Deformable Spatial Pyramid Matching for Fast Dense Correspondences

Jaechul Kim, Ce Liu, Fei Sha, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2307-2314

We introduce a fast deformable spatial pyramid (DSP) matching algorithm for computing dense pixel correspondences. Dense matching methods typically enforce both appearance agreement between matched pixels as well as geometric smoothness between neighboring pixels. Whereas the prevailing approaches operate at the pixel level, we propose a pyramid graph model that simultaneously regularizes match consistency at multiple spatial extents--ranging from an entire image, to coarse grid cells, to every single pixel. This novel regularization substantially improves pixel-level matching in the face of challenging image variations, while the "deformable" aspect of our model overcomes the strict rigidity of traditional spatial pyramids. Results on LabelMe and Caltech show our approach outperforms state-of-the-art methods (SIFT Flow [15] and PatchMatch [2]), both in terms of accuracy and run time.

****

A Genetic Algorithm-Based Solver for Very Large Jigsaw Puzzles

Dror Sholomon, Omid David, Nathan S. Netanyahu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1767-1774

In this paper we propose the first effective automated, genetic algorithm (GA)-based jigsaw puzzle solver. We introduce a novel procedure of merging two "parent" solutions to an improved "child" solution by detecting, extracting, and combining correctly assembled puzzle segments. The solver proposed exhibits state-of-the-art performance solving previously attempted puzzles faster and far more accurately, and also puzzles of size never before attempted. Other contributions include the creation of a benchmark of large images, previously unavailable. We share the data sets and all of our results for future testing and comparative evaluation of jigsaw puzzle solvers.

****

Exploring Compositional High Order Pattern Potentials for Structured Output Learning

Yujia Li, Daniel Tarlow, Richard Zemel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 49-56

When modeling structured outputs such as image segmentations, prediction can be improved by accurately modeling structure present in the labels. A key challenge is developing tractable models that are able to capture complex high level structure like shape. In this work, we study the learning of a general class of pattern-like high order potential, which we call Compositional High Order Pattern Potentials (CHOPPs). We show that CHOPPs include the linear deviation pattern potentials of Rother et al. [26] and also Restricted Boltzmann Machines (RBMs); we also establish the near equivalence of these two models. Experimentally, we show that performance is affected significantly by the degree of variability present in the datasets, and we define a quantitative variability measure to aid in studying this. We then improve CHOPPs performance in high variability datasets with two primary contributions: (a) developing a loss-sensitive joint learning procedure, so that internal pattern parameters can be learned in conjunction with other model potentials to minimize expected loss;and (b) learning an image-dependent mapping that encourages or inhibits patterns depending on image features. We also explore varying how multiple patterns are composed, and learning convolutional patterns. Quantitative results on challenging highly variable datasets show that the joint learning and image-dependent high order potentials can improve performance.

****

Hyperbolic Harmonic Mapping for Constrained Brain Surface Registration

Rui Shi, Wei Zeng, Zhengyu Su, Hanna Damasio, Zhonglin Lu, Yalin Wang, Shing-Tung Yau, Xianfeng Gu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2531-2538

Automatic computation of surface correspondence via harmonic map is an active research field in computer vision, computer graphics and computational geometry. It may help document and understand physical and biological phenomena and also ha s broad applications in biometrics, medical imaging and motion capture. Although numerous studies have been devoted to harmonic map research, limited progress has been made to compute a diffeomorphic harmonic map on general topology surfaces with landmark constraints. This work conquer this problem by changing the Riemannian metric on the target surface to a hyperbolic metric, so that the harmonic mapping is guaranteed to be a diffeomorphism under landmark constraints. The computational algorithms are based on the Ricci flow method and the method is general and robust. We apply our algorithm to study constrained human brain surface registration problem. Experimental results demonstrate that, by changing the Riemannian metric, the registrations are always diffeomorphic, and achieve relative high performance when evaluated with some popular cortical surface registration evaluation standards.

****

Dense Variational Reconstruction of Non-rigid Surfaces from Monocular Video

Ravi Garg, Anastasios Roussos, Lourdes Agapito; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1272-1279

This paper offers the first variational approach to the problem of dense 3D reconstruction of non-rigid surfaces from a monocular video sequence. We formulate nonrigid structure from motion (NRS f M) as a global variational energy minimization problem to estimate dense low-rank smooth 3D shapes for every frame along with the camera motion matrices, given dense 2D correspondences. Unlike traditional factorization based approaches to NRS f M , which model the low-rank non-rigid shape using a fixed number of basis shapes and corresponding coefficients, we minimize the rank of the matrix of time-varying shapes directly via trace norm minimization. In conjunction with this low-rank constraint, we use an edge preserving total-variation regularization term to obtain spatially smooth shapes for every frame. Thanks to proximal splitting techniques the optimization problem can be decomposed into many point-wise sub-problems and simple linear systems which can be easily solved on GPU hardware. We show results on real sequences of different objects (face, torso, beating heart) where, despite challenges in tracking, illumination changes and occlusions, our method reconstructs highly deforming smooth surfaces densely and accurately directly from video, without the need for any prior models or shape templates.

****

Fusing Depth from Defocus and Stereo with Coded Apertures

Yuichi Takeda, Shinsaku Hiura, Kosuke Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 209-216

In this paper we propose a novel depth measurement method by fusing depth from defocus (DFD) and stereo. One of the problems of passive stereo method is the difficulty of finding correct correspondence between images when an object has a repetitive pattern or edges parallel to the epipolar line. On the other hand, the accuracy of DFD method is inherently limited by the effective diameter of the lens. Therefore, we propose the fusion of stereo method and DFD by giving different focus distances for left and right cameras of a stereo camera with coded apertures. Two types of depth cues, defocus and disparity, are naturally integrated by the magnification and phase shift of a single point spread function (PSF) per camera. In this paper we give the proof of the proportional relationship between the diameter of defocus and disparity which makes the calibration easy. We also show the outstanding performance of our method which has both advantages of two depth cues through simulation and actual experiments.

****

A Non-parametric Framework for Document Bleed-through Removal

Roisin Rowley-Brooke, Francois Pitie, Anil Kokaram; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2954-2960

This paper presents recent work on a new framework for non-blind document bleed-through removal. The framework includes image preprocessing to remove local intensity variations, pixel region classification based on a segmentation of the joint recto-verso intensity histogram and connected component analysis on the subsequent image labelling. Finally restoration of the degraded regions is performed using exemplar-based image inpainting. The proposed method is evaluated visually and numerically on a freely available database of 25 scanned manuscript image pairs with ground truth, and is shown to outperform recent non-blind bleed-through removal techniques.

****

A Comparative Study of Modern Inference Techniques for Discrete Energy Minimization Problems

J. Kappes, B. Andres, F. Hamprecht, C. Schnorr, S. Nowozin, D. Batra, S. Kim, B. Kausler, J. Lellmann, N. Komodakis, C. Rother; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1328-1335

Seven years ago, Szeliski et al. published an influential study on energy minimization methods for Markov random fields (MRF). This study provided valuable insights in choosing the best optimization technique for certain classes of problems. While these insights remain generally useful today, the phenominal success of random field models means that the kinds of inference problems we solve have changed significantly. Specifically, the models today often include higher order interactions, flexible connectivity structures, large label-spaces of different cardinalities, or learned energy tables. To reflect these changes, we provide a modernized and enlarged study. We present an empirical comparison of 24 state-of-art techniques on a corpus of 2,300 energy minimization instances from 20 diverse computer vision applications. To ensure reproducibility, we evaluate all methods in the OpenGM2 framework and report extensive results regarding runtime and solution quality. Key insights from our study agree with the results of Szeliski et al. for the types of models they studied. However, on new and challenging types of models our findings disagree and suggest that polyhedral methods and integer programming solvers are competitive in terms of runtime and solution quality over a large range of model types.

****

Submodular Salient Region Detection

Zhuolin Jiang, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2043-2050

The problem of salient region detection is formulated as the well-studied facility location problem from operations research. High-level priors are combined with low-level features to detect salient regions. Salient region detection is achieved by maximizing a submodular objective function, which maximizes the total similarities (i.e., total profits) between the hypothesized salient region centers (i.e., facility locations) and their region elements (i.e., clients), and penalizes the number of potential salient regions (i.e., the number of open facilities). The similarities are efficiently computed by finding a closed-form harmonic solution on the constructed graph for an input image. The saliency of a selected region is modeled in terms of appearance and spatial location. By exploiting the submodularity properties of the objective function, a highly efficient greedy-based optimization algorithm can be employed. This algorithm is guaranteed to be at least a (e 1)/e ? 0.632-approximation to the optimum. Experimental results demonstrate that our approach outperforms several recently proposed saliency detection approaches.

****

Spatio-temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera

Lu Xia, J.K. Aggarwal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2834-2841

Local spatio-temporal interest points (STIPs) and the resulting features from RGB videos have been proven successful at activity recognition that can handle cluttered backgrounds and partial occlusions. In this paper, we propose its counterpart in depth video and show its efficacy on activity recognition. We present a filtering method to extract STIPs from depth videos (called DSTIP) that effectively suppress the noisy measurements. Further, we build a novel depth cuboid similarity feature (DCSF) to describe the local 3D depth cuboid around the DSTIPs with an adaptable supporting size. We test this feature on activity recognition application using the public MSRAction3D, MSRDailyActivity3D datasets and our own dataset. Experimental evaluation shows that the proposed approach outperforms st ateof-the-art activity recognition algorithms on depth videos, and the framework is more widely applicable than existing approaches. We also give detailed comparisons with other features and analysis of choice of parameters as a guidance for applications.

****

Bringing Semantics into Focus Using Visual Abstraction

C. L. Zitnick, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3009-3016

Relating visual information to its linguistic semantic meaning remains an open and challenging area of research. The semantic meaning of images depends on the presence of objects, their attributes and their relations to other objects. But precisely characterizing this dependence requires extracting complex visual information from an image, which is in general a difficult and yet unsolved problem. In this paper, we propose studying semantic information in abstract images created from collections of clip art. Abstract images provide several advantages. They allow for the direct study of how to infer high-level semantic information, since they remove the reliance on noisy low-level object, attribute and relation detectors, or the tedious hand-labeling of images. Importantly, abstract images also allow the ability to generate sets of semantically similar scenes. Finding analogous sets of semantically similar real images would be nearly impossible. We create 1,002 sets of 10 semantically similar abstract scenes with corresponding written descriptions. We thoroughly analyze this dataset to discover semantically important features, the relations of words to visual features and methods for measuring semantic similarity.

****

Fast Multiple-Part Based Object Detection Using KD-Ferns

Dan Levi, Shai Silberstein, Aharon Bar-Hillel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 947-954

In this work we present a new part-based object detection algorithm with hundreds of parts performing realtime detection. Part-based models are currently state-ofthe-art for object detection due to their ability to represent large appearance variations. However, due to their high computational demands such methods are limited to several parts only and are too slow for practical real-time implementation. Our algorithm is an accelerated version of the "Feature Synthesis" (FS) method [1], which uses multiple object parts for detection and is among state-of-theart methods on human detection benchmarks, but also suffers from a high computational cost. The proposed Accelerated Feature Synthesis (AFS) uses several strategies for reducing the number of locations searched for each part. The first strategy uses a novel algorithm for approximate nearest neighbor search which we developed, termed "KDFerns", to compare each image location to only a subset of the model parts. Candidate part locations for a specific part are further reduced using spatial inhibition, and using an object-level "coarse-to-fine" strategy. In our empirical evaluation on pedestrian detection benchmarks, AFS maintains almost fully the accuracy performance of the original FS, while running more than x4 faster than existing partbased methods which use only several parts. AFS is to our best knowledge the first part-based object detection method achieving real-time running performance: nearly 10 frames per-second on 640 x 480 images on a regular CPU.

****

Computing Diffeomorphic Paths for Large Motion Interpolation

Dohyung Seo, Jeffrey Ho, Baba C. Vemuri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1227-1232

In this paper, we introduce a novel framework for computing a path of diffeomorphisms between a pair of input diffeomorphisms. Direct computation of a geodesic path on the space of diffeomorphisms Diff(?) is difficult, and it can be attributed mainly to the infinite dimensionality of Diff(?). Our proposed framework, to some degree, bypasses this difficulty using the quotient map of Diff(?) to the quotient space Diff(M)/Diff(M) ? obtained by quotienting out the subgroup of volume-preserving diffeomorphisms Diff(M) ? . This quotient space was recently identified as the unit sphere in a Hilbert space in mathematics literature, a spa ce with well-known geometric properties. Our framework leverages this recent result by computing the diffeomorphic path in two stages. First, we project the given diffeomorphism pair onto this sphere and then compute the geodesic path between these projected points. Second, we lift the geodesic on the sphere back to the space of diffeomerphisms, by solving a quadratic programming problem with bilinear constraints using the augmented Lagrangian technique with penalty terms. In this way, we can estimate the path of diffeomorphisms, first, staying in the space of diffeomorphisms, and second, preserving shapes/volumes in the deformed images along the path as much as possible. We have applied our framework to interpolate intermediate frames of frame-sub-sampled video sequences. In the reported experiments, our approach compares favorably with the popular Large Deformation Diffeomorphic Metric Mapping framework (LDDMM).

****

Wide-Baseline Hair Capture Using Strand-Based Refinement

Linjie Luo, Cha Zhang, Zhengyou Zhang, Szymon Rusinkiewicz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 265-272

We propose a novel algorithm to reconstruct the 3D geometry of human hairs in wide-baseline setups using strand-based refinement. The hair strands are first extracted in each 2D view, and projected onto the 3D visual hull for initialization. The 3D positions of these strands are then refined by optimizing an objective function that takes into account cross-view hair orientation consistency, the visual hull constraint and smoothness constraints defined at the strand, wisp and global levels. Based on the refined strands, the algorithm can reconstruct an approximate hair surface: experiments with synthetic hair models achieve an accuracy of ~3mm. We also show real-world examples to demonstrate the capability to capture full-head hair styles as well as hair in motion with as few as 8 cameras.

****

Radial Distortion Self-Calibration

Jose Henrique Brito, Roland Angst, Kevin Koser, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1368-1375

In cameras with radial distortion, straight lines in space are in general mapped to curves in the image. Although epipolar geometry also gets distorted, there is a set of special epipolar lines that remain straight, namely those that go through the distortion center. By finding these straight epipolar lines in camera pairs we can obtain constraints on the distortion center(s) without any calibration object or plumbline assumptions in the scene. Although this holds for all radial distortion models we conceptually prove this idea using the division distortion model and the radial fundamental matrix which allow for a very simple closed form solution of the distortion center from two views (same distortion) or three views (different distortions). The non-iterative nature of our approach makes it immune to local minima and allows finding the distortion center also for cropped images or those where no good prior exists. Besides this, we give comprehensive relations between different undistortion models and discuss advantages and drawbacks.

****

Separating Signal from Noise Using Patch Recurrence across Scales

Maria Zontak, Inbar Mosseri, Michal Irani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1195-1202

Recurrence of small clean image patches across different scales of a natural image has been successfully used for solving ill-posed problems in clean images (e.g., superresolution from a single image). In this paper we show how this multi-scale property can be extended to solve ill-posed problems under noisy conditions, such as image denoising. While clean patches are obscured by severe noise in the original scale of a noisy image, noise levels drop dramatically at coarser image scales. This allows for the unknown hidden clean patches to "naturally emerge" in some coarser scale of the noisy image. We further show that patch recurrence across scales is strengthened when using directional pyramids (that blur and subsample only in one direction). Our statistical experiments show that for almo st any noisy image patch (more than 99%), there exists a "good" clean version of itself at the same relative image coordinates in some coarser scale of the image. This is a strong phenomenon of noise-contaminated natural images, which can serve as a strong prior for separating the signal from the noise. Finally, incorporating this multi-scale prior into a simple denoising algorithm yields state-of-the-art denoising results.

****

Detection Evolution with Multi-order Contextual Co-occurrence

Guang Chen, Yuanyuan Ding, Jing Xiao, Tony X. Han; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1798-1805

Context has been playing an increasingly important role to improve the object detection performance. In this paper we propose an effective representation, Multi-Order Contextual co-Occurrence (MOCO), to implicitly model the high level context using solely detection responses from a baseline object detector. The so-called (1 st -order) context feature is computed as a set of randomized binary comparisons on the response map of the baseline object detector. The statistics of the 1 st -order binary context features are further calculated to construct a high order co-occurrence descriptor. Combining the MOCO feature with the original image feature, we can evolve the baseline object detector to a stronger context aware detector. With the updated detector, we can continue the evolution till the contextual improvements saturate. Using the successful deformable-partmodel detector [13] as the baseline detector, we test the proposed MOCO evolution framework on the PASCAL VOC 2007 dataset [8] and Caltech pedestrian dataset [7]: The proposed MOCO detector outperforms all known state-ofthe-art approaches, contextually boosting deformable part models (ver.5) [13] by 3.3% in mean average precision on the PASCAL 2007 dataset. For the Caltech pedestrian dataset, our method further reduces the log-average miss rate from 48% to 46% and the miss rate at 1 FPPI from 2atfto m44%%compared with the best prior art [6].

****

Manhattan Scene Understanding via XSlit Imaging

Jinwei Ye, Yu Ji, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 81-88

A Manhattan World (MW) [3] is composed of planar surfaces and parallel lines aligned with three mutually orthogonal principal axes. Traditional MW understanding algorithms rely on geometry priors such as the vanishing points and reference (ground) planes for grouping coplanar structures. In this paper, we present a novel single-image MW reconstruction algorithm from the perspective of nonpinhole cameras. We show that by acquiring the MW using an XSlit camera, we can instantly resolve coplanarity ambiguities. Specifically, we prove that parallel 3D lines map to 2D curves in an XSlit image and they converge at an XSlit Vanishing Point (XVP). In addition, if the lines are coplanar, their curved images will intersect at a second common pixel that we call Coplanar Common Point (CCP). CCP is a unique image feature in XSlit cameras that does not exist in pinholes. We present a comprehensive theory to analyze XVPs and CCPs in a MW scene and study how to recover 3D geometry in a complex MW scene from XVPs and CCPs. Finally, we build a prototype XSlit camera by using two layers of cylindrical lenses. Experimental results on both synthetic and real data show that our new XSlitcamera-based solution provides an effective and reliable solution for MW understanding.

