pairs can be combined with images of natural p
oints. An imaged circular point-pair encodes the 2D Euclidean structure of a wor
ld plane and can easily be derived from the image of a planar shape, especially
those including circles. A classical SfM method generally runs two steps: first
a projective factorization of all matched image points (into projective cameras
and points) and second a camera selfcalibration that updates the obtained world
from projective to Euclidean. This work shows how to introduce images of circula
r points in these two SfM steps while its key contribution is to provide the the
oretical foundations for combining "classical" linear self-calibration constrain
ts with additional ones derived from such images. We show that the two proposed
SfM steps clearly contribute to better results than the classical approach. We v
alidate our contributions on synthetic and real images.
****

Understanding High-Level Semantics by Modeling Traffic Patterns
Hongyi Zhang, Andreas Geiger, Raquel Urtasun; Proceedings of the IEEE Internatio
nal Conference on Computer Vision (ICCV), 2013, pp. 3056-3063
In this paper, we are interested in understanding the semantics of outdoor scene
s in the context of autonomous driving. Towards this goal, we propose a generati
ve model of 3D urban scenes which is able to reason not only about the geometry
and objects present in the scene, but also about the high-level semantics in the
 form of traffic patterns. We found that a small number of patterns is sufficien
t to model the vast majority of traffic scenes and show how these patterns can b
e learned. As evidenced by our experiments, this high-level reasoning significan
tly improves the overall scene estimation as well as the vehicle-to-lane associa
tion when compared to state-of-the-art approaches [10].
****

PhotoOCR: Reading Text in Uncontrolled Conditions
Alessandro Bissacco, Mark Cummins, Yuval Netzer, Hartmut Neven; Proceedings of t
he IEEE International Conference on Computer Vision (ICCV), 2013, pp. 785-792
We describe PhotoOCR, a system for text extraction from images. Our particular f
ocus is reliable text extraction from smartphone imagery, with the goal of text
recognition as a user input modality similar to speech recognition. Commercially
 available OCR performs poorly on this task. Recent progress in machine learning
 has substantially improved isolated character classification; we build on this
progress by demonstrating a complete OCR system using these techniques. We also
incorporate modern datacenter-scale distributed language modelling. Our approach
 is capable of recognizing text in a variety of challenging imaging conditions w
here traditional OCR systems fail, notably in the presence of substantial blur,
low resolution, low contrast, high image noise and other distortions. It also op
erates with low latency; mean processing time is 600 ms per image. We evaluate o
ur system on public benchmark datasets for text extraction and outperform all pr
eviously reported results, more than halving the error rate on multiple benchmar
ks. The system is currently in use in many applications at Google, and is availa
ble as a user input modality in Google Translate for Android.
****

Support Surface Prediction in Indoor Scenes
Ruiqi Guo, Derek Hoiem; Proceedings of the IEEE International Conference on Comp
uter Vision (ICCV), 2013, pp. 2144-2151
In this paper, we present an approach to predict the extent and height of suppor
ting surfaces such as tables, chairs, and cabinet tops from a single RGBD image.
 We define support surfaces to be horizontal, planar surfaces that can physicall
y support objects and humans. Given a RGBD image, our goal is to localize the he

ight and full extent of such surfaces in 3D space. To achieve this, we created a labeling tool and annotated 1449 images with rich, complete 3D scene models in NYU dataset. We extract ground truth from the annotated dataset and developed a pipeline for predicting floor space, walls, the height and full extent of support surfaces. Finally we match the predicted extent with annotated scenes in training scenes and transfer the the support surface configuration from training scenes. We evaluate the proposed approach in our dataset and demonstrate its effectiveness in understanding scenes in 3D space.
****

Alternating Regression Forests for Object Detection and Pose Estimation
Samuel Schulter, Christian Leistner, Paul Wohlhart, Peter M. Roth, Horst Bischof; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 417-424
We present Alternating Regression Forests (ARFs), a novel regression algorithm that learns a Random Forest by optimizing a global loss function over all trees. This interrelates the information of single trees during the training phase and results in more accurate predictions. ARFs can minimize any differentiable regression loss without sacrificing the appealing properties of Random Forests, like low computational complexity during both, training and testing. Inspired by recent developments for classification [19], we derive a new algorithm capable of dealing with different regression loss functions, discuss its properties and investigate the relations to other methods like Boosted Trees. We evaluate ARFs on standard machine learning benchmarks, where we observe better generalization power compared to both standard Random Forests and Boosted Trees. Moreover, we apply the proposed regressor to two computer vision applications: object detection and head pose estimation from depth images. ARFs outperform the Random Forest baselines in both tasks, illustrating the importance of optimizing a common loss function for all trees.
****