****

Cumulative Attribute Space for Age and Crowd Density Estimation

Ke Chen, Shaogang Gong, Tao Xiang, Chen Change Loy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2467-2474

A number of computer vision problems such as human age estimation, crowd density estimation and body/face pose (view angle) estimation can be formulated as a regression problem by learning a mapping function between a high dimensional vector-formed feature input and a scalarvalued output. Such a learning problem is made difficult due to sparse and imbalanced training data and large feature variations caused by both uncertain viewing conditions and intrinsic ambiguities between observable visual features and the scalar values to be estimated. Encouraged by the recent success in using attributes for solving classification problems wit h sparse training data, this paper introduces a novel cumulative attribute concept for learning a regression model when only sparse and imbalanced data are available. More precisely, low-level visual features extracted from sparse and imbalanced image samples are mapped onto a cumulative attribute space where each dimension has clearly defined semantic interpretation (a label) that captures how the scalar output value (e.g. age, people count) changes continuously and cumulatively. Extensive experiments show that our cumulative attribute framework gains notable advantage on accuracy for both age estimation and crowd counting when compared against conventional regression models, especially when the labelled training data is sparse with imbalanced sampling.

****

Tensor-Based High-Order Semantic Relation Transfer for Semantic Scene Segmentation

Heesoo Myeong, Kyoung Mu Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3073-3080

We propose a novel nonparametric approach for semantic segmentation using high-order semantic relations. Conventional context models mainly focus on learning pairwise relationships between objects. Pairwise relations, however, are not enough to represent high-level contextual knowledge within images. In this paper, we propose semantic relation transfer, a method to transfer high-order semantic relations of objects from annotated images to unlabeled images analogous to label transfer techniques where label information are transferred. We first define semantic tensors representing high-order relations of objects. Semantic relation transfer problem is then formulated as semi-supervised learning using a quadratic objective function of the semantic tensors. By exploiting low-rank property of the semantic tensors and employing Kronecker sum similarity, an efficient approximation algorithm is developed. Based on the predicted high-order semantic relations, we reason semantic segmentation and evaluate the performance on several challenging datasets.

****

Accurate and Robust Registration of Nonrigid Surface Using Hierarchical Statistical Shape Model

Hidekata Hontani, Yuto Tsunekawa, Yoshihide Sawada; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2977-2984

In this paper, we propose a new non-rigid robust registration method that registers a point distribution model (PDM) of a surface to given 3D images. The contributions of the paper are (1) a new hierarchical statistical shape model (SSM) of the surface that has better generalization ability is introduced, (2) the registration algorithm of the hierarchical SSM that can estimate the marginal posterior distribution of the surface location is proposed, and (3) the registration performance is improved by (3-1) robustly registering each local shape of the surface with the sparsity regularization and by (3-2) referring to the appearance between the neighboring model points in the likelihood computation. The SSM of a liver was constructed from a set of clinical CT images, and the performance of the proposed method was evaluated. Experimental results demonstrated that the proposed method outperformed some existing methods that use non-hierarchical SSMs.

****

POOF: Part-Based One-vs.-One Features for Fine-Grained Categorization, Face Verification, and Attribute Estimation

Thomas Berg, Peter N. Belhumeur; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 955-962

From a set of images in a particular domain, labeled with part locations and class, we present a method to automatically learn a large and diverse set of highly discriminative intermediate features that we call Part-based One-vs-One Features (POOFs). Each of these features specializes in discrimination between two particular classes based on the appearance at a particular part. We demonstrate the particular usefulness of these features for fine-grained visual categorization with new state-of-the-art results on bird species identification using the Caltech UCSD Birds (CUB) dataset and parity with the best existing results in face verification on the Labeled Faces in the Wild (LFW) dataset. Finally, we demonstrat e the particular advantage of POOFs when training data is scarce.

****

Sparse Quantization for Patch Description

Xavier Boix, Michael Gygli, Gemma Roig, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2842-2849

The representation of local image patches is crucial for the good performance and efficiency of many vision tasks. Patch descriptors have been designed to generalize towards diverse variations, depending on the application, as well as the desired compromise between accuracy and efficiency. We present a novel formulation of patch description, that serves such issues well. Sparse quantization lies at its heart. This allows for efficient encodings, leading to powerful, novel binary descriptors, yet also to the generalization of existing descriptors like SIFT or BRIEF. We demonstrate the capabilities of our formulation for both keypoint matching and image classification. Our binary descriptors achieve state-of-the-art results for two keypoint matching benchmarks, namely those by Brown [6] and Mikolajczyk [18]. For image classification, we propose new descriptors that perform similar to SIFT on Caltech101 [10] and PASCAL VOC07 [9].

****

What's in a Name? First Names as Facial Attributes

Huizhong Chen, Andrew C. Gallagher, Bernd Girod; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3366-3373

This paper introduces a new idea in describing people using their first names, i.e., the name assigned at birth. We show that describing people in terms of similarity to a vector of possible first names is a powerful description of facial appearance that can be used for face naming and building facial attribute classifiers. We build models for 100 common first names used in the United States and for each pair, construct a pairwise firstname classifier. These classifiers are built using training images downloaded from the internet, with no additional user interaction. This gives our approach important advantages in building practical systems that do not require additional human intervention for labeling. We use the scores from each pairwise name classifier as a set of facial attributes. We show several surprising results. Our name attributes predict the correct first names of test faces at rates far greater than chance. The name attributes are applied to gender recognition and to age classification, outperforming state-of-the-art methods with all training images automatically gathered from the internet.

****

Context-Aware Modeling and Recognition of Activities in Video

Yingying Zhu, Nandita M. Nayak, Amit K. Roy-Chowdhury; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2491-2498

In this paper, rather than modeling activities in videos individually, we propose a hierarchical framework that jointly models and recognizes related activities using motion and various context features. This is motivated from the observations that the activities related in space and time rarely occur independently and can serve as the context for each other. Given a video, action segments are automatically detected using motion segmentation based on a nonlinear dynamical model. We aim to merge these segments into activities of interest and generate optimum labels for the activities. Towards this goal, we utilize a structural model in a max-margin framework that jointly models the underlying activities which are related in space and time. The model explicitly learns the duration, motion and context patterns for each activity class, as well as the spatio-temporal relationships for groups of them. The learned model is then used to optimally label the activities in the testing videos using a greedy search method. We show promising results on the VIRAT Ground Dataset demonstrating the benefit of joint modeling and recognizing activities in a wide-area scene.

****

Learning to Detect Partially Overlapping Instances

Carlos Arteta, Victor Lempitsky, J. A. Noble, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3230-3237

The objective of this work is to detect all instances of a class (such as cells or people) in an image. The instances may be partially overlapping and clustered, and hence quite challenging for traditional detectors, which aim at localizing individual instances. Our approach is to propose a set of candidate regions, and then select regions based on optimizing a global classification score, subject to the constraint that the selected regions are non-overlapping. Our novel contribution is to extend standard object detection by introducing separate classes for tuples of objects into the detection process. For example, our detector can pick a region containing two or three object instances, while assigning such region an appropriate label. We show that this formulation can be learned within the structured output SVM framework, and that the inference in such model can be accomplished using dynamic programming on a tree structured region graph. Furthermore, the learning only requires weak annotations a dot on each instance. The improvement resulting from the addition of the capability to detect tuples of objects is demonstrated on quite disparate data sets: fluorescence microscopy images and UCSD pedestrians.

****

Exemplar-Based Face Parsing

Brandon M. Smith, Li Zhang, Jonathan Brandt, Zhe Lin, Jianchao Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3484-3491

In this work, we propose an exemplar-based face image segmentation algorithm. We take inspiration from previous works on image parsing for general scenes. Our approach assumes a database of exemplar face images, each of which is associated with a hand-labeled segmentation map. Given a test image, our algorithm first selects a subset of exemplar images from the database, Our algorithm then computes a nonrigid warp for each exemplar image to align it with the test image. Finally, we propagate labels from the exemplar images to the test image in a pixel-wise manner, using trained weights to modulate and combine label maps from different exemplars. We evaluate our method on two challenging datasets and compare with two face parsing algorithms and a general scene parsing algorithm. We also compare our segmentation results with contour-based face alignment results; that is, we first run the alignment algorithms to extract contour points and then derive segments from the contours. Our algorithm compares favorably with all previous works on all datasets evaluated.

****

Multipath Sparse Coding Using Hierarchical Matching Pursuit

Liefeng Bo, Xiaofeng Ren, Dieter Fox; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 660-667

Complex real-world signals, such as images, contain discriminative structures that differ in many aspects including scale, invariance, and data channel. While progress in deep learning shows the importance of learning features through multiple layers, it is equally important to learn features through multiple paths. We propose Multipath Hierarchical Matching Pursuit (M-HMP), a novel feature learning architecture that combines a collection of hierarchical sparse features for image classification to capture multiple aspects of discriminative structures. Our building blocks are MI-KSVD, a codebook learning algorithm that balances the reconstruction error and the mutual incoherence of the codebook, and batch orthogonal matching pursuit (OMP); we apply them recursively at varying layers and scales. The result is a highly discriminative image representation that leads to large improvements to the state-of-the-art on many standard benchmarks, e.g., Caltech-101, Caltech-256, MITScenes, Oxford-IIIT Pet and Caltech-UCSD Bird-200.

****

Visual Tracking via Locality Sensitive Histograms

Shengfeng He, Qingxiong Yang, Rynson W.H. Lau, Jiang Wang, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2427-2434

This paper presents a novel locality sensitive histogram algorithm for visual tracking. Unlike the conventional image histogram that counts the frequency of occurrences of each intensity value by adding ones to the corresponding bin, a locality sensitive histogram is computed at each pixel location and a floating-point value is added to the corresponding bin for each occurrence of an intensity value. The floating-point value declines exponentially with respect to the distance to the pixel location where the histogram is computed; thus every pixel is considered but those that are far away can be neglected due to the very small weights assigned. An efficient algorithm is proposed that enables the locality sensitive histograms to be computed in time linear in the image size and the number of bins. A robust tracking framework based on the locality sensitive histograms is proposed, which consists of two main components: a new feature for tracking that is robust to illumination changes and a novel multi-region tracking algorithm that runs in realtime even with hundreds of regions. Extensive experiments demonstrate that the proposed tracking framework outperforms the state-of-the-art methods in challenging scenarios, especially when the illumination changes dramatically.

****

Optimized Product Quantization for Approximate Nearest Neighbor Search

Tiezheng Ge, Kaiming He, Qifa Ke, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2946-2953

Product quantization is an effective vector quantization approach to compactly encode high-dimensional vectors for fast approximate nearest neighbor (ANN) search. The essence of product quantization is to decompose the original high-dimensional space into the Cartesian product of a finite number of low-dimensional subspaces that are then quantized separately. Optimal space decomposition is important for the performance of ANN search, but still remains unaddressed. In this paper, we optimize product quantization by minimizing quantization distortions w.r.t. the space decomposition and the quantization codebooks. We present two novel methods for optimization: a nonparametric method that alternatively solves two smaller sub-problems, and a parametric method that is guaranteed to achieve the optimal solution if the input data follows some Gaussian distribution. We show by experiments that our optimized approach substantially improves the accuracy of product quantization for ANN search.

****

Tracking People and Their Objects

Tobias Baumgartner, Dennis Mitzel, Bastian Leibe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3658-3665

Current pedestrian tracking approaches ignore important aspects of human behavior. Humans are not moving independently, but they closely interact with their environment, which includes not only other persons, but also different scene objects. Typical everyday scenarios include people moving in groups, pushing child strollers, or pulling luggage. In this paper, we propose a probabilistic approach for classifying such person-object interactions, associating objects to persons, and predicting how the interaction will most likely continue. Our approach relies on stereo depth information in order to track all scene objects in 3D, while simultaneously building up their 3D shape models. These models and their relative spatial arrangement are then fed into a probabilistic graphical model which jointly infers pairwise interactions and object classes. The inferred interactions can then be used to support tracking by recovering lost object tracks. We evaluate our approach on a novel dataset containing more than 15,000 frames of personobject interactions in 325 video sequences and demonstrate good performance in challenging real-world scenarios.

****

Multi-target Tracking by Lagrangian Relaxation to Min-cost Network Flow

Asad A. Butt, Robert T. Collins; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1846-1853

We propose a method for global multi-target tracking that can incorporate higher-order track smoothness constraints such as constant velocity. Our problem formulation readily lends itself to path estimation in a trellis graph, but unlike previous methods, each node in our network represents a candidate pair of matching observations between consecutive frames. Extra constraints on binary flow variables in the graph result in a problem that can no longer be solved by min-cost network flow. We therefore propose an iterative solution method that relaxes thes e extra constraints using Lagrangian relaxation, resulting in a series of problems that ARE solvable by min-cost flow, and that progressively improve towards a high-quality solution to our original optimization problem. We present experimental results showing that our method outperforms the standard network-flow formulation as well as other recent algorithms that attempt to incorporate higher-order smoothness constraints.

****

In Defense of 3D-Label Stereo

Carl Olsson, Johannes Ulen, Yuri Boykov; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1730-1737

It is commonly believed that higher order smoothness should be modeled using higher order interactions. For example, 2nd order derivatives for deformable (active) contours are represented by triple cliques. Similarly, the 2nd order regularization methods in stereo predominantly use MRF models with scalar (1D) disparity labels and triple clique interactions. In this paper we advocate a largely overlooked alternative approach to stereo where 2nd order surface smoothness is represented by pairwise interactions with 3D-labels, e.g. tangent planes. This general paradigm has been criticized due to perceived computational complexity of optimization in higher-dimensional label space. Contrary to popular beliefs, we demonstrate that representing 2nd order surface smoothness with 3D labels leads to simpler optimization problems with (nearly) submodular pairwise interactions. Our theoretical and experimental results demonstrate advantages over state-of-the-art methods for 2nd order smoothness stereo. 1

****

Compressible Motion Fields

Giuseppe Ottaviano, Pushmeet Kohli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2251-2258

Traditional video compression methods obtain a compact representation for image frames by computing coarse motion fields defined on patches of pixels called blocks, in order to compensate for the motion in the scene across frames. This piecewise constant approximation makes the motion field efficiently encodable, but it introduces block artifacts in the warped image frame. In this paper, we address the problem of estimating dense motion fields that, while accurately predicting one frame from a given reference frame by warping it with the field, are also compressible. We introduce a representation for motion fields based on wavelet bases, and approximate the compressibility of their coefficients with a piecewise smooth surrogate function that yields an objective function similar to classical optical flow formulations. We then show how to quantize and encode such coefficients with adaptive precision. We demonstrate the effectiveness of our approach by comparing its performance with a state-of-the-art wavelet video encoder. Experimental results on a number of standard flow and video datasets reveal that our method significantly outperforms both block-based and optical-flow-based motion compensation algorithms.

****

Dense Object Reconstruction with Semantic Priors

Sid Yingze Bao, Manmohan Chandraker, Yuanqing Lin, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1264-1271

We present a dense reconstruction approach that overcomes the drawbacks of traditional multiview stereo by incorporating semantic information in the form of learned category-level shape priors and object detection. Given training data comprised of 3D scans and images of objects from various viewpoints, we learn a prior comprised of a mean shape and a set of weighted anchor points. The former captures the commonality of shapes across the category, while the latter encodes similarities between instances in the form of appearance and spatial consistency. We propose robust algorithms to match anchor points across instances that enable learning a mean shape for the category, even with large shape variations across instances. We model the shape of an object instance as a warped version of the category mean, along with instance-specific details. Given multiple images of an unseen instance, we collate information from 2D object detectors to align the str ucture from motion point cloud with the mean shape, which is subsequently warped and refined to approach the actual shape. Extensive experiments demonstrate that our model is general enough to learn semantic priors for different object categories, yet powerful enough to reconstruct individual shapes with large variations. Qualitative and quantitative evaluations show that our framework can produce more accurate reconstructions than alternative state-of-the-art multiview stereo systems.

****

Large-Scale Video Summarization Using Web-Image Priors

Aditya Khosla, Raffay Hamid, Chih-Jen Lin, Neel Sundaresan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2698-2705

Given the enormous growth in user-generated videos, it is becoming increasingly important to be able to navigate them efficiently. As these videos are generally of poor quality, summarization methods designed for well-produced videos do not generalize to them. To address this challenge, we propose to use web-images as a prior to facilitate summarization of user-generated videos. Our main intuition is that people tend to take pictures of objects to capture them in a maximally informative way. Such images could therefore be used as prior information to summarize videos containing a similar set of objects. In this work, we apply our novel insight to develop a summarization algorithm that uses the web-image based prior information in an unsupervised manner. Moreover, to automatically evaluate summarization algorithms on a large scale, we propose a framework that relies on multiple summaries obtained through crowdsourcing. We demonstrate the effectiveness of our evaluation framework by comparing its performance to that of multiple human evaluators. Finally, we present results for our framework tested on hundreds of user-generated videos.

****

Deformable Graph Matching

Feng Zhou, Fernando De la Torre; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2922-2929

Graph matching (GM) is a fundamental problem in computer science, and it has been successfully applied to many problems in computer vision. Although widely used, existing GM algorithms cannot incorporate global consistence among nodes, which is a natural constraint in computer vision problems. This paper proposes deformable graph matching (DGM), an extension of GM for matching graphs subject to global rigid and non-rigid geometric constraints. The key idea of this work is a new factorization of the pair-wise affinity matrix. This factorization decouples the affinity matrix into the local structure of each graph and the pair-wise affinity edges. Besides the ability to incorporate global geometric transformations, this factorization offers three more benefits. First, there is no need to compute the costly (in space and time) pair-wise affinity matrix. Second, it provides a unified view of many GM methods and extends the standard iterative closest point algorithm. Third, it allows to use the path-following optimization algorithm that leads to improved optimization strategies and matching performance. Experimental results on synthetic and real databases illustrate how DGM outperforms state-of-the-art algorithms for GM. The code is available at http://humansensing.cs.cmu.edu/fgm .

****

3D Visual Proxemics: Recognizing Human Interactions in 3D from a Single Image

Ishani Chakraborty, Hui Cheng, Omar Javed; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3406-3413

We present a unified framework for detecting and classifying people interactions in unconstrained user generated images. g Unlike previous approaches that directly map people/face locations in 2D image space into features for classification, we first estimate camera viewpoint and people positions in 3D space and then extract spatial configuration features from explicit 3D people positions. This approach has several advantages. First, it can accurately estimate relative distances and orientations between people in 3D. Second, it encodes spatial arrangements of people into a richer set of shape descriptors than afforded in 2D. Our 3D shape descriptors are invariant to camera pose variations often seen in web images and videos. The proposed approach also estimates camera pose and uses it to capture the intent of the photo. To achieve accurate 3D people layout estimation, we develop an algorithm that robustly fuses semantic constraints about human interpositions into a linear camera model. This enables our model to handle large variations in people size, heights (e.g. age) and poses. An accurate 3D layout also allows us to construct features informed by Proxemics that improves our semantic classification. To characterize the human interaction space, we introduce visual proxemes; a set of prototypical patterns that represent commonly occurring social interactions in events. We train a discriminative classifier that classifies 3D arrangements of people into visual proxemes and quantitatively evaluate the performance on a large, challenging dataset.

****

Dictionary Learning from Ambiguously Labeled Data

Yi-Chen Chen, Vishal M. Patel, Jaishanker K. Pillai, Rama Chellappa, P. J. Phillips; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 353-360

We propose a novel dictionary-based learning method for ambiguously labeled multiclass classification, where each training sample has multiple labels and only one of them is the correct label. The dictionary learning problem is solved using an iterative alternating algorithm. At each iteration of the algorithm, two alternating steps are performed: a confidence update and a dictionary update. The confidence of each sample is defined as the probability distribution on its ambiguous labels. The dictionaries are updated using either soft (EM-based) or hard decision rules. Extensive evaluations on existing datasets demonstrate that the proposed method performs significantly better than state-of-the-art ambiguously labeled learning approaches.

****

Graph-Based Optimization with Tubularity Markov Tree for 3D Vessel Segmentation

Ning Zhu, Albert C.S. Chung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2219-2226

In this paper, we propose a graph-based method for 3D vessel tree structure segmentation based on a new tubularity Markov tree model (TMT), which works as both new energy function and graph construction method. With the help of power-watershed implementation [7], a global optimal segmentation can be obtained with low computational cost. Different with other graph-based vessel segmentation methods, the proposed method does not depend on any skeleton and ROI extraction method. The classical issues of the graph-based methods, such as shrinking bias and sensitivity to seed point location, can be solved with the proposed method thanks to vessel data fidelity obtained with TMT . The proposed method is compared with some classical graph-based image segmentation methods and two up-to-date 3D vessel segmentation methods, and is demonstrated to be more accurate than these methods for 3D vessel tree segmentation. Although the segmentation is done without ROI extraction, the computational cost for the proposed method is low (within 20 seconds for 256*256*144 image).

****

Fast Convolutional Sparse Coding

Hilton Bristow, Anders Eriksson, Simon Lucey; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 391-398

Sparse coding has become an increasingly popular method in learning and vision for a variety of classification, reconstruction and coding tasks. The canonical approach intrinsically assumes independence between observations during learning. For many natural signals however, sparse coding is applied to sub-elements (i.e. patches) of the signal, where such an assumption is invalid. Convolutional sparse coding explicitly models local interactions through the convolution operator, however the resulting optimization problem is considerably more complex than traditional sparse coding. In this paper, we draw upon ideas from signal processing and Augmented Lagrange Methods (ALMs) to produce a fast algorithm with globally optimal subproblems and super-linear convergence.

****

Block and Group Regularized Sparse Modeling for Dictionary Learning

Yu-Tseh Chi, Mohsen Ali, Ajit Rajwade, Jeffrey Ho; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 377-382

This paper proposes a dictionary learning framework that combines the proposed block/group (BGSC) or reconstructed block/group (R-BGSC) sparse coding schemes with the novel Intra-block Coherence Suppression Dictionary Learning (ICS-DL) algorithm. An important and distinguishing feature of the proposed framework is that all dictionary blocks are trained simultaneously with respect to each data group while the intra-block coherence being explicitly minimized as an important objective. We provide both empirical evidence and heuristic support for this feature that can be considered as a direct consequence of incorporating both the group structure for the input data and the block structure for the dictionary in the learning process. The optimization problems for both the dictionary learning and sparse coding can be solved efficiently using block-gradient descent, and the details of the optimization algorithms are presented. We evaluate the proposed methods using well-known datasets, and favorable comparisons with state-of-the-art dictionary learning methods demonstrate the viability and validity of the proposed framework.

****

Compressed Hashing

Yue Lin, Rong Jin, Deng Cai, Shuicheng Yan, Xuelong Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 446-451

Recent studies have shown that hashing methods are effective for high dimensional nearest neighbor search. A common problem shared by many existing hashing methods is that in order to achieve a satisfied performance, a large number of hash tables (i.e., long codewords) are required. To address this challenge, in this paper we propose a novel approach called Compressed Hashing by exploring the techniques of sparse coding and compressed sensing. In particular, we introduce a sparse coding scheme, based on the approximation theory of integral operator, that generate sparse representation for high dimensional vectors. We then project sparse codes into a low dimensional space by effectively exploring the Restricted Isometry Property (RIP), a key property in compressed sensing theory. Both of the theoretical analysis and the empirical studies on two large data sets show that the proposed approach is more effective than the state-of-the-art hashing algorithms.

****

Part Discovery from Partial Correspondence

Subhransu Maji, Gregory Shakhnarovich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 931-938

We study the problem of part discovery when partial correspondence between instances of a category are available. For visual categories that exhibit high diversity in structure such as buildings, our approach can be used to discover parts that are hard to name, but can be easily expressed as a correspondence between pairs of images. Parts naturally emerge from point-wise landmark matches across many instances within a category. We propose a learning framework for automatic discovery of parts in such weakly supervised settings, and show the utility of the rich part library learned in this way for three tasks: object detection, category-specific saliency estimation, and fine-grained image parsing.

****

Alternating Decision Forests

Samuel Schulter, Paul Wohlhart, Christian Leistner, Amir Saffari, Peter M. Roth, Horst Bischof; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 508-515

This paper introduces a novel classification method termed Alternating Decision Forests (ADFs), which formulates the training of Random Forests explicitly as a global loss minimization problem. During training, the losses are minimized via keeping an adaptive weight distribution over the training samples, similar to Boosting methods. In order to keep the method as flexible and general as possible, we adopt the principle of employing gradient descent in function space, which allows to minimize arbitrary losses. Contrary to Boosted Trees, in our method the loss minimization is an inherent part of the tree growing process, thus allowing to keep the benefits of common Random Forests, such as, parallel processing. We derive the new classifier and give a discussion and evaluation on standard machine learning data sets. Furthermore, we show how ADFs can be easily integrated into an object detection application. Compared to both, standard Random Forests and Boosted Trees, ADFs give better performance in our experiments, while yielding more compact models in terms of tree depth.

****

SWIGS: A Swift Guided Sampling Method
Victor Fragoso, Matthew Turk; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2770-2777

We present SWIGS, a Swift and efficient Guided Sampling method for robust model estimation from image feature correspondences. Our method leverages the accuracy of our new confidence measure (MR-Rayleigh), which assigns a correctness-confidence to a putative correspondence in an online fashion. MR-Rayleigh is inspired by Meta-Recognition (MR), an algorithm that aims to predict when a classifier's outcome is correct. We demonstrate that by using a Rayleigh distribution, the prediction accuracy of MR can be improved considerably. Our experiments show that MR-Rayleigh tends to predict better than the often-used Lowe's ratio, Brown's ratio, and the standard MR under a range of imaging conditions. Furthermore, our homography estimation experiment demonstrates that SWIGS performs similarly or better than other guided sampling methods while requiring fewer iterations, leading to fast and accurate model estimates.

****

Recognize Human Activities from Partially Observed Videos
Yu Cao, Daniel Barrett, Andrei Barbu, Siddharth Narayanaswamy, Haonan Yu, Aaron Michaux, Yuewei Lin, Sven Dickinson, Jeffrey Mark Siskind, Song Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2658-2665

Recognizing human activities in partially observed videos is a challenging problem and has many practical applications. When the unobserved subsequence is at the end of the video, the problem is reduced to activity prediction from unfinished activity streaming, which has been studied by many researchers. However, in the general case, an unobserved subsequence may occur at any time by yielding a temporal gap in the video. In this paper, we propose a new method that can recognize human activities from partially observed videos in the general case. Specifically, we formulate the problem into a probabilistic framework: 1) dividing each activity into multiple ordered temporal segments, 2) using spatiotemporal features of the training video samples in each segment as bases and applying sparse coding (SC) to derive the activity likelihood of the test video sample at each segment, and 3) finally combining the likelihood at each segment to achieve a global posterior for the activities. We further extend the proposed method to include more bases that correspond to a mixture of segments with different temporal lengths (MSSC), which can better represent the activities with large intra-class variations. We evaluate the proposed methods (SC and MSSC) on various real videos. We also evaluate the proposed methods on two special cases: 1) activity prediction where the unobserved subsequence is at the end of the video, and 2) human activity recognition on fully observed videos. Experimental results show that the proposed methods outperform existing state-of-the-art comparison methods.

****

A Convex Regularizer for Reducing Color Artifact in Color Image Recovery
Shunsuke Ono, Isao Yamada; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1775-1781

We propose a new convex regularizer, named the local color nuclear norm (LCNN), for color image recovery. The LCNN is designed to promote a property inherent in natural color images - in which their local color distributions often exhibit strong linearity - and is thus expected to reduce color artifact effectively. In addition, the very nature of LCNN allows us to incorporate it into various types of color image recovery formulations, with the associated convex optimization problems solvable using proximal splitting techniques. Applicatinos of LCNN are d emonstrated with illustrative numerical examples.

****

Maximum Cohesive Grid of Superpixels for Fast Object Localization

Liang Li, Wei Feng, Liang Wan, Jiawan Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3174-3181

This paper addresses a challenging problem of regularizing arbitrary superpixels into an optimal grid structure, which may significantly extend current low-level vision algorithms by allowing them to use superpixels (SPs) conveniently as using pixels. For this purpose, we aim at constructing maximum cohesive SP-grid, which is composed of real nodes, i.e. SPs, and dummy nodes that are meaningless in the image with only position-taking function in the grid. For a given formation of image SPs and proper number of dummy nodes, we first dynamically align them into a grid based on the centroid localities of SPs. We then define the SP-grid coherence as the sum of edge weights, with SP locality and appearance encoded, along all direct paths connecting any two nearest neighboring real nodes in the grid. We finally maximize the SP-grid coherence via cascade dynamic programming. Our approach can take the regional objectness as an optional constraint to produce more semantically reliable SP-grids. Experiments on object localization show that our approach outperforms state-of-the-art methods in terms of both detection accuracy and speed. We also find that with the same searching strategy and features, object localization at SP-level is about 100-500 times faster than pixel-level, with usually better detection accuracy.

****

Action Recognition by Hierarchical Sequence Summarization

Yale Song, Louis-Philippe Morency, Randall Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3562-3569

Recent progress has shown that learning from hierarchical feature representations leads to improvements in various computer vision tasks. Motivated by the observation that human activity data contains information at various temporal resolutions, we present a hierarchical sequence summarization approach for action recognition that learns multiple layers of discriminative feature representations at different temporal granularities. We build up a hierarchy dynamically and recursively by alternating sequence learning and sequence summarization. For sequence learning we use CRFs with latent variables to learn hidden spatiotemporal dynamics; for sequence summarization we group observations that have similar semantic meaning in the latent space. For each layer we learn an abstract feature representation through non-linear gate functions. This procedure is repeated to obtain a hierarchical sequence summary representation. We develop an efficient learning method to train our model and show that its complexity grows sublinearly with the size of the hierarchy. Experimental results show the effectiveness of our approach, achieving the best published results on the ArmGesture and Canal9 datasets.

****

An Iterated L1 Algorithm for Non-smooth Non-convex Optimization in Computer Vision

Peter Ochs, Alexey Dosovitskiy, Thomas Brox, Thomas Pock; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1759-1766

Natural image statistics indicate that we should use nonconvex norms for most regularization tasks in image processing and computer vision. Still, they are rarely used in practice due to the challenge to optimize them. Recently, iteratively reweighed 1 minimization has been proposed as a way to tackle a class of non-convex functions by solving a sequence of convex 2 1 problems. Here we extend the problem class to linearly constrained optimization of a Lipschitz continuous function, which is the sum of a convex function and a function being concave and increasing on the non-negative orthant (possibly non-convex and nonconcave on the whole space). This allows to apply the algorithm to many computer vision tasks. We show the effect of non-convex regularizers on image denoising, deconvolution, optical flow, and depth map fusion. Non-convexity is particularly interesting in combination with total generalized variation and learned image priors. Efficie nt optimization is made possible by some important properties that are shown to hold.

****

Ensemble Video Object Cut in Highly Dynamic Scenes

Xiaobo Ren, Tony X. Han, Zhihai He; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1947-1954

We consider video object cut as an ensemble of framelevel background-foreground object classifiers which fuses information across frames and refine their segmentation results in a collaborative and iterative manner. Our approach addresses the challenging issues of modeling of background with dynamic textures and segmentation of foreground objects from cluttered scenes. We construct patch-level bagof-words background models to effectively capture the background motion and texture dynamics. We propose a foreground salience graph (FSG) to characterize the similarity of an image patch to the bag-of-words background models in the temporal domain and to neighboring image patches in the spatial domain. We incorporate this similarity information into a graph-cut energy minimization framework for foreground object segmentation. The background-foreground classification results at neighboring frames are fused together to construct a foreground probability map to update the graph weights. The resulting object shapes at neighboring frames are also used as constraints to guide the energy minimization process during graph cut. Our extensive experimental results and performance comparisons over a diverse set of challenging videos with dynamic scenes, including the new Change Detection Challenge Dataset, demonstrate that the proposed ensemble video object cut method outperforms various state-ofthe-art algorithms.

****

Learning for Structured Prediction Using Approximate Subgradient Descent with Working Sets

Aurelien Lucchi, Yunpeng Li, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1987-1994

We propose a working set based approximate subgradient descent algorithm to minimize the margin-sensitive hinge loss arising from the soft constraints in max-margin learning frameworks, such as the structured SVM. We focus on the setting of general graphical models, such as loopy MRFs and CRFs commonly used in image segmentation, where exact inference is intractable and the most violated constraints can only be approximated, voiding the optimality guarantees of the structured SVM's cutting plane algorithm as well as reducing the robustness of existing subgradient based methods. We show that the proposed method obtains better approximate subgradients through the use of working sets, leading to improved convergence properties and increased reliability. Furthermore, our method allows new constraints to be randomly sampled instead of computed using the more expensive approximate inference techniques such as belief propagation and graph cuts, which can be used to reduce learning time at only a small cost of performance. We demonstrate the strength of our method empirically on the segmentation of a new publicly available electron microscopy dataset as well as the popular MSRC data set and show state-of-the-art results.

****

Exploring Implicit Image Statistics for Visual Representativeness Modeling

Xiaoshuai Sun, Xin-Jing Wang, Hongxun Yao, Lei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 516-523

In this paper, we propose a computational model of visual representativeness by integrating cognitive theories of representativeness heuristics with computer vision and machine learning techniques. Unlike previous models that build their representativeness measure based on the visible data, our model takes the initial inputs as explicit positive reference and extend the measure by exploring the implicit negatives. Given a group of images that contains obvious visual concepts, we create a customized image ontology consisting of both positive and negative instances by mining the most related and confusable neighbors of the positive concept in ontological semantic knowledge bases. The representativeness of a new item is then determined by its likelihoods for both the positive and negative references. To ensure the effectiveness of probability inference as well as the cog nitive plausibility, we discover the potential prototypes and treat them as an intermediate representation of semantic concepts. In the experiment, we evaluate the performance of representativeness models based on both human judgements and user-click logs of commercial image search engine. Experimental results on both ImageNet and image sets of general concepts demonstrate the superior performance of our model against the state-of-the-arts.

****

Reconstructing Gas Flows Using Light-Path Approximation

Yu Ji, Jinwei Ye, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2507-2514

Transparent gas flows are difficult to reconstruct: the refractive index field (RIF) within the gas volume is uneven and rapidly evolving, and correspondence matching under distortions is challenging. We present a novel computational imaging solution by exploiting the light field probe (LFProbe). A LF-probe resembles a view-dependent pattern where each pixel on the pattern maps to a unique ray. By observing the LF-probe through the gas flow, we acquire a dense set of ray-ray correspondences and then reconstruct their light paths. To recover the RIF, we use Fermat's Principle to correlate each light path with the RIF via a Partial Differential Equation (PDE). We then develop an iterative optimization scheme to solve for all light-path PDEs in conjunction. Specifically, we initialize the light paths by fitting Hermite splines to ray-ray correspondences, discretize their PDEs onto voxels, and solve a large, over-determined PDE system for the RIF. The RIF can then be used to refine the light paths. Finally, we alternate the RIF and light-path estimations to improve the reconstruction. Experiments on synthetic and real data show that our approach can reliably reconstruct small to medium scale gas flows. In particular, when the flow is acquired by a small number of cameras, the use of ray-ray correspondences can greatly improve the reconstruction.

****

Learning Multiple Non-linear Sub-spaces Using K-RBMs

Siddhartha Chandra, Shailesh Kumar, C.V. Jawahar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2778-2785

Understanding the nature of data is the key to building good representations. In domains such as natural images, the data comes from very complex distributions which are hard to capture. Feature learning intends to discover or best approximate these underlying distributions and use their knowledge to weed out irrelevant information, preserving most of the relevant information. Feature learning can thus be seen as a form of dimensionality reduction. In this paper, we describe a feature learning scheme for natural images. We hypothesize that image patches do not all come from the same distribution, they lie in multiple nonlinear subspaces. We propose a framework that uses K Restricted Boltzmann Machines (K-RBM S) to learn multiple non-linear subspaces in the raw image space. Projections of the image patches into these subspaces gives us features, which we use to build image representations. Our algorithm solves the coupled problem of finding the right non-linear subspaces in the input space and associating image patches with those subspaces in an iterative EM like algorithm to minimize the overall reconstruction error. Extensive empirical results over several popular image classification datasets show that representations based on our framework outperform the traditional feature representations such as the SIFT based Bag-of-Words (BoW) and convolutional deep belief networks.

****

Articulated and Restricted Motion Subspaces and Their Signatures

Bastien Jacquet, Roland Angst, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1506-1513

Articulated objects represent an important class of objects in our everyday environment. Automatic detection of the type of articulated or otherwise restricted motion and extraction of the corresponding motion parameters are therefore of high value, e.g. in order to augment an otherwise static 3D reconstruction with dynamic semantics, such as rotation axes and allowable translation directions for certain rigid parts or objects. Hence, in this paper, a novel theory to analyse relative transformations between two motion-restricted parts will be presented. The analysis is based on linear subspaces spanned by relative transformations. Moreover, a signature for relative transformations will be introduced which uniquely specifies the type of restricted motion encoded in these relative transformations. This theoretic framework enables the derivation of novel algebraic constraints, such as low-rank constraints for subsequent rotations around two fixed axes for example. Lastly, given the type of restricted motion as predicted by the signature, the paper shows how to extract all the motion parameters with matrix manipulations from linear algebra. Our theory is verified on several real data sets, such as a rotating blackboard or a wheel rolling on the floor amongst others.