Multi-attributed Dictionary Learning for Sparse Coding
Chen-Kuo Chiang, Te-Feng Su, Chih Yen, Shang-Hong Lai; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1137-1144
We present a multi-attributed dictionary learning algorithm for sparse coding. Considering training samples with multiple attributes, a new distance matrix is proposed by jointly incorporating data and attribute similarities. Then, an objective function is presented to learn categorydependent dictionaries that are compact (closeness of dictionary atoms based on data distance and attribute similarity), reconstructive (low reconstruction error with correct dictionary) and label-consistent (encouraging the labels of dictionary atoms to be similar). We have demonstrated our algorithm on action classification and face recognition tasks on several publicly available datasets. Experimental results with improved performance over previous dictionary learning methods are shown to validate the effectiveness of the proposed algorithm.
****

Similarity Metric Learning for Face Recognition
Qiong Cao, Yiming Ying, Peng Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2408-2415
Recently, there is a considerable amount of efforts devoted to the problem of unconstrained face verification, where the task is to predict whether pairs of images are from the same person or not. This problem is challenging and difficult due to the large variations in face images. In this paper, we develop a novel regularization framework to learn similarity metrics for unconstrained face verification. We formulate its objective function by incorporating the robustness to the large intra-personal variations and the discriminative power of novel similarity metrics. In addition, our formulation is a convex optimization problem which guarantees the existence of its global solution. Experiments show that our proposed method achieves the state-of-the-art results on the challenging Labeled Faces in the Wild (LFW) database [10].
****

Efficient Image Dehazing with Boundary Constraint and Contextual Regularization

Gaofeng Meng, Ying Wang, Jiangyong Duan, Shiming Xiang, Chunhong Pan; Proceeding
s of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 617-
624
suffer from bad visibility. In this paper, we propose an efficient regularizatio
n method to remove hazes from a single input image. Our method benefits much fro
m an exploration on the inherent boundary constraint on the transmission functio
n. This constraint, combined with a weighted ms1 sonffnbased contextual regulari
zation, is modeled into an optimization problem to estimate the unknown scene tr
ansmission. A quite efficient algorithm based on variable splitting is also pres
ented to solve the problem. The proposed method requires only a few general assu
mptions and can restore a high-quality haze-free image with faithful colors and
fine image details. Experimental results on a variety of haze images demonstrate
 the effectiveness and efficiency of the proposed method. Keywords-image process
ing; single image dehazing; visibility enhancement; I. I NTRODUCTION When one ta
kes a picture in foggy weather conditions, the obtained image often suffers from
 poor visibility. The distant objects in the fog lose the contrasts and get blur
red with
****

Robust Face Landmark Estimation under Occlusion
Xavier P. Burgos-Artizzu, Pietro Perona, Piotr Dollar; Proceedings of the IEEE I
nternational Conference on Computer Vision (ICCV), 2013, pp. 1513-1520
Human faces captured in real-world conditions present large variations in shape
and occlusions due to differences in pose, expression, use of accessories such a
s sunglasses and hats and interactions with objects (e.g. food). Current face la
ndmark estimation approaches struggle under such conditions since they fail to p
rovide a principled way of handling outliers. We propose a novel method, called
Robust Cascaded Pose Regression (RCPR) which reduces exposure to outliers by det
ecting occlusions explicitly and using robust shape-indexed features. We show th
at RCPR improves on previous landmark estimation methods on three popular face d
atasets (LFPW, LFW and HELEN). We further explore RCPR's performance by introduc
ing a novel face dataset focused on occlusion, composed of 1,007 faces presentin
g a wide range of occlusion patterns. RCPR reduces failure cases by half on all
four datasets, at the same time as it detects face occlusions with a 80/40% prec
ision/recall.
****