****

Simultaneous Active Learning of Classifiers & Attributes via Relative Feedback

Arijit Biswas, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 644-651

Active learning provides useful tools to reduce annotation costs without compromising classifier performance. However it traditionally views the supervisor simply as a labeling machine. Recently a new interactive learning paradigm was introduced that allows the supervisor to additionally convey useful domain knowledge using attributes. The learner first conveys its belief about an actively chosen image e.g. "I think this is a forest, what do you think?". If the learner is wrong, the supervisor provides an explanation e.g. "No, this is too open to be a forest". With access to a pre-trained set of relative attribute predictors, the learner fetches all unlabeled images more open than the query image, and uses them as negative examples of forests to update its classifier. This rich human-machine communication leads to better classification performance. In this work, we propose three improvements over this set-up. First, we incorporate a weighting scheme that instead of making a hard decision reasons about the likelihood of an image being a negative example. Second, we do away with pre-trained attributes and instead learn the attribute models on the fly, alleviating overhead and restrictions of a pre-determined attribute vocabulary. Finally, we propose an active learning framework that accounts for not just the labelbut also the attributes-based feedback while selecting the next query image. We demonstrate significant improvement in classification accuracy on faces and shoes. We also collect and make available the largest relative attributes dataset containing 29 attributes of faces from 60 categories.

****

Monocular Template-Based 3D Reconstruction of Extensible Surfaces with Local Linear Elasticity

Abed Malti, Richard Hartley, Adrien Bartoli, Jae-Hak Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1522-1529

We propose a new approach for template-based extensible surface reconstruction from a single view. We extend the method of isometric surface reconstruction and more recent work on conformal surface reconstruction. Our approach relies on the minimization of a proposed stretching energy formalized with respect to the Poisson ratio parameter of the surface. We derive a patch-based formulation of this stretching energy by assuming local linear elasticity. This formulation unifies geometrical and mechanical constraints in a single energy term. We prevent local scale ambiguities by imposing a set of fixed boundary 3D points. We experimentally prove the sufficiency of this set of boundary points and demonstrate the effectiveness of our approach on different developable and non-developable surfaces with a wide range of extensibility.

****

Multi-view Photometric Stereo with Spatially Varying Isotropic Materials

Zhenglong Zhou, Zhe Wu, Ping Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1482-1489

We present a method to capture both 3D shape and spatially varying reflectance with a multi-view photometric stereo technique that works for general isotropic materials. Our data capture setup is simple, which consists of only a digital cam era and a handheld light source. From a single viewpoint, we use a set of photometric stereo images to identify surface points with the same distance to the camera. We collect this information from multiple viewpoints and combine it with structure-from-motion to obtain a precise reconstruction of the complete 3D shape. The spatially varying isotropic bidirectional reflectance distribution function (BRDF) is captured by simultaneously inferring a set of basis BRDFs and their mixing weights at each surface point. According to our experiments, the captured shapes are accurate to 0.3 millimeters. The captured reflectance has relative root-mean-square error (RMSE) of 9%.

****

A New Model and Simple Algorithms for Multi-label Mumford-Shah Problems

Byung-Woo Hong, Zhaojin Lu, Ganesh Sundaramoorthi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1219-1226

In this work, we address the multi-label Mumford-Shah problem, i.e., the problem of jointly estimating a partitioning of the domain of the image, and functions defined within regions of the partition. We create algorithms that are efficient, robust to undesirable local minima, and are easy-toimplement. Our algorithms are formulated by slightly modifying the underlying statistical model from which the multilabel Mumford-Shah functional is derived. The advantage of this statistical model is that the underlying variables: the labels and the functions are less coupled than in the original formulation, and the labels can be computed from the functions with more global updates. The resulting algorithms can be tuned to the desired level of locality of the solution: from fully global updates to more local updates. We demonstrate our algorithm on two applications: joint multi-label segmentation and denoising, and joint multi-label motion segmentation and flow estimation. We compare to the stateof-the-art in multi-label Mumford-Shah problems and show that we achieve more promising results.

****

Kernel Learning for Extrinsic Classification of Manifold Features

Raviteja Vemulapalli, Jaishanker K. Pillai, Rama Chellappa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1782-1789

In computer vision applications, features often lie on Riemannian manifolds with known geometry. Popular learning algorithms such as discriminant analysis, partial least squares, support vector machines, etc., are not directly applicable to such features due to the non-Euclidean nature of the underlying spaces. Hence, classification is often performed in an extrinsic manner by mapping the manifolds to Euclidean spaces using kernels. However, for kernel based approaches, poor choice of kernel often results in reduced performance. In this paper, we address the issue of kernelselection for the classification of features that lie on Riemannian manifolds using the kernel learning approach. We propose two criteria for jointly learning the kernel and the classifier using a single optimization problem. Specifically, for the SVM classifier, we formulate the problem of learning a good kernel-classifier combination as a convex optimization problem and solve it efficiently following the multiple kernel learning approach. Experimental results on image set-based classification and activity recognition clearly demonstrate the superiority of the proposed approach over existing methods for classification of manifold features.

****

Finding Things: Image Parsing with Regions and Per-Exemplar Detectors

Joseph Tighe, Svetlana Lazebnik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3001-3008

This paper presents a system for image parsing, or labeling each pixel in an image with its semantic category, aimed at achieving broad coverage across hundreds of object categories, many of them sparsely sampled. The system combines region-level features with per-exemplar sliding window detectors. Per-exemplar detectors are better suited for our parsing task than traditional bounding box detectors: they perform well on classes with little training data and high intra-class variation, and they allow object masks to be transferred into the test image for pixel-level segmentation. The proposed system achieves state-of-theart accuracy on three challenging datasets, the largest of which contains 45,676 images and 232 labels.

****

Complex Event Detection via Multi-source Video Attributes

Zhigang Ma, Yi Yang, Zhongwen Xu, Shuicheng Yan, Nicu Sebe, Alexander G. Hauptmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2627-2633

Complex events essentially include human, scenes, objects and actions that can be summarized by visual attributes, so leveraging relevant attributes properly could be helpful for event detection. Many works have exploited attributes at image level for various applications. However, attributes at image level are possibly insufficient for complex event detection in videos due to their limited capability in characterizing the dynamic properties of video data. Hence, we propose to leverage attributes at video level (named as video attributes in this work), i.e., the semantic labels of external videos are used as attributes. Compared to complex event videos, these external videos contain simple contents such as objects, scenes and actions which are the basic elements of complex events. Specifically, building upon a correlation vector which correlates the attributes and the complex event, we incorporate video attributes latently as extra informative cues into the event detector learnt from complex event videos. Extensive experiments on a real-world large-scale dataset validate the efficacy of the proposed approach.

****

Learning Collections of Part Models for Object Recognition

Ian Endres, Kevin J. Shih, Johnston Jiaa, Derek Hoiem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 939-946

We propose a method to learn a diverse collection of discriminative parts from object bounding box annotations. Part detectors can be trained and applied individually, which simplifies learning and extension to new features or categories. We apply the parts to object category detection, pooling part detections within bottom-up proposed regions and using a boosted classifier with proposed sigmoid weak learners for scoring. On PASCAL VOC 2010, we evaluate the part detectors' ability to discriminate and localize annotated keypoints. Our detection system is competitive with the best-existing systems, outperforming other HOG-based detectors on the more deformable categories.

****

FrameBreak: Dramatic Image Extrapolation by Guided Shift-Maps

Yinda Zhang, Jianxiong Xiao, James Hays, Ping Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1171-1178

We significantly extrapolate the field of view of a photograph by learning from a roughly aligned, wide-angle guide image of the same scene category. Our method can extrapolate typical photos into complete panoramas. The extrapolation problem is formulated in the shift-map image synthesis framework. We analyze the self-similarity of the guide image to generate a set of allowable local transformations and apply them to the input image. Our guided shift-map method preserves to the scene layout of the guide image when extrapolating a photograph. While conventional shiftmap methods only support translations, this is not expressive enough to characterize the self-similarity of complex scenes. Therefore we additionally allow image transformations of rotation, scaling and reflection. To handle this increase in complexity, we introduce a hierarchical graph optimization method to choose the optimal transformation at each output pixel. We demonstrate our approach on a variety of indoor, outdoor, natural, and man-made scenes.

****

Bayesian Grammar Learning for Inverse Procedural Modeling

Andelo Martinovic, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 201-208

Within the fields of urban reconstruction and city modeling, shape grammars have emerged as a powerful tool for both synthesizing novel designs and reconstructing buildings. Traditionally, a human expert was required to write grammars for specific building styles, which limited the scope of method applicability. We pre

sent an approach to automatically learn two-dimensional attributed stochastic context-free grammars (2D-ASCFGs) from a set of labeled building facades. To this end, we use Bayesian Model Merging, a technique originally developed in the field of natural language processing, which we extend to the domain of two-dimensional languages. Given a set of labeled positive examples, we induce a grammar which can be sampled to create novel instances of the same building style. In addition, we demonstrate that our learned grammar can be used for parsing existing facade imagery. Experiments conducted on the dataset of Haussmannian buildings in Paris show that our parsing with learned grammars not only outperforms bottom-up classifiers but is also on par with approaches that use a manually designed style grammar.
****

Single Image Calibration of Multi-axial Imaging Systems
Amit Agrawal, Srikumar Ramalingam; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1399-1406

Imaging systems consisting of a camera looking at multiple spherical mirrors (reflection) or multiple refractive spheres (refraction) have been used for wide-angle imaging applications. We describe such setups as multi-axial imaging systems, since a single sphere results in an axial system. Assuming an internally calibrated camera, calibration of such multi-axial systems involves estimating the sphere radii and locations in the camera coordinate system. However, previous calibration approaches require manual intervention or constrained setups. We present a fully automatic approach using a single photo of a 2D calibration grid. The pose of the calibration grid is assumed to be unknown and is also recovered. Our approach can handle unconstrained setups, where the mirrors/refractive balls can be arranged in any fashion, not necessarily on a grid. The axial nature of rays allows us to compute the axis of each sphere separately. We then show that by choosing rays from two or more spheres, the unknown pose of the calibration grid can be obtained linearly and independently of sphere radii and locations. Knowing the pose, we derive analytical solutions for obtaining the sphere radius and location. This leads to an interesting result that 6-DOF pose estimation of a multi-axial camera can be done without the knowledge of full calibration. Simulations and real experiments demonstrate the applicability of our algorithm.
****

3D R Transform on Spatio-temporal Interest Points for Action Recognition
Chunfeng Yuan, Xi Li, Weiming Hu, Haibin Ling, Stephen Maybank; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 724-730

Spatio-temporal interest points serve as an elementary building block in many modern action recognition algorithms, and most of them exploit the local spatio-temporal volume features using a Bag of Visual Words (BOVW) representation. Such representation, however, ignores potentially valuable information about the global spatio-temporal distribution of interest points. In this paper, we propose a new global feature to capture the detailed geometrical distribution of interest points. It is calculated by using the R transform which is defined as an extended 3D discrete Radon transform, followed by applying a two-directional two-dimensional principal component analysis. Such R feature captures the geometrical information of the interest points and keeps invariant to geometry transformation and robust to noise. In addition, we propose a new fusion strategy to combine the R feature with the BOVW representation for further improving recognition accuracy. We utilize a context-aware fusion method to capture both the pairwise similarities and higher-order contextual interactions of the videos. Experimental results on several publicly available datasets demonstrate the effectiveness of the proposed approach for action recognition.
****

First-Person Activity Recognition: What Are They Doing to Me?
Michael S. Ryoo, Larry Matthies; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2730-2737

This paper discusses the problem of recognizing interaction-level human activities from a first-person viewpoint. The goal is to enable an observer (e.g., a rob ot or a wearable camera) to understand 'what activity others are performing to it' from continuous video inputs. These include friendly interactions such as 'a person hugging the observer' as well as hostile interactions like 'punching the observer' or 'throwing objects to the observer', whose videos involve a large amount of camera ego-motion caused by physical interactions. The paper investigates multichannel kernels to integrate global and local motion information, and presents a new activity learning/recognition methodology that explicitly considers temporal structures displayed in first-person activity videos. In our experiments, we not only show classification results with segmented videos, but also confirm that our new approach is able to detect activities from continuous videos reliably.

****

Sparse Subspace Denoising for Image Manifolds

Bo Wang, Zhuowen Tu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 468-475

With the increasing availability of high dimensional data and demand in sophisticated data analysis algorithms, manifold learning becomes a critical technique to perform dimensionality reduction, unraveling the intrinsic data structure. The real-world data however often come with noises and outliers; seldom, all the data live in a single linear subspace. Inspired by the recent advances in sparse subspace learning and diffusion-based approaches, we propose a new manifold denoising algorithm in which data neighborhoods are adaptively inferred via sparse subspace reconstruction; we then derive a new formulation to perform denoising to the original data. Experiments carried out on both toy and real applications demonstrate the effectiveness of our method; it is insensitive to parameter tuning and we show significant improvement over the competing algorithms.

****

Adding Unlabeled Samples to Categories by Learned Attributes

Jonghyun Choi, Mohammad Rastegari, Ali Farhadi, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 875-882

We propose a method to expand the visual coverage of training sets that consist of a small number of labeled examples using learned attributes. Our optimization formulation discovers category specific attributes as well as the images that have high confidence in terms of the attributes. In addition, we propose a method to stably capture example-specific attributes for a small sized training set. Our method adds images to a category from a large unlabeled image pool, and leads to significant improvement in category recognition accuracy evaluated on a large-scale dataset, ImageNet.

****

Auxiliary Cuts for General Classes of Higher Order Functionals

Ismail Ben Ayed, Lena Gorelick, Yuri Boykov; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1304-1311

Several recent studies demonstrated that higher order (non-linear) functionals can yield outstanding performances in the contexts of segmentation, co-segmentation and tracking. In general, higher order functionals result in difficult problems that are not amenable to standard optimizers, and most of the existing works investigated particular forms of such functionals. In this study, we derive general bounds for a broad class of higher order functionals. By introducing auxiliary variables and invoking the Jensen's inequality as well as some convexity arguments, we prove that these bounds are auxiliary functionals for various non-linear terms, which include but are not limited to several affinity measures on the distributions or moments of segment appearance and shape, as well as soft constraints on segment volume. From these general-form bounds, we state various non-linear problems as the optimization of auxiliary functionals by graph cuts. The proposed bound optimizers are derivative-free, and consistently yield very steep functional decreases, thereby converging within a few graph cuts. We report several experiments on color and medical data, along with quantitative comparisons to stateof-the-art methods. The results demonstrate competitive performances of the proposed algorithms in regard to accuracy and convergence speed, and confirm t heir potential in various vision and medical applications.

****

Template-Based Isometric Deformable 3D Reconstruction with Sampling-Based Focal Length Self-Calibration

Adrien Bartoli, Toby Collins; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1514-1521

It has been shown that a surface deforming isometrically can be reconstructed from a single image and a template 3D shape. Methods from the literature solve this problem efficiently. However, they all assume that the camera model is calibrated, which drastically limits their applicability. We propose (i) a general variational framework that applies to (calibrated and uncalibrated) general camera models and (ii) self-calibrating 3D reconstruction algorithms for the weak-perspective and full-perspective camera models. In the former case, our algorithm returns the normal field and camera's scale factor. In the latter case, our algorithm returns the normal field, depth and camera's focal length. Our algorithms are the first to achieve deformable 3D reconstruction including camera self-calibration. They apply to much more general setups than existing methods. Experimental results on simulated and real data show that our algorithms give results with the same level of accuracy as existing methods (which use the true focal length) on perspective images, and correctly find the normal field on affine images for which the existing methods fail.

****

Binary Code Ranking with Weighted Hamming Distance

Lei Zhang, Yongdong Zhang, Jinhu Tang, Ke Lu, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1586-1593

Binary hashing has been widely used for efficient similarity search due to its query and storage efficiency. In most existing binary hashing methods, the high-dimensional data are embedded into Hamming space and the distance or similarity of two points are approximated by the Hamming distance between their binary codes. The Hamming distance calculation is efficient, however, in practice, there are often lots of results sharing the same Hamming distance to a query, which makes this distance measure ambiguous and poses a critical issue for similarity search where ranking is important. In this paper, we propose a weighted Hamming distance ranking algorithm (WhRank) to rank the binary codes of hashing methods. By assigning different bit-level weights to different hash bits, the returned binary codes are ranked at a finer-grained binary code level. We give an algorithm to learn the data-adaptive and query-sensitive weight for each hash bit. Evaluations on two large-scale image data sets demonstrate the efficacy of our weighted Hamming distance for binary code ranking.

****

Video Editing with Temporal, Spatial and Appearance Consistency

Xiaojie Guo, Xiaochun Cao, Xiaowu Chen, Yi Ma; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2283-2290

Given an area of interest in a video sequence, one may want to manipulate or edit the area, e.g. remove occlusions from or replace with an advertisement on it. Such a task involves three main challenges including temporal consistency, spatial pose, and visual realism. The proposed method effectively seeks an optimal solution to simultaneously deal with temporal alignment, pose rectification, as well as precise recovery of the occlusion. To make our method applicable to long video sequences, we propose a batch alignment method for automatically aligning and rectifying a small number of initial frames, and then show how to align the remaining frames incrementally to the aligned base images. From the error residual of the robust alignment process, we automatically construct a trimap of the region for each frame, which is used as the input to alpha matting methods to extract the occluding foreground. Experimental results on both simulated and real data demonstrate the accurate and robust performance of our method.

****

Unsupervised Joint Object Discovery and Segmentation in Internet Images

Michael Rubinstein, Armand Joulin, Johannes Kopf, Ce Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1939-1

We present a new unsupervised algorithm to discover and segment out common objects from large and diverse image collections. In contrast to previous co-segmentation methods, our algorithm performs well even in the presence of significant amounts of noise images (images not containing a common object), as typical for datasets collected from Internet search. The key insight to our algorithm is that common object patterns should be salient within each image, while being sparse with respect to smooth transformations across images. We propose to use dense correspondences between images to capture the sparsity and visual variability of the common object over the entire database, which enables us to ignore noise objects that may be salient within their own images but do not commonly occur in others. We performed extensive numerical evaluation on established co-segmentation datasets, as well as several new datasets generated using Internet search. Our approach is able to effectively segment out the common object for diverse object categories, while naturally identifying images where the common object is not present.

****

Learning SURF Cascade for Fast and Accurate Object Detection

Jianguo Li, Yimin Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3468-3475

This paper presents a novel learning framework for training boosting cascade based object detector from large scale dataset. The framework is derived from the wellknown Viola-Jones (VJ) framework but distinguished by three key differences. First, the proposed framework adopts multi-dimensional SURF features instead of single dimensional Haar features to describe local patches. In this way, the number of used local patches can be reduced from hundreds of thousands to several hundreds. Second, it adopts logistic regression as weak classifier for each local patch instead of decision trees in the VJ framework. Third, we adopt AUC as a single criterion for the convergence test during cascade training rather than the two trade-off criteria (false-positive-rate and hit-rate) in the VJ framework. The benefit is that the false-positive-rate can be adaptive among different cascade stages, and thus yields much faster convergence speed of SURF cascade. Combining these points together, the proposed approach has three good properties. First, the boosting cascade can be trained very efficiently. Experiments show that the proposed approach can train object detectors from billions of negative samples within one hour even on personal computers. Second, the built detector is comparable to the stateof-the-art algorithm not only on the accuracy but also on the processing speed. Third, the built detector is small in model-size due to short cascade stages.

****

Efficient Computation of Shortest Path-Concavity for 3D Meshes

Henrik Zimmer, Marcel Campen, Leif Kobbelt; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2155-2162

In the context of shape segmentation and retrieval object-wide distributions of measures are needed to accurately evaluate and compare local regions of shapes. Lien et al. [16] proposed two point-wise concavity measures in the context of Approximate Convex Decompositions of polygons measuring the distance from a point to the polygon's convex hull: an accurate Shortest Path-Concavity (SPC) measure and a Straight Line-Concavity (SLC) approximation of the same. While both are practicable on 2D shapes, the exponential costs of SPC in 3D makes it inhibitively expensive for a generalization to meshes [14]. In this paper we propose an efficient and straight forward approximation of the Shortest Path-Concavity measure to 3D meshes. Our approximation is based on discretizing the space between mesh and convex hull, thereby reducing the continuous Shortest Path search to an efficiently solvable graph problem. Our approach works outof-the-box on complex mesh topologies and requires no complicated handling of genus. Besides presenting a rigorous evaluation of our method on a variety of input meshes, we also define an SPC-based Shape Descriptor and show its superior retrieval and runtime performance compared with the recently presented results on the Convexity Distribution by Lian et al. [12].

****

Learning Discriminative Illumination and Filters for Raw Material Classification with Optimal Projections of Bidirectional Texture Functions

Chao Liu, Geifei Yang, Jinwei Gu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1430-1437

We present a computational imaging method for raw material classification using features of Bidirectional Texture Functions (BTF). Texture is an intrinsic feature for many materials, such as wood, fabric, and granite. At appropriate scales, even "uniform" materials will also exhibit texture features that can be helpful for recognition, such as paper, metal, and ceramic. To cope with the high-dimensionality of BTFs, in this paper, we proposed to learn discriminative illumination patterns and texture filters, with which we can directly measure optimal projections of BTFs for classification. We also studied the effects of texture rotation and scale variation for material classification. We built an LED-based multispectral dome, with which we have acquired a BTF database of a variety of materials and demonstrated the effectiveness of the proposed approach for material classification.

****

Illumination Estimation Based on Bilayer Sparse Coding

Bing Li, Weihua Xiong, Weiming Hu, Houwen Peng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1423-1429

Computational color constancy is a very important topic in computer vision and has attracted many researchers' attention. Recently, lots of research has shown the effects of using high level visual content cues for improving illumination estimation. However, nearly all the existing methods are essentially combinational strategies in which image's content analysis is only used to guide the combination or selection from a variety of individual illumination estimation methods. In this paper, we propose a novel bilayer sparse coding model for illumination estimation that considers image similarity in terms of both low level color distribution and high level image scene content simultaneously. For the purpose, the image's scene content information is integrated with its color distribution to obtain optimal illumination estimation model. The experimental results on real-world image sets show that our algorithm is superior to some prevailing illumination estimation methods, even better than some combinational methods.

****

Leveraging Structure from Motion to Learn Discriminative Codebooks for Scalable Landmark Classification

Alessandro Bergamo, Sudipta N. Sinha, Lorenzo Torresani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 763-770

In this paper we propose a new technique for learning a discriminative codebook for local feature descriptors, specifically designed for scalable landmark classification. The key contribution lies in exploiting the knowledge of correspondences within sets of feature descriptors during codebook learning. Feature correspondences are obtained using structure from motion (SfM) computation on Internet photo collections which serve as the training data. Our codebook is defined by a random forest that is trained to map corresponding feature descriptors into identical codes. Unlike prior forest-based codebook learning methods, we utilize fine-grained descriptor labels and address the challenge of training a forest with an extremely large number of labels. Our codebook is used with various existing feature encoding schemes and also a variant we propose for importanceweighted aggregation of local features. We evaluate our approach on a public dataset of 25 landmarks and our new dataset of 620 landmarks (614K images). Our approach significantly outperforms the state of the art in landmark classification. Furthermore, our method is memory efficient and scalable.

****

Efficient 2D-to-3D Correspondence Filtering for Scalable 3D Object Recognition

Qiang Hao, Rui Cai, Zhiwei Li, Lei Zhang, Yanwei Pang, Feng Wu, Yong Rui; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 899-906

3D model-based object recognition has been a noticeable research trend in recent years. Common methods find 2D-to-3D correspondences and make recognition decisions by pose estimation, whose efficiency usually suffers from noisy correspondences caused by the increasing number of target objects. To overcome this scalability bottleneck, we propose an efficient 2D-to-3D correspondence filtering approach, which combines a light-weight neighborhoodbased step with a finer-grained pairwise step to remove spurious correspondences based on 2D/3D geometric cues. On a dataset of 300 3D objects, our solution achieves ~10 times speed improvement over the baseline, with a comparable recognition accuracy. A parallel implementation on a quad-core CPU can run at ~3fps for 1280x720 images.

****

Weakly Supervised Learning for Attribute Localization in Outdoor Scenes

Shuo Wang, Jungseock Joo, Yizhou Wang, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3111-3118

In this paper, we propose a weakly supervised method for simultaneously learning scene parts and attributes from a collection of images associated with attributes in text, where the precise localization of the each attribute left unknown. Our method includes three aspects. (i) Compositional scene configuration. We learn the spatial layouts of the scene by Hierarchical Space Tiling (HST) representation, which can generate an excessive number of scene configurations through the hierarchical composition of a relatively small number of parts. (ii) Attribute association. The scene attributes contain nouns and adjectives corresponding to the objects and their appearance descriptions respectively. We assign the nouns to the nodes (parts) in HST using nonmaximum suppression of their correlation, then train an appearance model for each noun+adjective attribute pair. (iii) Joint inference and learning. For an image, we compute the most probable parse tree with the attributes as an instantiation of the HST by dynamic programming. Then update the HST and attribute association based on the inferred parse trees. We evaluate the proposed method by (i) showing the improvement of attribute recognition accuracy; and (ii) comparing the average precision of localizing attributes to the scene parts.

****

Jointly Aligning and Segmenting Multiple Web Photo Streams for the Inference of Collective Photo Storylines

Gunhee Kim, Eric P. Xing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 620-627

With an explosion of popularity of online photo sharing, we can trivially collect a huge number of photo streams for any interesting topics such as scuba diving as an outdoor recreational activity class. Obviously, the retrieved photo streams are neither aligned nor calibrated since they are taken in different temporal, spatial, and personal perspectives. However, at the same time, they are likely to share common storylines that consist of sequences of events and activities frequently recurred within the topic. In this paper, as a first technical step to detect such collective storylines, we propose an approach to jointly aligning and segmenting uncalibrated multiple photo streams. The alignment task discovers the matched images between different photo streams, and the image segmentation task parses each image into multiple meaningful regions to facilitate the image understanding. We close a loop between the two tasks so that solving one task helps enhance the performance of the other in a mutually rewarding way. To this end, we design a scalable message-passing based optimization framework to jointly achieve both tasks for the whole input image set at once. With evaluation on the new Flickr dataset of 15 outdoor activities that consist of 1.5 millions of images of 13 thousands of photo streams, our empirical results show that the proposed algorithms are more successful than other candidate methods for both tasks.

****

Studying Relationships between Human Gaze, Description, and Computer Vision

Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, Tamara L. Berg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 739-746

We posit that user behavior during natural viewing of images contains an abundance of information about the content of images as well as information related to user intent and user defined content importance. In this paper, we conduct experiments to better understand the relationship between images, the eye movements people make while viewing images, and how people construct natural language to describe images. We explore these relationships in the context of two commonly used computer vision datasets. We then further relate human cues with outputs of current visual recognition systems and demonstrate prototype applications for gaze-enabled detection and annotation.

****

SLAM++: Simultaneous Localisation and Mapping at the Level of Objects
Renato F. Salas-Moreno, Richard A. Newcombe, Hauke Strasdat, Paul H.J. Kelly, Andrew J. Davison; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1352-1359

We present the major advantages of a new 'object oriented' 3D SLAM paradigm, which takes full advantage in the loop of prior knowledge that many scenes consist of repeated, domain-specific objects and structures. As a hand-held depth camera browses a cluttered scene, realtime 3D object recognition and tracking provides 6DoF camera-object constraints which feed into an explicit graph of objects, continually refined by efficient pose-graph optimisation. This offers the descriptive and predictive power of SLAM systems which perform dense surface reconstruction, but with a huge representation compression. The object graph enables predictions for accurate ICP-based camera to model tracking at each live frame, and efficient active search for new objects in currently undescribed image regions. We demonstrate real-time incremental SLAM in large, cluttered environments, including loop closure, relocalisation and the detection of moved objects, and of course the generation of an object level scene description with the potential to enable interaction.

****

A Theory of Refractive Photo-Light-Path Triangulation
Visesh Chari, Peter Sturm; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1438-1445

3D reconstruction of transparent refractive objects like a plastic bottle is challenging: they lack appearance related visual cues and merely reflect and refract light from the surrounding environment. Amongst several approaches to reconstruct such objects, the seminal work of Light-Path triangulation [17] is highly popular because of its general applicability and analysis of minimal scenarios. A lightpath is defined as the piece-wise linear path taken by a ray of light as it passes from source, through the object and into the camera. Transparent refractive objects not only affect the geometric configuration of light-paths but also their radiometric properties. In this paper, we describe a method that combines both geometric and radiometric information to do reconstruction. We show two major consequences of the addition of radiometric cues to the light-path setup. Firstly, we extend the case of scenarios in which reconstruction is plausible while reducing the minimal requirements for a unique reconstruction. This happens as a consequence of the fact that radiometric cues add an additional known variable to the already existing system of equations. Secondly, we present a simple algorithm for reconstruction, owing to the nature of the radiometric cue. We present several synthetic experiments to validate our theories, and show high quality reconstructions in challenging scenarios.

****

Learning Structured Low-Rank Representations for Image Classification
Yangmuzi Zhang, Zhuolin Jiang, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 676-683

An approach to learn a structured low-rank representation for image classification is presented. We use a supervised learning method to construct a discriminative and reconstructive dictionary. By introducing an ideal regularization term, we perform low-rank matrix recovery for contaminated training data from all categories simultaneously without losing structural information. A discriminative low-rank representation for images with respect to the constructed dictionary is obtained. With semantic structure information and strong identification capability, this representation is good for classification tasks even using a simple linea r multi-classifier. Experimental results demonstrate the effectiveness of our approach.

****

Detecting and Aligning Faces by Image Retrieval

Xiaohui Shen, Zhe Lin, Jonathan Brandt, Ying Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3460-3467

Detecting faces in uncontrolled environments continues to be a challenge to traditional face detection methods[24] due to the large variation in facial appearances, as well as occlusion and clutter. In order to overcome these challenges, we present a novel and robust exemplarbased face detector that integrates image retrieval and discriminative learning. A large database of faces with bounding rectangles and facial landmark locations is collected, and simple discriminative classifiers are learned from each of them. A voting-based method is then proposed to let these classifiers cast votes on the test image through an efficient image retrieval technique. As a result, faces can be very efficiently detected by selecting the modes from the voting maps, without resorting to exhaustive sliding window-style scanning. Moreover, due to the exemplar-based framework, our approach can detect faces under challenging conditions without explicitly modeling their variations. Evaluation on two public benchmark datasets shows that our new face detection approach is accurate and efficient, and achieves the state-of-the-art performance. We further propose to use image retrieval for face validation (in order to remove false positives) and for face alignment/landmark localization. The same methodology can also be easily generalized to other facerelated tasks, such as attribute recognition, as well as general object detection.

****

Towards Contactless, Low-Cost and Accurate 3D Fingerprint Identification

Ajay Kumar, Cyril Kwong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3438-3443

In order to avail the benefits of higher user convenience, hygiene, and improved accuracy, contactless 3D fingerprint recognition techniques have recently been introduced. One of the key limitations of these emerging 3D fingerprint technologies to replace the conventional 2D fingerprint system is their bulk and high cost, which mainly results from the use of multiple imaging cameras or structured lighting employed in these systems. This paper details the development of a contactless 3D fingerprint identification system that uses only single camera. We develop a new representation of 3D finger surface features using Finger Surface Codes and illustrate its effectiveness in matching 3D fingerprints. Conventional minutiae representation is extended in 3D space to accurately match the recovered 3D minutiae. Multiple 2D fingerprint images (with varying illumination profile) acquired to build 3D fingerprints can themselves be used recover 2D features for further improving 3D fingerprint identification and has been illustrated in this paper. The experimental results are shown on a database of 240 client fingerprints and confirm the advantages of the single camera based 3D fingerprint identification.

****

Augmenting CRFs with Boltzmann Machine Shape Priors for Image Labeling

Andrew Kae, Kihyuk Sohn, Honglak Lee, Erik Learned-Miller; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2019-2026

Conditional random fields (CRFs) provide powerful tools for building models to label image segments. They are particularly well-suited to modeling local interactions among adjacent regions (e.g., superpixels). However, CRFs are limited in dealing with complex, global (long-range) interactions between regions. Complementary to this, restricted Boltzmann machines (RBMs) can be used to model global shapes produced by segmentation models. In this work, we present a new model that uses the combined power of these two network types to build a state-of-the-art labeler. Although the CRF is a good baseline labeler, we show how an RBM can be added to the architecture to provide a global shape bias that complements the local modeling provided by the CRF. We demonstrate its labeling performance for the parts of complex face images from the Labeled Faces in the Wild data set. This hybrid model produces results that are both quantitatively and qualitatively better than the CRF alone. In addition to high-quality labeling results, we demonstrate that the hidden units in the RBM portion of our model can be interpreted as face attributes that have been learned without any attribute-level supervision.

****

It's Not Polite to Point: Describing People with Uncertain Attributes
Amir Sadovnik, Andrew Gallagher, Tsuhan Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3089-3096

Visual attributes are powerful features for many different applications in computer vision such as object detection and scene recognition. Visual attributes present another application that has not been examined as rigorously: verbal communication from a computer to a human. Since many attributes are nameable, the computer is able to communicate these concepts through language. However, this is not a trivial task. Given a set of attributes, selecting a subset to be communicated is task dependent. Moreover, because attribute classifiers are noisy, it is important to find ways to deal with this uncertainty. We address the issue of communication by examining the task of composing an automatic description of a person in a group photo that distinguishes him from the others. We introduce an efficient, principled method for choosing which attributes are included in a short description to maximize the likelihood that a third party will correctly guess to which person the description refers. We compare our algorithm to computer baselines and human describers, and show the strength of our method in creating effective descriptions.

****

Reconstructing Loopy Curvilinear Structures Using Integer Programming
Engin Turetken, Fethallah Benmansour, Bjoern Andres, Hanspeter Pfister, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1822-1829

We propose a novel approach to automated delineation of linear structures that form complex and potentially loopy networks. This is in contrast to earlier approaches that usually assume a tree topology for the networks. At the heart of our method is an Integer Programming formulation that allows us to find the global optimum of an objective function designed to allow cycles but penalize spurious junctions and early terminations. We demonstrate that it outperforms state-of-the-art techniques on a wide range of datasets.

****

Weakly-Supervised Dual Clustering for Image Semantic Segmentation
Yang Liu, Jing Liu, Zechao Li, Jinhui Tang, Hanqing Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2075-2082

In this paper, we propose a novel Weakly-Supervised Dual Clustering (WSDC) approach for image semantic segmentation with image-level labels, i.e., collaboratively performing image segmentation and tag alignment with those regions. The proposed approach is motivated from the observation that superpixels belonging to an object class usually exist across multiple images and hence can be gathered via the idea of clustering. In WSDC, spectral clustering is adopted to cluster the superpixels obtained from a set of over-segmented images. At the same time, a linear transformation between features and labels as a kind of discriminative clustering is learned to select the discriminative features among different classes. The both clustering outputs should be consistent as much as possible. Besides, weakly-supervised constraints from image-level labels are imposed to restrict the labeling of superpixels. Finally, the non-convex and non-smooth objective function are efficiently optimized using an iterative CCCP procedure. Extensive experiments conducted on MSRC and LabelMe datasets demonstrate the encouraging performance of our method in comparison with some state-of-the-arts.

****

Multi-target Tracking by Rank-1 Tensor Approximation
Xinchu Shi, Haibin Ling, Junling Xing, Weiming Hu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2387-2394

In this paper we formulate multi-target tracking (MTT) as a rank-1 tensor approximation problem and propose an 1 norm tensor power iteration solution. In particular, a high order tensor is constructed based on trajectories in the time window, with each tensor element as the affinity of the corresponding trajectory candidate. The local assignment variables are the 1 normalized vectors, which are used to approximate the rank-1 tensor. Our approach provides a flexible and effective formulation where both pairwise and high-order association energies can be used expediently. We also show the close relation between our formulation and the multi-dimensional assignment (MDA) model. To solve the optimization in the rank-1 tensor approximation, we propose an algorithm that iteratively powers the intermediate solution followed by an 1 normalization. Aside from effectively capturing high-order motion information, the proposed solver runs efficiently with proved convergence. The experimental validations are conducted on two challenging datasets and our method demonstrates promising performances on both.