Finding Actors and Actions in Movies
P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, J. Sivic; Proceedings of
 the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2280-228
7
We address the problem of learning a joint model of actors and actions in movies
 using weak supervision provided by scripts. Specifically, we extract actor/acti
on pairs from the script and use them as constraints in a discriminative cluster
ing framework. The corresponding optimization problem is formulated as a quadrat
ic program under linear constraints. People in video are represented by automati
cally extracted and tracked faces together with corresponding motion features. F
irst, we apply the proposed framework to the task of learning names of character
s in the movie and demonstrate significant improvements over previous methods us
ed for this task. Second, we explore the joint actor/action constraint and show
its advantage for weakly supervised action learning. We validate our method in t
he challenging setting of localizing and recognizing characters and their action
s in feature length movies Casablanca and American Beauty.
****

Deblurring by Example Using Dense Correspondence
Yoav Hacohen, Eli Shechtman, Dani Lischinski; Proceedings of the IEEE Internatio
nal Conference on Computer Vision (ICCV), 2013, pp. 2384-2391
This paper presents a new method for deblurring photos using a sharp reference e
xample that contains some shared content with the blurry photo. Most previous de
blurring methods that exploit information from other photos require an accuratel
y registered photo of the same static scene. In contrast, our method aims to exp
loit reference images where the shared content may have undergone substantial ph

otometric and non-rigid geometric transformations, as these are the kind of refe
rence images most likely to be found in personal photo albums. Our approach buil
ds upon a recent method for examplebased deblurring using non-rigid dense corres
pondence (NRDC) [11] and extends it in two ways. First, we suggest exploiting in
formation from the reference image not only for blur kernel estimation, but also
 as a powerful local prior for the non-blind deconvolution step. Second, we intr
oduce a simple yet robust technique for spatially varying blur estimation, rathe
r than assuming spatially uniform blur. Unlike the above previous method, which
has proven successful only with simple deblurring scenarios, we demonstrate that
 our method succeeds on a variety of real-world examples. We provide quantitativ
e and qualitative evaluation of our method and show that it outperforms the stat
e-of-the-art.
****

High Quality Shape from a Single RGB-D Image under Uncalibrated Natural Illumina
tion

Yudeog Han, Joon-Young Lee, In So Kweon; Proceedings of the IEEE International C
onference on Computer Vision (ICCV), 2013, pp. 1617-1624

We present a novel framework to estimate detailed shape of diffuse objects with
uniform albedo from a single RGB-D image. To estimate accurate lighting in natur
al illumination environment, we introduce a general lighting model consisting of
 two components: global and local models. The global lighting model is estimated
 from the RGB-D input using the low-dimensional characteristic of a diffuse refl
ectance model. The local lighting model represents spatially varying illuminatio
n and it is estimated by using the smoothlyvarying characteristic of illuminatio
n. With both the global and local lighting model, we can estimate complex lighti
ng variations in uncontrolled natural illumination conditions accurately. For hi
gh quality shape capture, a shapefrom-shading approach is applied with the estim
ated lighting model. Since the entire process is done with a single RGB-D input,
 our method is capable of capturing the high quality shape details of a dynamic
object under natural illumination. Experimental results demonstrate the feasibil
ity and effectiveness of our method that dramatically improves shape details of
the rough depth input.
****

Discriminative Label Propagation for Multi-object Tracking with Sporadic Appeara
nce Features

K.C. Amit Kumar, Christophe De Vleeschouwer; Proceedings of the IEEE Internation
al Conference on Computer Vision (ICCV), 2013, pp. 2000-2007

Given a set of plausible detections, detected at each time instant independently
, we investigate how to associate them across time. This is done by propagating
labels on a set of graphs that capture how the spatio-temporal and the appearanc
e cues promote the assignment of identical or distinct labels to a pair of nodes
. The graph construction is driven by the locally linear embedding (LLE) of eith
er the spatio-temporal or the appearance features associated to the detections.
Interestingly, the neighborhood of a node in each appearance graph is defined to
 include all nodes for which the appearance feature is available (except the one
s that coexist at the same time). This allows to connect the nodes that share th
e same appearance even if they are temporally distant, which gives our framework
 the uncommon ability to exploit the appearance features that are available only
 sporadically along the sequence of detections. Once the graphs have been define
d, the multi-object tracking is formulated as the problem of finding a label ass
ignment that is consistent with the constraints captured by each of the graphs.
This results into a difference of convex program that can be efficiently solved.
 Experiments are performed on a basketball and several well-known pedestrian dat
asets in order to validate the effectiveness of the proposed solution.
****