****

Multi-image Blind Deblurring Using a Coupled Adaptive Sparse Prior

Haichao Zhang, David Wipf, Yanning Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1051-1058

This paper presents a robust algorithm for estimating a single latent sharp image given multiple blurry and/or noisy observations. The underlying multi-image blind deconvolution problem is solved by linking all of the observations together via a Bayesian-inspired penalty function which couples the unknown latent image, blur kernels, and noise levels together in a unique way. This coupled penalty function enjoys a number of desirable properties, including a mechanism whereby the relative-concavity or shape is adapted as a function of the intrinsic quality of each blurry observation. In this way, higher quality observations may automatically contribute more to the final estimate than heavily degraded ones. The resulting algorithm, which requires no essential tuning parameters, can recover a high quality image from a set of observations containing potentially both blurry and noisy examples, without knowing a priori the degradation type of each observation. Experimental results on both synthetic and real-world test images clearly demonstrate the efficacy of the proposed method.

****

Templateless Quasi-rigid Shape Modeling with Implicit Loop-Closure

Ming Zeng, Jiaxiang Zheng, Xuan Cheng, Xinguo Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 145-152

This paper presents a method for quasi-rigid objects modeling from a sequence of depth scans captured at different time instances. As quasi-rigid objects, such as human bodies, usually have shape motions during the capture procedure, it is difficult to reconstruct their geometries. We represent the shape motion by a deformation graph, and propose a model-to-part method to gradually integrate sampled points of depth scans into the deformation graph. Under an as-rigid-as-possible assumption, the model-to-part method can adjust the deformation graph non-rigidly, so as to avoid error accumulation in alignment, which also implicitly achieves loop-closure. To handle the drift and topological error for the deformation graph, two algorithms are introduced. First, we use a two-stage registration to largely keep the rigid motion part. Second, in the step of graph integration, we topology-adaptively integrate new parts and dynamically control the regularization effect of the deformation graph. We demonstrate the effectiveness and robustness of our method by several depth sequences of quasi-rigid objects, and an application in human shape modeling.

****

Cross-View Action Recognition via a Continuous Virtual Path

Zhong Zhang, Chunheng Wang, Baihua Xiao, Wen Zhou, Shuang Liu, Cunzhao Shi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2690-2697

In this paper, we propose a novel method for cross-view action recognition via a continuous virtual path which connects the source view and the target view. Each point on this virtual path is a virtual view which is obtained by a linear transformation of the action descriptor. All the virtual views are concatenated int o an infinite-dimensional feature to characterize continuous changes from the source to the target view. However, these infinite-dimensional features cannot be used directly. Thus, we propose a virtual view kernel to compute the value of similarity between two infinite-dimensional features, which can be readily used to construct any kernelized classifiers. In addition, there are a lot of unlabeled samples from the target view, which can be utilized to improve the performance of classifiers. Thus, we present a constraint strategy to explore the information contained in the unlabeled samples. The rationality behind the constraint is that any action video belongs to only one class. Our method is verified on the IXMAS dataset, and the experimental results demonstrate that our method achieves better performance than the state-of-the-art methods.

****

Non-rigid Structure from Motion with Diffusion Maps Prior
Lili Tao, Bogdan J. Matuszewski; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1530-1537

In this paper, a novel approach based on a non-linear manifold learning technique is proposed to recover 3D nonrigid structures from 2D image sequences captured by a single camera. Most of the existing approaches assume that 3D shapes can be accurately modelled in a linear subspace. These techniques perform well when the deformations are relatively small or simple, but fail when more complex deformations need to be recovered. The non-linear deformations are often observed in highly flexible objects for which the use of the linear model is impractical. A specific type of shape variations might be governed by only a small number of parameters, therefore can be wellrepresented in a low dimensional manifold. We learn a nonlinear shape prior using diffusion maps method. The key

****

Discriminative Non-blind Deblurring
Uwe Schmidt, Carsten Rother, Sebastian Nowozin, Jeremy Jancsary, Stefan Roth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 604-611

Non-blind deblurring is an integral component of blind approaches for removing image blur due to camera shake. Even though learning-based deblurring methods exist, they have been limited to the generative case and are computationally expensive. To this date, manually-defined models are thus most widely used, though limiting the attained restoration quality. We address this gap by proposing a discriminative approach for non-blind deblurring. One key challenge is that the blur kernel in use at test time is not known in advance. To address this, we analyze existing approaches that use half-quadratic regularization. From this analysis, we derive a discriminative model cascade for image deblurring. Our cascade model consists of a Gaussian CRF at each stage, based on the recently introduced regression tree fields. We train our model by loss minimization and use synthetically generated blur kernels to generate training data. Our experiments show that the proposed approach is efficient and yields state-of-the-art restoration quality on images corrupted with synthetic and real blur.

****

Prostate Segmentation in CT Images via Spatial-Constrained Transductive Lasso
Yinghuan Shi, Shu Liao, Yaozong Gao, Daoqiang Zhang, Yang Gao, Dinggang Shen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2227-2234

Accurate prostate segmentation in CT images is a significant yet challenging task for image guided radiotherapy. In this paper, a novel semi-automated prostate segmentation method is presented. Specifically, to segment the prostate in the current treatment image, the physician first takes a few seconds to manually specify the first and last slices of the prostate in the image space. Then, the prostate is segmented automatically by the proposed two steps: (i) The first step of prostate-likelihood estimation to predict the prostate likelihood for each voxel in the current treatment image, aiming to generate the 3-D prostate-likelihood map by the proposed Spatial-COnstrained Transductive LassO (SCOTO); (ii) The second step of multi-atlases based label fusion to generate the final segmentation result by using the prostate shape information obtained from the planning and p revious treatment images. The experimental result shows that the proposed method outperforms several state-of-the-art methods on prostate segmentation in a real prostate CT dataset, consisting of 24 patients with 330 images. Moreover, it is also clinically feasible since our method just requires the physician to spend a few seconds on manual specification of the first and last slices of the prostate.

****

Optimized Pedestrian Detection for Multiple and Occluded People

Sitapa Rujikietgumjorn, Robert T. Collins; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3690-3697

We present a quadratic unconstrained binary optimization (QUBO) framework for reasoning about multiple object detections with spatial overlaps. The method maximizes an objective function composed of unary detection confidence scores and pairwise overlap constraints to determine which overlapping detections should be suppressed, and which should be kept. The framework is flexible enough to handle the problem of detecting objects as a shape covering of a foreground mask, and to handle the problem of filtering confidence weighted detections produced by a traditional sliding window object detector. In our experiments, we show that our method outperforms two existing state-ofthe-art pedestrian detectors.

****

Rotation, Scaling and Deformation Invariant Scattering for Texture Discrimination

Laurent Sifre, Stephane Mallat; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1233-1240

An affine invariant representation is constructed with a cascade of invariants, which preserves information for classification. A joint translation and rotation invariant representation of image patches is calculated with a scattering transform. It is implemented with a deep convolution network, which computes successive wavelet transforms and modulus non-linearities. Invariants to scaling, shearing and small deformations are calculated with linear operators in the scattering domain. State-of-the-art classification results are obtained over texture databases with uncontrolled viewing conditions.

****

A Minimum Error Vanishing Point Detection Approach for Uncalibrated Monocular Images of Man-Made Environments

Yiliang Xu, Sangmin Oh, Anthony Hoogs; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1376-1383

We present a novel vanishing point detection algorithm for uncalibrated monocular images of man-made environments. We advance the state-of-the-art by a new model of measurement error in the line segment extraction and minimizing its impact on the vanishing point estimation. Our contribution is twofold: 1) Beyond existing hand-crafted models, we formally derive a novel consistency measure, which captures the stochastic nature of the correlation between line segments and vanishing points due to the measurement error, and use this new consistency measure to improve the line segment clustering. 2) We propose a novel minimum error vanishing point estimation approach by optimally weighing the contribution of each line segment pair in the cluster towards the vanishing point estimation. Unlike existing works, our algorithm provides an optimal solution that minimizes the uncertainty of the vanishing point in terms of the trace of its covariance, in a closed-form. We test our algorithm and compare it with the state-of-the-art on two public datasets: York Urban Dataset and Eurasian Cities Dataset. The experiments show that our approach outperforms the state-of-the-art.

****

Poselet Key-Framing: A Model for Human Activity Recognition

Michalis Raptis, Leonid Sigal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2650-2657

In this paper, we develop a new model for recognizing human actions. An action is modeled as a very sparse sequence of temporally local discriminative keyframes collections of partial key-poses of the actor(s), depicting key states in the action sequence. We cast the learning of keyframes in a max-margin discriminative framework, where we treat keyframes as latent variables. This allows us to (jointly) learn a set of most discriminative keyframes while also learning the local temporal context between them. Keyframes are encoded using a spatially-localizable poselet-like representation with HoG and BoW components learned from weak annotations; we rely on structured SVM formulation to align our components and mine for hard negatives to boost localization performance. This results in a model that supports spatio-temporal localization and is insensitive to dropped frames or partial observations. We show classification performance that is competitive with the state of the art on the benchmark UT-Interaction dataset and illustrate that our model outperforms prior methods in an on-line streaming setting.

****

Probabilistic Label Trees for Efficient Large Scale Image Classification

Baoyuan Liu, Fereshteh Sadeghi, Marshall Tappen, Ohad Shamir, Ce Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 843-850

Large-scale recognition problems with thousands of classes pose a particular challenge because applying the classifier requires more computation as the number of classes grows. The label tree model integrates classification with the traversal of the tree so that complexity grows logarithmically. In this paper, we show how the parameters of the label tree can be found using maximum likelihood estimation. This new probabilistic learning technique produces a label tree with significantly improved recognition accuracy.

****

Depth Super Resolution by Rigid Body Self-Similarity in 3D

Michael Hornacek, Christoph Rhemann, Margrit Gelautz, Carsten Rother; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1123-1130

We tackle the problem of jointly increasing the spatial resolution and apparent measurement accuracy of an input low-resolution, noisy, and perhaps heavily quantized depth map. In stark contrast to earlier work, we make no use of ancillary data like a color image at the target resolution, multiple aligned depth maps, or a database of highresolution depth exemplars. Instead, we proceed by identifying and merging patch correspondences within the input depth map itself, exploiting patchwise scene self-similarity across depth such as repetition of geometric primitives or object symmetry. While the notion of 'single-image' super resolution has successfully been applied in the context of color and intensity images, we are to our knowledge the first to present a tailored analogue for depth images. Rather than reason in terms of patches of 2D pixels as others have before us, our key contribution is to proceed by reasoning in terms of patches of 3D points, with matched patch pairs related by a respective 6 DoF rigid body motion in 3D. In support of obtaining a dense correspondence field in reasonable time, we introduce a new 3D variant of PatchMatch. A third contribution is a simple, yet effective patch upscaling and merging technique, which predicts sharp object boundaries at the target resolution. We show that our results are highly competitive with those of alternative techniques leveraging even a color image at the target resolution or a database of high-resolution depth exemplars.

****

SCALPEL: Segmentation Cascades with Localized Priors and Efficient Learning

David Weiss, Ben Taskar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2035-2042

We propose SCALPEL, a flexible method for object segmentation that integrates rich region-merging cues with midand high-level information about object layout, class, and scale into the segmentation process. Unlike competing approaches, SCALPEL uses a cascade of bottom-up segmentation models that is capable of learning to ignore boundaries early on, yet use them as a stopping criterion once the object has been mostly segmented. Furthermore, we show how such cascades can be learned efficiently. When paired with a novel method that generates better localized shape priors than our competitors, our method leads to a concise, accurate set of segmentation proposals; these proposals are more accurate on the PASCAL VOC2010 dataset than state-of-the-art methods that use re-ranking to filter much lar ger bags of proposals. The code for our algorithm is available online.

****

Lost! Leveraging the Crowd for Probabilistic Visual Self-Localization

Marcus A. Brubaker, Andreas Geiger, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3057-3064

In this paper we propose an affordable solution to selflocalization, which utilizes visual odometry and road maps as the only inputs. To this end, we present a probabilistic model as well as an efficient approximate inference algorithm, which is able to utilize distributed computation to meet the real-time requirements of autonomous systems. Because of the probabilistic nature of the model we are able to cope with uncertainty due to noisy visual odometry and inherent ambiguities in the map (e.g., in a Manhattan world). By exploiting freely available, community developed maps and visual odometry measurements, we are able to localize a vehicle up to 3m after only a few seconds of driving on maps which contain more than 2,150km of drivable roads.

****

Class Generative Models Based on Feature Regression for Pose Estimation of Object Categories

Michele Fenzi, Laura Leal-Taixe, Bodo Rosenhahn, Jorn Ostermann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 755-762

In this paper, we propose a method for learning a class representation that can return a continuous value for the pose of an unknown class instance using only 2D data and weak 3D labelling information. Our method is based on generative feature models, i.e., regression functions learnt from local descriptors of the same patch collected under different viewpoints. The individual generative models are then clustered in order to create class generative models which form the class representation. At run-time, the pose of the query image is estimated in a maximum a posteriori fashion by combining the regression functions belonging to the matching clusters. We evaluate our approach on the EPFL car dataset [17] and the Pointing'04 face dataset [8]. Experimental results show that our method outperforms by 10% the state-of-the-art in the first dataset and by 9% in the second.

****

Event Retrieval in Large Video Collections with Circulant Temporal Encoding

Jerome Revaud, Matthijs Douze, Cordelia Schmid, Herve Jegou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2459-2466

This paper presents an approach for large-scale event retrieval. Given a video clip of a specific event, e.g., the wedding of Prince William and Kate Middleton, the goal is to retrieve other videos representing the same event from a dataset of over 100k videos. Our approach encodes the frame descriptors of a video to jointly represent their appearance and temporal order. It exploits the properties of circulant matrices to compare the videos in the frequency domain. This offers a significant gain in complexity and accurately localizes the matching parts of videos. Furthermore, we extend product quantization to complex vectors in order to compress our descriptors, and to compare them in the compressed domain. Our method outperforms the state of the art both in search quality and query time on two large-scale video benchmarks for copy detection, T RECVID and CC WEB . Finally, we introduce a challenging dataset for event retrieval, EVVE, and report the performance on this dataset.

****

Looking Beyond the Image: Unsupervised Learning for Object Saliency and Detection

Parthipan Siva, Chris Russell, Tao Xiang, Lourdes Agapito; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3238-3245

We propose a principled probabilistic formulation of object saliency as a sampling problem. This novel formulation allows us to learn, from a large corpus of unlabelled images, which patches of an image are of the greatest interest and most likely to correspond to an object. We then sample the object saliency map to pr opose object locations. We show that using only a single object location proposal per image, we are able to correctly select an object in over 42% of the images in the P ASCAL VOC 2007 dataset, substantially outperforming existing approaches. Furthermore, we show that our object proposal can be used as a simple unsupervised approach to the weakly supervised annotation problem. Our simple unsupervised approach to annotating objects of interest in images achieves a higher annotation accuracy than most weakly supervised approaches.

****

Selective Transfer Machine for Personalized Facial Action Unit Detection

Wen-Sheng Chu, Fernando De La Torre, Jeffery F. Cohn; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3515-3522

Automatic facial action unit (AFA) detection from video is a long-standing problem in facial expression analysis. Most approaches emphasize choices of features and classifiers. They neglect individual differences in target persons. People vary markedly in facial morphology (e.g., heavy versus delicate brows, smooth versus deeply etched wrinkles) and behavior. Individual differences can dramatically influence how well generic classifiers generalize to previously unseen persons. While a possible solution would be to train person-specific classifiers, that often is neither feasible nor theoretically compelling. The alternative that we propose is to personalize a generic classifier in an unsupervised manner (no additional labels for the test subjects are required). We introduce a transductive learning method, which we refer to Selective Transfer Machine (STM), to personalize a generic classifier by attenuating person-specific biases. STM achieves this effect by simultaneously learning a classifier and re-weighting the training samples that are most relevant to the test subject. To evaluate the effectiveness of STM, we compared STM to generic classifiers and to cross-domain learning methods in three major databases: CK+ [20], GEMEP-FERA [32] and RU-FACS [2]. STM outperformed generic classifiers in all.

****

Procrustean Normal Distribution for Non-rigid Structure from Motion

Minsik Lee, Jungchan Cho, Chong-Ho Choi, Songhwai Oh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1280-1287

Non-rigid structure from motion is a fundamental problem in computer vision, which is yet to be solved satisfactorily. The main difficulty of the problem lies in choosing the right constraints for the solution. In this paper, we propose new constraints that are more effective for non-rigid shape recovery. Unlike the other proposals which have mainly focused on restricting the deformation space using rank constraints, our proposal constrains the motion parameters so that the 3D shapes are most closely aligned to each other, which makes the rank constraints unnecessary. Based on these constraints, we define a new class of probability distribution called the Procrustean normal distribution and propose a new NRSfM algorithm, EM-PND. The experimental results show that the proposed method outperforms the existing methods, and it works well even if there is no temporal dependence between the observed samples.

****

Blur Processing Using Double Discrete Wavelet Transform

Yi Zhang, Keigo Hirakawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1091-1098

We propose a notion of double discrete wavelet transform (DDWT) that is designed to sparsify the blurred image and the blur kernel simultaneously. DDWT greatly enhances our ability to analyze, detect, and process blur kernels and blurry images--the proposed framework handles both global and spatially varying blur kernels seamlessly, and unifies the treatment of blur caused by object motion, optical defocus, and camera shake. To illustrate the potential of DDWT in computer vision and image processing, we develop example applications in blur kernel estimation, deblurring, and near-blur-invariant image feature extraction.

****

Video Enhancement of People Wearing Polarized Glasses: Darkening Reversal and Reflection Reduction

Mao Ye, Cha Zhang, Ruigang Yang; Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition (CVPR), 2013, pp. 1179-1186

With the wide-spread of consumer 3D-TV technology, stereoscopic videoconferencing systems are emerging. However, the special glasses participants wear to see 3D can create distracting images. This paper presents a computational framework to reduce undesirable artifacts in the eye regions caused by these 3D glasses. More specifically, we add polarized filters to the stereo camera so that partial images of reflection can be captured. A novel Bayesian model is then developed to describe the imaging process of the eye regions including darkening and reflection, and infer the eye regions based on Classification ExpectationMaximization (EM). The recovered eye regions under the glasses are brighter and with little reflections, leading to a more nature videoconferencing experience. Qualitative evaluations and user studies are conducted to demonstrate the substantial improvement our approach can achieve.

****

Joint Geodesic Upsampling of Depth Images

Ming-Yu Liu, Oncel Tuzel, Yuichi Taguchi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 169-176

We propose an algorithm utilizing geodesic distances to upsample a low resolution depth image using a registered high resolution color image. Specifically, it computes depth for each pixel in the high resolution image using geodesic paths to the pixels whose depths are known from the low resolution one. Though this is closely related to the all-pairshortest-path problem which has O(n lalog n) complexity, we develop a novel approximation algorithm whose complexity grows linearly with the image size and achieve realtime performance. We compare our algorithm with the state of the art on the benchmark dataset and show that our approach provides more accurate depth upsampling with fewer artifacts. In addition, we show that the proposed algorithm is well suited for upsampling depth images using binary edge maps, an important sensor fusion application.

****

Discriminative Re-ranking of Diverse Segmentations

Payman Yadollahpour, Dhruv Batra, Gregory Shakhnarovich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1923-1930

This paper introduces a two-stage approach to semantic image segmentation. In the first stage a probabilistic model generates a set of diverse plausible segmentations. In the second stage, a discriminatively trained re-ranking model selects the best segmentation from this set. The re-ranking stage can use much more complex features than what could be tractably used in the probabilistic model, allowing a better exploration of the solution space than possible by simply producing the most probable solution from the probabilistic model. While our proposed approach already achieves state-of-the-art results (48.1%) on the challenging VOC 2012 dataset, our machine and human analyses suggest that even larger gains are possible with such an approach.

****

Incorporating User Interaction and Topological Constraints within Contour Completion via Discrete Calculus

Jia Xu, Maxwell D. Collins, Vikas Singh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1886-1893

We study the problem of interactive segmentation and contour completion for multiple objects. The form of constraints our model incorporates are those coming from user scribbles (interior or exterior constraints) as well as information regarding the topology of the 2-D space after partitioning (number of closed contours desired). We discuss how concepts from discrete calculus and a simple identity using the Euler characteristic of a planar graph can be utilized to derive a practical algorithm for this problem. We also present specialized branch and bound methods for the case of single contour completion under such constraints. On an extensive dataset of ~ 1000 images, our experiments suggest that a small amount of side knowledge can give strong improvements over fully unsupervised contour completion methods. We show that by interpreting user indications topologically, user effort is substantially reduced.

****

Shading-Based Shape Refinement of RGB-D Images

Lap-Fai Yu, Sai-Kit Yeung, Yu-Wing Tai, Stephen Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1415-1422

We present a shading-based shape refinement algorithm which uses a noisy, incomplete depth map from Kinect to help resolve ambiguities in shape-from-shading. In our framework, the partial depth information is used to overcome bas-relief ambiguity in normals estimation, as well as to assist in recovering relative albedos, which are needed to reliably estimate the lighting environment and to separate shading from albedo. This refinement of surface normals using a noisy depth map leads to high-quality 3D surfaces. The effectiveness of our algorithm is demonstrated through several challenging real-world examples.

****

Active Contours with Group Similarity

Xiaowei Zhou, Xiaojie Huang, James S. Duncan, Weichuan Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2969-2976

Active contours are widely used in image segmentation. To cope with missing or misleading features in images, researchers have introduced various ways to model the prior of shapes and use the prior to constrain active contours. However, the shape prior is usually learnt from a large set of annotated data, which is not always accessible in practice. Moreover, it is often doubted that the existing shapes in the training set will be sufficient to model the new instance in the testing image. In this paper, we propose to use the group similarity of object shapes in multiple images as a prior to aid segmentation, which can be interpreted as an unsupervised approach of shape prior modeling. We show that the rank of the matrix consisting of multiple shapes is a good measure of the group similarity of the shapes, and the nuclear norm minimization is a simple and effective way to impose the proposed constraint on existing active contour models. Moreover, we develop a fast algorithm to solve the proposed model by using the accelerated proximal method. Experiments using echocardiographic image sequences acquired from acute canine experiments demonstrate that the proposed method can consistently improve the performance of active contour models and increase the robustness against image defects such as missing boundaries.

****

Diffusion Processes for Retrieval Revisited

Michael Donoser, Horst Bischof; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1320-1327

In this paper we revisit diffusion processes on affinity graphs for capturing the intrinsic manifold structure defined by pairwise affinity matrices. Such diffusion processes have already proved the ability to significantly improve subsequent applications like retrieval. We give a thorough overview of the state-of-the-art in this field and discuss obvious similarities and differences. Based on our observations, we are then able to derive a generic framework for diffusion processes in the scope of retrieval applications, where the related work represents specific instances of our generic formulation. We evaluate our framework on several retrieval tasks and are able to derive algorithms that e. g. achieve a 100% bullseye score on the popular MPEG7 shape retrieval data set.

****

From N to N+1: Multiclass Transfer Incremental Learning

Ilja Kuzborskij, Francesco Orabona, Barbara Caputo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3358-3365

Since the seminal work of Thrun [17], the learning to learn paradigm has been defined as the ability of an agent to improve its performance at each task with experience, with the number of tasks. Within the object categorization domain, the visual learning community has actively declined this paradigm in the transfer learning setting. Almost all proposed methods focus on category detection problems, addressing how to learn a new target class from few samples by leveraging over the known source. But if one thinks of learning over multiple tasks, there is a need for multiclass transfer learning algorithms able to exploit previous sour ce knowledge when learning a new class, while at the same time optimizing their overall performance. This is an open challenge for existing transfer learning algorithms. The contribution of this paper is a discriminative method that addresses this issue, based on a Least-Squares Support Vector Machine formulation. Our approach is designed to balance between transferring to the new class and preserving what has already been learned on the source models. Extensive experiments on subsets of publicly available datasets prove the effectiveness of our approach.

****

The SVM-Minus Similarity Score for Video Face Recognition
Lior Wolf, Noga Levy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3523-3530

Face recognition in unconstrained videos requires specialized tools beyond those developed for still images: the fact that the confounding factors change state during the video sequence presents a unique challenge, but also an opportunity to eliminate spurious similarities. Luckily, a major source of confusion in visual similarity of faces is the 3D head orientation, for which image analysis tools provide an accurate estimation. The method we propose belongs to a family of classifierbased similarity scores. We present an effective way to discount pose induced similarities within such a framework, which is based on a newly introduced classifier called SVMminus. The presented method is shown to outperform existing techniques on the most challenging and realistic publicly available video face recognition benchmark, both by itself, and in concert with other methods.

****

Human Pose Estimation Using Body Parts Dependent Joint Regressors
Matthias Dantone, Juergen Gall, Christian Leistner, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3041-3048

In this work, we address the problem of estimating 2d human pose from still images. Recent methods that rely on discriminatively trained deformable parts organized in a tree model have shown to be very successful in solving this task. Within such a pictorial structure framework, we address the problem of obtaining good part templates by proposing novel, non-linear joint regressors. In particular, we employ two-layered random forests as joint regressors. The first layer acts as a discriminative, independent body part classifier. The second layer takes the estimated class distributions of the first one into account and is thereby able to predict joint locations by modeling the interdependence and co-occurrence of the parts. This results in a pose estimation framework that takes dependencies between body parts already for joint localization into account and is thus able to circumvent typical ambiguities of tree structures, such as for legs and arms. In the experiments, we demonstrate that our body parts dependent joint regressors achieve a higher joint localization accuracy than tree-based state-of-the-art methods.

****

A Principled Deep Random Field Model for Image Segmentation
Pushmeet Kohli, Anton Osokin, Stefanie Jegelka; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1971-1978

We discuss a model for image segmentation that is able to overcome the short-boundary bias observed in standard pairwise random field based approaches. To wit, we show that a random field with multi-layered hidden units can encode boundary preserving higher order potentials such as the ones used in the cooperative cuts model of [11] while still allowing for fast and exact MAP inference. Exact inference allows our model to outperform previous image segmentation methods, and to see the true effect of coupling graph edges. Finally, our model can be easily extended to handle segmentation instances with multiple labels, for which it yields promising results.

****

Hash Bit Selection: A Unified Solution for Selection Problems in Hashing
Xianglong Liu, Junfeng He, Bo Lang, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1570-1577

Recent years have witnessed the active development of hashing techniques for nearest neighbor search over big datasets. However, to apply hashing techniques successfully, there are several important issues remaining open in selecting features, hashing algorithms, parameter settings, kernels, etc. In this work, we unify all these selection problems into a hash bit selection framework, i.e. selecting the most informative hash bits from a pool of candidate bits generated by different types of hashing methods using different feature spaces and/or parameter settings, etc. We represent the bit pool as a vertx- and edge-weighted graph with the candidate bits as vertices. The vertex weight represents the bit quality in terms of similarity preservation, and the edge weight reflects independence (non-redundancy) between bits. Then we formulate the bit selection problem as quadratic programming on the graph, and solve it efficiently by replicator dynamics. Moreover, a theoretical study is provided to reveal a very interesting insight: the selected bits actually are the normalized ominant set of the candidate bit graph. We conducted extensive large-scale experiments for three important application scenarios of hash techniques, i.e., hashing with multiple features, multiple hashing algorithms, and multiple bit hashing. We demonstrate that our bit selection approach can achieve superior performance over both naive selection methods and state-of-the-art hashing methods under each scenario, with significant accuracy gains ranging from 10% to 50% relatively.

****

HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences

Omar Oreifej, Zicheng Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 716-723

We present a new descriptor for activity recognition from videos acquired by a depth sensor. Previous descriptors mostly compute shape and motion features independently; thus, they often fail to capture the complex joint shapemotion cues at pixel-level. In contrast, we describe the depth sequence using a histogram capturing the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates. To build the histogram, we create 4D projectors, which quantize the 4D space and represent the possible directions for the 4D normal. We initialize the projectors using the vertices of a regular polychoron. Consequently, we refine the projectors using a discriminative density measure, such that additional projectors are induced in the directions where the 4D normals are more dense and discriminative. Through extensive experiments, we demonstrate that our descriptor better captures the joint shape-motion cues in the depth sequence, and thus outperforms the state-of-the-art on all relevant benchmarks.

****

Principal Observation Ray Calibration for Tiled-Lens-Array Integral Imaging Display

Weiming Li, Haitao Wang, Mingcai Zhou, Shandong Wang, Shaohui Jiao, Xing Mei, Tao Hong, Hoyoung Lee, Jiyeun Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1019-1026

Integral imaging display (IID) is a promising technology to provide realistic 3D image without glasses. To achieve a large screen IID with a reasonable fabrication cost, a potential solution is a tiled-lens-array IID (TLA-IID). However, TLA-IIDs are subject to 3D image artifacts when there are even slight misalignments between the lens arrays. This work aims at compensating these artifacts by calibrating the lens array poses with a camera and including them in a ray model used for rendering the 3D image. Since the lens arrays are transparent, this task is challenging for traditional calibration methods. In this paper, we propose a novel calibration method based on defining a set of principle observation rays that pass lens centers of the TLA and the camera's optical center. The method is able to determine the lens array poses with only one camera at an arbitrary unknown position without using any additional markers. The principle observation rays are automatically extracted using a structured light based method from a dense correspondence map between the displayed and captured pixels. Experiments show that lens array misalignments can be estimated with a standard deviation smaller than 0.4 pixels. Based on this, 3D image artifacts are shown to be effectively r emoved in a test TLA-IID with challenging misalignments.

****

Exploring Weak Stabilization for Motion Feature Extraction

Dennis Park, C. L. Zitnick, Deva Ramanan, Piotr Dollar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2882-2889

We describe novel but simple motion features for the problem of detecting objects in video sequences. Previous approaches either compute optical flow or temporal differences on video frame pairs with various assumptions about stabilization. We describe a combined approach that uses coarse-scale flow and fine-scale temporal difference features. Our approach performs weak motion stabilization by factoring out camera motion and coarse object motion while preserving nonrigid motions that serve as useful cues for recognition. We show results for pedestrian detection and human pose estimation in video sequences, achieving state-of-the-art results in both. In particular, given a fixed detection rate our method achieves a five-fold reduction in false positives over prior art on the Caltech Pedestrian benchmark. Finally, we perform extensive diagnostic experiments to reveal what aspects of our system are crucial for good performance. Proper stabilization, long time-scale features, and proper normalization are all critical.

****

Discovering the Structure of a Planar Mirror System from Multiple Observations of a Single Point

Ilya Reshetouski, Alkhazur Manakov, Ayush Bandhari, Ramesh Raskar, Hans-Peter Seidel, Ivo Ihrke; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 89-96

We investigate the problem of identifying the position of a viewer inside a room of planar mirrors with unknown geometry in conjunction with the room's shape parameters. We consider the observations to consist of angularly resolved depth measurements of a single scene point that is being observed via many multi-bounce interactions with the specular room geometry. Applications of this problem statement include areas such as calibration, acoustic echo cancelation and time-of-flight imaging. We theoretically analyze the problem and derive sufficient conditions for a combination of convex room geometry, observer, and scene point to be reconstructable. The resulting constructive algorithm is exponential in nature and, therefore, not directly applicable to practical scenarios. To counter the situation, we propose theoretically devised geometric constraints that enable an efficient pruning of the solution space and develop a heuristic randomized search algorithm that uses these constraints to obtain an effective solution. We demonstrate the effectiveness of our algorithm on extensive simulations as well as in a challenging real-world calibration scenario.

****

Fine-Grained Crowdsourcing for Fine-Grained Recognition

Jia Deng, Jonathan Krause, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 580-587

Fine-grained recognition concerns categorization at sub-ordinate levels, where the distinction between object classes is highly local. Compared to basic level recognition, fine-grained categorization can be more challenging as there are in general less data and fewer discriminative features. This necessitates the use of stronger prior for feature selection. In this work, we include humans in the loop to help computers select discriminative features. We introduce a novel online game called "Bubbles" that reveals discriminative features humans use. The player's goal is to identify the category of a heavily blurred image. During the game, the player can choose to reveal full details of circular regions ("bubbles"), with a certain penalty. With proper setup the game generates discriminative bubbles with assured quality. We next propose the "BubbleBank" algorithm that uses the human selected bubbles to improve machine recognition performance. Experiments demonstrate that our approach yields large improvements over the previous state of the art on challenging benchmarks.

****

Joint 3D Scene Reconstruction and Class Segmentation

Christian Hane, Christopher Zach, Andrea Cohen, Roland Angst, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 97-104

Both image segmentation and dense 3D modeling from images represent an intrinsically ill-posed problem. Strong regularizers are therefore required to constrain the solutions from being 'too noisy'. Unfortunately, these priors generally yield overly smooth reconstructions and/or segmentations in certain regions whereas they fail in other areas to constrain the solution sufficiently. In this paper we argue that image segmentation and dense 3D reconstruction contribute valuable information to each other's task. As a consequence, we propose a rigorous mathematical framework to formulate and solve a joint segmentation and dense reconstruction problem. Image segmentations provide geometric cues about which surface orientations are more likely to appear at a certain location in space whereas a dense 3D reconstruction yields a suitable regularization for the segmentation problem by lifting the labeling from 2D images to 3D space. We show how appearance-based cues and 3D surface orientation priors can be learned from training data and subsequently used for class-specific regularization. Experimental results on several real data sets highlight the advantages of our joint formulation.

****

Kernel Null Space Methods for Novelty Detection

Paul Bodesheim, Alexander Freytag, Erik Rodner, Michael Kemmler, Joachim Denzler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3374-3381

Detecting samples from previously unknown classes is a crucial task in object recognition, especially when dealing with real-world applications where the closed-world assumption does not hold. We present how to apply a null space method for novelty detection, which maps all training samples of one class to a single point. Beside the possibility of modeling a single class, we are able to treat multiple known classes jointly and to detect novelties for a set of classes with a single model. In contrast to modeling the support of each known class individually, our approach makes use of a projection in a joint subspace where training samples of all known classes have zero intra-class variance. This subspace is called the null space of the training data. To decide about novelty of a test sample, our null space approach allows for solely relying on a distance measure instead of performing density estimation directly. Therefore, we derive a simple yet powerful method for multi-class novelty detection, an important problem not studied sufficiently so far. Our novelty detection approach is assessed in comprehensive multi-class experiments using the publicly available datasets Caltech-256 and ImageNet. The analysis reveals that our null space approach is perfectly suited for multi-class novelty detection since it outperforms all other methods.

****

Information Consensus for Distributed Multi-target Tracking

Ahmed T. Kamal, Jay A. Farrell, Amit K. Roy-Chowdhury; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2403-2410

Due to their high fault-tolerance, ease of installation and scalability to large networks, distributed algorithms have recently gained immense popularity in the sensor networks community, especially in computer vision. Multitarget tracking in a camera network is one of the fundamental problems in this domain. Distributed estimation algorithms work by exchanging information between sensors that are communication neighbors. Since most cameras are directional sensors, it is often the case that neighboring sensors may not be sensing the same target. Such sensors that do not have information about a target are termed as "naive" with respect to that target. In this paper, we propose consensus-based distributed multi-target tracking algorithms in a camera network that are designed to address this issue of naivety. The estimation errors in tracking and data association, as well as the effect of naivety, are jointly addressed leading to the development of an informationweighted consensus algorithm, which we term as the Multitarget Information Consensus (MTIC) algorithm. The incorporation of the probabilistic data association mechanism makes the MTIC algorithm very robust to false measurements/clutter. Experimental analysis is provided to support the theoretical results

.

****

CLAM: Coupled Localization and Mapping with Efficient Outlier Handling

Jonathan Balzer, Stefano Soatto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1554-1561

We describe a method to efficiently generate a model (map) of small-scale objects from video. The map encodes sparse geometry as well as coarse photometry, and could be used to initialize dense reconstruction schemes as well as to support recognition and localization of three-dimensional objects. Self-occlusions and the predominance of outliers present a challenge to existing online Structure From Motion and Simultaneous Localization and Mapping systems. We propose a unified inference criterion that encompasses map building and localization (object detection) relative to the map in a coupled fashion. We establish correspondence in a computationally efficient way without resorting to combinatorial matching or random-sampling techniques. Instead, we use a simpler M-estimator that exploits putative correspondence from tracking after photometric and topological validation. We have collected a new dataset to benchmark model building in the small scale, which we test our algorithm on in comparison to others. Although our system is significantly leaner than previous ones, it compares favorably to the state of the art in terms of accuracy and robustness.