Attribute Adaptation for Personalized Image Search

Adriana Kovashka, Kristen Grauman; Proceedings of the IEEE International Confere
nce on Computer Vision (ICCV), 2013, pp. 3432-3439

Current methods learn monolithic attribute predictors, with the assumption that
a single model is sufficient to reflect human understanding of a visual attribut

e. However, in reality, humans vary in how they perceive the association between a named property and image content. For example, two people may have slightly different internal models for what makes a shoe look "formal", or they may disagree on which of two scenes looks "more cluttered". Rather than discount these differences as noise, we propose to learn user-specific attribute models. We adapt a generic model trained with annotations from multiple users, tailoring it to satisfy user-specific labels. Furthermore, we propose novel techniques to infer user-specific labels based on transitivity and contradictions in the user's search history. We demonstrate that adapted attributes improve accuracy over both existing monolithic models as well as models that learn from scratch with user-specific data alone. In addition, we show how adapted attributes are useful to personalize image search, whether with binary or relative attributes.
****

Regionlets for Generic Object Detection
Xiaoyu Wang, Ming Yang, Shenghuo Zhu, Yuanqing Lin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 17-24
Generic object detection is confronted by dealing with different degrees of variations in distinct object classes with tractable computations, which demands for descriptive and flexible object representations that are also efficient to evaluate for many locations. In view of this, we propose to model an object class by a cascaded boosting classifier which integrates various types of features from competing local regions, named as regionlets. A regionlet is a base feature extraction region defined proportionally to a detection window at an arbitrary resolution (i.e. size and aspect ratio). These regionlets are organized in small groups with stable relative positions to delineate fine-grained spatial layouts inside objects. Their features are aggregated to a one-dimensional feature within one group so as to tolerate deformations. Then we evaluate the object bounding box proposal in selective search from segmentation cues, limiting the evaluation locations to thousands. Our approach significantly outperforms the state-of-the-art on popular multi-class detection benchmark datasets with a single method, without any contexts. It achieves the detection mean average precision of 41.7% on the PASCAL VOC 2007 dataset and 39.7% on the VOC 2010 for 20 object categories. It achieves 14.7% mean average precision on the ImageNet dataset for 200 object categories, outperforming the latest deformable part-based model (DPM) by 4.7%.
****

Event Recognition in Photo Collections with a Stopwatch HMM
Lukas Bossard, Matthieu Guillaumin, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1193-1200
The task of recognizing events in photo collections is central for automatically organizing images. It is also very challenging, because of the ambiguity of photos across different event classes and because many photos do not convey enough relevant information. Unfortunately, the field still lacks standard evaluation data sets to allow comparison of different approaches. In this paper, we introduce and release a novel data set of personal photo collections containing more than 61,000 images in 807 collections, annotated with 14 diverse social event classes. Casting collections as sequential data, we build upon recent and state-of-the-art work in event recognition in videos to propose a latent sub-event approach for event recognition in photo collections. However, photos in collections are sparsely sampled over time and come in bursts from which transpires the importance of specific moments for the photographers. Thus, we adapt a discriminative hidden Markov model to allow the transitions between states to be a function of the time gap between consecutive images, which we coin as Stopwatch Hidden Markov model (SHMM). In our experiments, we show that our proposed model outperforms approaches based only on feature pooling or a classical hidden Markov model. With an average accuracy of 56%, we also highlight the difficulty of the data set and the need for future advances in event recognition in photo collections.
****

Handling Occlusions with Franken-Classifiers
Markus Mathias, Rodrigo Benenson, Radu Timofte, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1505-1512

Detecting partially occluded pedestrians is challenging. A common practice to ma
ximize detection quality is to train a set of occlusion-specific classifiers, ea
ch for a certain amount and type of occlusion. Since training classifiers is exp
ensive, only a handful are typically trained. We show that by using many occlusi
on-specific classifiers, we outperform previous approaches on three pedestrian d
atasets; INRIA, ETH, and Caltech USA. We present a new approach to train such cl
assifiers. By reusing computations among different training stages, 16 occlusion
-specific classifiers can be trained at only one tenth the cost of one full trai
ning. We show that also test time cost grows sub-linearly.
****

Linear Sequence Discriminant Analysis: A Model-Based Dimensionality Reduction Me
thod for Vector Sequences
Bing Su, Xiaoqing Ding; Proceedings of the IEEE International Conference on Comp
uter Vision (ICCV), 2013, pp. 889-896
Dimensionality reduction for vectors in sequences is challenging since labels ar
e attached to sequences as a whole. This paper presents a model-based dimensiona
lity reduction method for vector sequences, namely linear sequence discriminant
analysis (LSDA) , which attempts to find a subspace in which sequences of the sa
me class are projected together while those of different classes are projected a
s far as possible. For each sequence class, an HMM is built from states of which
 statistics are extracted. Means of these states are linked in order to form a m
ean sequence, and the variance of the sequence class is defined as the sum of al
l variances of component states. LSDA then learns a transformation by maximizing
 the separability between sequence classes and at the same time minimizing the w
ithin-sequence class scatter. DTW distance between mean sequences is used to mea
sure the separability between sequence classes. We show that the optimization pr
oblem can be approximately transformed into an eigen decomposition problem. LDA
can be seen as a special case of LSDA by considering non-sequential vectors as s
equences of length one. The effectiveness of the proposed LSDA is demonstrated o
n two individual sequence datasets from UCI machine learning repository as well
as two concatenate sequence datasets: APTI Arabic printed text database and IFN/
ENIT Arabic handwriting database.
****

Learning Coupled Feature Spaces for Cross-Modal Matching
Kaiye Wang, Ran He, Wei Wang, Liang Wang, Tieniu Tan; Proceedings of the IEEE In
ternational Conference on Computer Vision (ICCV), 2013, pp. 2088-2095
Cross-modal matching has recently drawn much attention due to the widespread exi
stence of multimodal data. It aims to match data from different modalities, and
generally involves two basic problems: the measure of relevance and coupled feat
ure selection. Most previous works mainly focus on solving the first problem. In
 this paper, we propose a novel coupled linear regression framework to deal with
 both problems. Our method learns two projection matrices to map multimodal data
 into a common feature space, in which cross-modal data matching can be performe
d. And in the learning procedure, the 21 -norm penalties are imposed on the two
projection matrices separately, which leads to select relevant and discriminativ
e features from coupled feature spaces simultaneously. A trace norm is further i
mposed on the projected data as a low-rank constraint, which enhances the releva
nce of different modal data with connections. We also present an iterative algor
ithm based on halfquadratic minimization to solve the proposed regularized linea
r regression problem. The experimental results on two challenging cross-modal da
tasets demonstrate that the proposed method outperforms the state-of-the-art app
roaches.
****

Structured Light in Sunlight
Mohit Gupta, Qi Yin, Shree K. Nayar; Proceedings of the IEEE International Confe
rence on Computer Vision (ICCV), 2013, pp. 545-552
Strong ambient illumination severely degrades the performance of structured ligh
t based techniques. This is especially true in outdoor scenarios, where the stru
ctured light sources have to compete with sunlight, whose power is often 2-5 ord
ers of magnitude larger than the projected light. In this paper, we propose the

concept of light-concentration to overcome strong ambient illumination. Our key observation is that given a fixed light (power) budget, it is always better to allocate it sequentially in several portions of the scene, as compared to spreading it over the entire scene at once. For a desired level of accuracy, we show that by distributing light appropriately, the proposed approach requires 1-2 orders lower acquisition time than existing approaches. Our approach is illumination-adaptive as the optimal light distribution is determined based on a measurement of the ambient illumination level. Since current light sources have a fixed light distribution, we have built a prototype light source that supports flexible light distribution by controlling the scanning speed of a laser scanner. We show several high quality 3D scanning results in a wide range of outdoor scenarios. The proposed approach will benefit 3D vision systems that need to operate outdoors under extreme ambient illumination levels on a limited time and power budget.
****

Probabilistic Elastic Part Model for Unsupervised Face Detector Adaptation
Haoxiang Li, Gang Hua, Zhe Lin, Jonathan Brandt, Jianchao Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 793-800
We propose an unsupervised detector adaptation algorithm to adapt any offline trained face detector to a specific collection of images, and hence achieve better accuracy. The core of our detector adaptation algorithm is a probabilistic elastic part (PEP) model, which is offline trained with a set of face examples. It produces a statisticallyaligned part based face representation, namely the PEP representation. To adapt a general face detector to a collection of images, we compute the PEP representations of the candidate detections from the general face detector, and then train a discriminative classifier with the top positives and negatives. Then we re-rank all the candidate detections with this classifier. This way, a face detector tailored to the statistics of the specific image collection is adapted from the original detector. We present extensive results on three datasets with two state-of-the-art face detectors. The significant improvement of detection accuracy over these stateof-the-art face detectors strongly demonstrates the efficacy of the proposed face detector adaptation algorithm.
****

Robust Non-parametric Data Fitting for Correspondence Modeling
Wen-Yan Lin, Ming-Ming Cheng, Shuai Zheng, Jiangbo Lu, Nigel Crook; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2376-2383
We propose a generic method for obtaining nonparametric image warps from noisy point correspondences. Our formulation integrates a huber function into a motion coherence framework. This makes our fitting function especially robust to piecewise correspondence noise (where an image section is consistently mismatched). By utilizing over parameterized curves, we can generate realistic nonparametric image warps from very noisy correspondence. We also demonstrate how our algorithm can be used to help stitch images taken from a panning camera by warping the images onto a virtual push-broom camera imaging plane.
****

A Convex Optimization Framework for Active Learning
Ehsan Elhamifar, Guillermo Sapiro, Allen Yang, S. Shankar Sasrty; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 209-216
In many image/video/web classification problems, we have access to a large number of unlabeled samples. However, it is typically expensive and time consuming to obtain labels for the samples. Active learning is the problem of progressively selecting and annotating the most informative unlabeled samples, in order to obtain a high classification performance. Most existing active learning algorithms select only one sample at a time prior to retraining the classifier. Hence, they are computationally expensive and cannot take advantage of parallel labeling systems such as Mechanical Turk. On the other hand, algorithms that allow the selection of multiple samples prior to retraining the classifier, may select samples that have significant information overlap or they involve solving a non-convex optimization. More importantly, the majority of active learning algorithms are developed for a certain classifier type such as SVM. In this paper, we develop an

efficient active learning framework based on convex programming, which can select multiple samples at a time for annotation. Unlike the state of the art, our algorithm can be used in conjunction with any type of classifiers, including those of the family of the recently proposed Sparse Representation-based Classification (SRC). We use the two principles of classifier uncertainty and sample diversity in order to guide the optimization program towards selecting the most informative unlabeled samples, which have the least information overlap. Our method can incorporate the data distribution in the selection process by using the appropriate dissimilarity between pairs of samples. We show the effectiveness of our framework in person detection, scene categorization and face recognition on real-world datasets.
****
Joint Noise Level Estimation from Personal Photo Collections
Yichang Shih, Vivek Kwatra, Troy Chinen, Hui Fang, Sergey Ioffe; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2896-2903
Personal photo albums are heavily biased towards faces of people, but most state-of-the-art algorithms for image denoising and noise estimation do not exploit facial information. We propose a novel technique for jointly estimating noise levels of all face images in a photo collection. Photos in a personal album are likely to contain several faces of the same people. While some of these photos would be clean and high quality, others may be corrupted by noise. Our key idea is to estimate noise levels by comparing multiple images of the same content that differ predominantly in their noise content. Specifically, we compare geometrically and photometrically aligned face images of the same person. Our estimation algorithm is based on a probabilistic formulation that seeks to maximize the joint probability of estimated noise levels across all images. We propose an approximate solution that decomposes this joint maximization into a two-stage optimization. The first stage determines the relative noise between pairs of images by pooling estimates from corresponding patch pairs in a probabilistic fashion. The second stage then jointly optimizes for all absolute noise parameters by conditioning them upon relative noise levels, which allows for a pairwise factorization of the probability distribution. We evaluate our noise estimation method using quantitative experiments to measure accuracy on synthetic data. Additionally, we employ the estimated noise levels for automatic denoising using "BM3D", and evaluate the quality of denoising on real-world photos through a user study.
****