****

Semi-supervised Domain Adaptation with Instance Constraints

Jeff Donahue, Judy Hoffman, Erik Rodner, Kate Saenko, Trevor Darrell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 668-675

Most successful object classification and detection methods rely on classifiers trained on large labeled datasets. However, for domains where labels are limited, simply borrowing labeled data from existing datasets can hurt performance, a phenomenon known as "dataset bias." We propose a general framework for adapting classifiers from "borrowed" data to the target domain using a combination of available labeled and unlabeled examples. Specifically, we show that imposing smoothness constraints on the classifier scores over the unlabeled data can lead to improved adaptation results. Such constraints are often available in the form of instance correspondences, e.g. when the same object or individual is observed simultaneously from multiple views, or tracked between video frames. In these cases, the object labels are unknown but can be constrained to be the same or similar. We propose techniques that build on existing domain adaptation methods by explicitly modeling these relationships, and demonstrate empirically that they improve recognition accuracy in two scenarios, multicategory image classification and object detection in video.

****

Detecting and Naming Actors in Movies Using Generative Appearance Models

Vineet Gandhi, Remi Ronfard; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3706-3713

We introduce a generative model for learning person and costume specific detectors from labeled examples. We demonstrate the model on the task of localizing and naming actors in long video sequences. More specifically, the actor's head and shoulders are each represented as a constellation of optional color regions. Detection can proceed despite changes in view-point and partial occlusions. We explain how to learn the models from a small number of labeled keyframes or video tracks, and how to detect novel appearances of the actors in a maximum likelihood framework. We present results on a challenging movie example, with 81% recall in actor detection (coverage) and 89% precision in actor identification (naming).

****

Rolling Shutter Camera Calibration

Luc Oth, Paul Furgale, Laurent Kneip, Roland Siegwart; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1360-1367

Rolling Shutter (RS) cameras are used across a wide range of consumer electronic devices--from smart-phones to high-end cameras. It is well known, that if a RS camera is used with a moving camera or scene, significant image distortions are introduced. The quality or even success of structure from motion on rolling shutter images requires the usual intrinsic parameters such as focal length and distortion coefficients as well as accurate modelling of the shutter timing. The current state-of-the-art technique for calibrating the shutter timings requires specialised hardware. We present a new method that only requires video of a known calibration pattern. Experimental results on over 60 real datasets show that our method is more accurate than the current state of the art.

****

A Linear Approach to Matching Cuboids in RGBD Images

Hao Jiang, Jianxiong Xiao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2171-2178

We propose a novel linear method to match cuboids in indoor scenes using RGBD images from Kinect. Beyond depth maps, these cuboids reveal important structures of a scene. Instead of directly fitting cuboids to 3D data, we first construct cuboid candidates using superpixel pairs on a RGBD image, and then we optimize the configuration of the cuboids to satisfy the global structure constraints. The optimal configuration has low local matching costs, small object intersection and occlusion, and the cuboids tend to project to a large region in the image; the number of cuboids is optimized simultaneously. We formulate the multiple cuboid matching problem as a mixed integer linear program and solve the optimization efficiently with a branch and bound method. The optimization guarantees the global optimal solution. Our experiments on the Kinect RGBD images of a variety of indoor scenes show that our proposed method is efficient, accurate and robust against object appearance variations, occlusions and strong clutter.

****

Discriminative Segment Annotation in Weakly Labeled Video

Kevin Tang, Rahul Sukthankar, Jay Yagnik, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2483-2490

The ubiquitous availability of Internet video offers the vision community the exciting opportunity to directly learn localized visual concepts from real-world imagery. Unfortunately, most such attempts are doomed because traditional approaches are ill-suited, both in terms of their computational characteristics and their inability to robustly contend with the label noise that plagues uncurated Internet content. We present CRANE, a weakly supervised algorithm that is specifically designed to learn under such conditions. First, we exploit the asymmetric availability of real-world training data, where small numbers of positive videos tagged with the concept are supplemented with large quantities of unreliable negative data. Second, we ensure that CRANE is robust to label noise, both in terms of tagged videos that fail to contain the concept as well as occasional negative videos that do. Finally, CRANE is highly parallelizable, making it practical to deploy at large scale without sacrificing the quality of the learned solution. Although CRANE is general, this paper focuses on segment annotation, where we show state-of-the-art pixel-level segmentation results on two datasets, one of which includes a training set of spatiotemporal segments from more than 20,000 videos.

****

Multi-agent Event Detection: Localization and Role Assignment

Suha Kwak, Bohyung Han, Joon Hee Han; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2682-2689

We present a joint estimation technique of event localization and role assignment when the target video event is described by a scenario. Specifically, to detect multi-agent events from video, our algorithm identifies agents involved in an event and assigns roles to the participating agents. Instead of iterating through all possible agent-role combinations, we formulate the joint optimization problem as two efficient subproblems--quadratic programming for role assignment followed by linear programming for event localization. Additionally, we reduce the computational complexity significantly by applying role-specific event detectors to each agent independently. We test the performance of our algorithm in natural videos, which contain multiple target events and nonparticipating agents.

****

Incorporating Structural Alternatives and Sharing into Hierarchy for Multiclass Object Recognition and Detection

Xiaolong Wang, Liang Lin, Lichao Huang, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3334-3341

This paper proposes a reconfigurable model to recognize and detect multiclass (or multiview) objects with large variation in appearance. Compared with well acknowledged hierarchical models, we study two advanced capabilities in hierarchy for object modeling: (i)"switch" variables(i.e. or-nodes) for specifying alternative compositions, and (ii) making local classifiers (i.e. leaf-nodes) shared among different classes. These capabilities enable us to account well for structural variabilities while preserving the model compact. Our model, in the form of an And-Or Graph, comprises four layers: a batch of leaf-nodes with collaborative edges in bottom for localizing object parts; the or-nodes over bottom to activate their children leaf-nodes; the andnodes to classify objects as a whole; one root-node on the top for switching multiclass classification, which is also an or-node. For model training, we present an EM-type algorithm, namely dynamical structural optimization (DSO), to iteratively determine the structural configuration, (e.g., leaf-node generation associated with their parent or-nodes and shared across other classes), along with optimizing multi-layer parameters. The proposed method is valid on challenging databases, e.g., PASCAL VOC 2007 and UIUCPeople, and it achieves state-of-the-arts performance.

****

Correspondence-Less Non-rigid Registration of Triangular Surface Meshes

Zsolt Santa, Zoltan Kato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2275-2282

A novel correspondence-less approach is proposed to find a thin plate spline map between a pair of deformable 3D objects represented by triangular surface meshes. The proposed method works without landmark extraction and feature correspondences. The aligning transformation is found simply by solving a system of nonlinear equations. Each equation is generated by integrating a nonlinear function over the object's domains. We derive recursive formulas for the efficient computation of these integrals. Based on a series of comparative tests on a large synthetic dataset, our triangular mesh-based algorithm outperforms state of the art methods both in terms of computing time and accuracy. The applicability of the proposed approach has been demonstrated on the registration of 3D lung CT volumes.

****

Globally Consistent Multi-label Assignment on the Ray Space of 4D Light Fields

Sven Wanner, Christoph Straehle, Bastian Goldluecke; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1011-1018

We present the first variational framework for multi-label segmentation on the ray space of 4D light fields. For traditional segmentation of single images, features need to be extracted from the 2D projection of a three-dimensional scene. The associated loss of geometry information can cause severe problems, for example if different objects have a very similar visual appearance. In this work, we show that using a light field instead of an image not only enables to train classifiers which can overcome many of these problems, but also provides an optimal data structure for label optimization by implicitly providing scene geometry information. It is thus possible to consistently optimize label assignment over all views simultaneously. As a further contribution, we make all light fields available online with complete depth and segmentation ground truth data where available, and thus establish the first benchmark data set for light field analysis to facilitate competitive further development of algorithms.

****

Top-Down Segmentation of Non-rigid Visual Objects Using Derivative-Based Search on Sparse Manifolds

Jacinto C. Nascimento, Gustavo Carneiro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1963-1970

The solution for the top-down segmentation of non-rigid visual objects using machine learning techniques is generally regarded as too complex to be solved in its full generality given the large dimensionality of the search space of the expl icit representation of the segmentation contour. In order to reduce this complexity, the problem is usually divided into two stages: rigid detection and non-rigid segmentation. The rationale is based on the fact that the rigid detection can be run in a lower dimensionality space (i.e., less complex and faster) than the original contour space, and its result is then used to constrain the non-rigid segmentation. In this paper, we propose the use of sparse manifolds to reduce the dimensionality of the rigid detection search space of current stateof-the-art top-down segmentation methodologies. The main goals targeted by this smaller dimensionality search space are the decrease of the search running time complexity and the reduction of the training complexity of the rigid detector. These goals are attainable given that both the search and training complexities are function of the dimensionality of the rigid search space. We test our approach in the segmentation of the left ventricle from ultrasound images and lips from frontal face images. Compared to the performance of state-of-the-art non-rigid segmentation system, our experiments show that the use of sparse manifolds for the rigid detection leads to the two goals mentioned above.

****

Harry Potter's Marauder's Map: Localizing and Tracking Multiple Persons-of-Interest by Nonnegative Discretization
Shoou-I Yu, Yi Yang, Alexander Hauptmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3714-3720
A device just like Harry Potter's Marauder's Map, which pinpoints the location of each person-of-interest at all times, provides invaluable information for analysis of surveillance videos. To make this device real, a system would be required to perform robust person localization and tracking in real world surveillance scenarios, especially for complex indoor environments with many walls causing occlusion and long corridors with sparse surveillance camera coverage. We propose a tracking-by-detection approach with nonnegative discretization to tackle this problem. Given a set of person detection outputs, our framework takes advantage of all important cues such as color, person detection, face recognition and non-background information to perform tracking. Local learning approaches are used to uncover the manifold structure in the appearance space with spatio-temporal constraints. Nonnegative discretization is used to enforce the mutual exclusion constraint, which guarantees a person detection output to only belong to exactly one individual. Experiments show that our algorithm performs robust localization and tracking of persons-of-interest not only in outdoor scenes, but also in a complex indoor real-world nursing home environment.

****

Fast Image Super-Resolution Based on In-Place Example Regression
Jianchao Yang, Zhe Lin, Scott Cohen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1059-1066
We propose a fast regression model for practical single image super-resolution based on in-place examples, by leveraging two fundamental super-resolution approaches-learning from an external database and learning from selfexamples. Our in-place self-similarity refines the recently proposed local self-similarity by proving that a patch in the upper scale image have good matches around its origin location in the lower scale image. Based on the in-place examples, a first-order approximation of the nonlinear mapping function from lowto high-resolution image patches is learned. Extensive experiments on benchmark and realworld images demonstrate that our algorithm can produce natural-looking results with sharp edges and preserved fine details, while the current state-of-the-art algorithms are prone to visual artifacts. Furthermore, our model can easily extend to deal with noise by combining the regression results on multiple in-place examples for robust estimation. The algorithm runs fast and is particularly useful for practical applications, where the input images typically contain diverse textures and they are potentially contaminated by noise or compression artifacts.

****

Query Adaptive Similarity for Large Scale Object Retrieval
Danfeng Qin, Christian Wengert, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1610-1617

Many recent object retrieval systems rely on local features for describing an image. The similarity between a pair of images is measured by aggregating the similarity between their corresponding local features. In this paper we present a probabilistic framework for modeling the feature to feature similarity measure. We then derive a query adaptive distance which is appropriate for global similarity evaluation. Furthermore, we propose a function to score the individual contributions into an image to image similarity within the probabilistic framework. Experimental results show that our method improves the retrieval accuracy significantly and consistently. Moreover, our result compares favorably to the state-of-the-art.

****

Winding Number for Region-Boundary Consistent Salient Contour Extraction

Yansheng Ming, Hongdong Li, Xuming He; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2818-2825

This paper aims to extract salient closed contours from an image. For this vision task, both region segmentation cues (e.g. color/texture homogeneity) and boundary detection cues (e.g. local contrast, edge continuity and contour closure) play important and complementary roles. In this paper we show how to combine both cues in a unified framework. The main focus is given to how to maintain the consistency (compatibility) between the region cues and the boundary cues. To this ends, we introduce the use of winding number-a well-known concept in topology-as a powerful mathematical device. By this device, the region-boundary consistency is represented as a set of simple linear relationships. Our method is applied to the figure-ground segmentation problem. The experiments show clearly improved results.

****

Analytic Bilinear Appearance Subspace Construction for Modeling Image Irradiance under Natural Illumination and Non-Lambertian Reflectance

Shireen Y. Elhabian, Aly A. Farag; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1446-1451

Conventional subspace construction approaches suffer from the need of "large-enough" image ensemble rendering numerical methods intractable. In this paper, we propose an analytic formulation for low-dimensional subspace construction in which shading cues lie while preserving the natural structure of an image sample. Using the frequencyspace representation of the image irradiance equation, the process of finding such subspace is cast as establishing a relation between its principal components and that of a deterministic set of basis functions, termed as irradiance harmonics. Representing images as matrices further lessen the number of parameters to be estimated to define a bilinear projection which maps the image sample to a lowerdimensional bilinear subspace. Results show significant impact on dimensionality reduction with minimal loss of information as well as robustness against noise.

****

A Fully-Connected Layered Model of Foreground and Background Flow

Deqing Sun, Jonas Wulff, Erik B. Sudderth, Hanspeter Pfister, Michael J. Black; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2451-2458

Layered models allow scene segmentation and motion estimation to be formulated together and to inform one another. Traditional layered motion methods, however, employ fairly weak models of scene structure, relying on locally connected Ising/Potts models which have limited ability to capture long-range correlations in natural scenes. To address this, we formulate a fully-connected layered model that enables global reasoning about the complicated segmentations of real objects. Optimization with fully-connected graphical models is challenging, and our inference algorithm leverages recent work on efficient mean field updates for fully-connected conditional random fields. These methods can be implemented efficiently using high-dimensional Gaussian filtering. We combine these ideas with a layered flow model, and find that the long-range connections greatly improve segmentation into figure-ground layers when compared with locally connected MRF models. Experiments on several benchmark datasets show that the method can recover fine s

tructures and large occlusion regions, with good flow accuracy and much lower computational cost than previous locally-connected layered models.
****

Bilinear Programming for Human Activity Recognition with Unknown MRF Graphs
Zhenhua Wang, Qinfeng Shi, Chunhua Shen, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1690-1697

Markov Random Fields (MRFs) have been successfully applied to human activity modelling, largely due to their ability to model complex dependencies and deal with local uncertainty. However, the underlying graph structure is often manually specified, or automatically constructed by heuristics. We show, instead, that learning an MRF graph and performing MAP inference can be achieved simultaneously by solving a bilinear program. Equipped with the bilinear program based MAP inference for an unknown graph, we show how to estimate parameters efficiently and effectively with a latent structural SVM. We apply our techniques to predict sport moves (such as serve, volley in tennis) and human activity in TV episodes (such as kiss, hug and Hi-Five). Experimental results show the proposed method outperforms the state-of-the-art.
****

What Object Motion Reveals about Shape with Unknown BRDF and Lighting
Manmohan Chandraker, Dikpal Reddy, Yizhou Wang, Ravi Ramamoorthi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2523-2530

We present a theory that addresses the problem of determining shape from the (small or differential) motion of an object with unknown isotropic reflectance, under arbitrary unknown distant illumination, for both orthographic and perpsective projection. Our theory imposes fundamental limits on the hardness of surface reconstruction, independent of the method involved. Under orthographic projection, we prove that three differential motions suffice to yield an invariant that relates shape to image derivatives, regardless of BRDF and illumination. Under perspective projection, we show that four differential motions suffice to yield depth and a linear constraint on the surface gradient, with unknown BRDF and lighting. Further, we delineate the topological classes up to which reconstruction may be achieved using the invariants. Finally, we derive a general stratification that relates hardness of shape recovery to scene complexity. Qualitatively, our invariants are homogeneous partial differential equations for simple lighting and inhomogeneous for complex illumination. Quantitatively, our framework shows that the minimal number of motions required to resolve shape is greater for more complex scenes. Prior works that assume brightness constancy, Lambertian BRDF or a known directional light source follow as special cases of our stratification. We illustrate with synthetic and real data how potential reconstruction methods may exploit our framework.
****

Learning by Associating Ambiguously Labeled Images
Zinan Zeng, Shijie Xiao, Kui Jia, Tsung-Han Chan, Shenghua Gao, Dong Xu, Yi Ma; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 708-715

We study in this paper the problem of learning classifiers from ambiguously labeled images. For instance, in the collection of new images, each image contains some samples of interest (e.g., human faces), and its associated caption has labels with the true ones included, while the samplelabel association is unknown. The task is to learn classifiers from these ambiguously labeled images and generalize to new images. An essential consideration here is how to make use of the information embedded in the relations between samples and labels, both within each image and across the image set. To this end, we propose a novel framework to address this problem. Our framework is motivated by the observation that samples from the same class repetitively appear in the collection of ambiguously labeled training images, while they are just ambiguously labeled in each image. If we can identify samples of the same class from each image and associate them across the image set, the matrix formed by the samples from the same class would be ideal ly low-rank. By leveraging such a low-rank assumption, we can simultaneously optimize a partial permutation matrix (PPM) for each image, which is formulated in order to exploit all information between samples and labels in a principled way. The obtained PPMs can be readily used to assign labels to samples in training images, and then a standard SVM classifier can be trained and used for unseen data. Experiments on benchmark datasets show the effectiveness of our proposed method.

****

As-Projective-As-Possible Image Stitching with Moving DLT

Julio Zaragoza, Tat-Jun Chin, Michael S. Brown, David Suter; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2339-2346

We investigate projective estimation under model inadequacies, i.e., when the underpinning assumptions of the projective model are not fully satisfied by the data. We focus on the task of image stitching which is customarily solved by estimating a projective warp -a model that is justified when the scene is planar or when the views differ purely by rotation. Such conditions are easily violated in practice, and this yields stitching results with ghosting artefacts that necessitate the usage of deghosting algorithms. To this end we propose as-projective-as-possible warps, i.e., warps that aim to be globally projective, yet allow local non-projective deviations to account for violations to the assumed imaging conditions. Based on a novel estimation technique called Moving Direct Linear Transformation (Moving DLT), our method seamlessly bridges image regions that are inconsistent with the projective model. The result is highly accurate image stitching, with significantly reduced ghosting effects, thus lowering the dependency on post hoc deghosting.

****

Light Field Distortion Feature for Transparent Object Recognition

Kazuki Maeno, Hajime Nagahara, Atsushi Shimada, Rin-Ichiro Taniguchi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2786-2793

Current object-recognition algorithms use local features, such as scale-invariant feature transform (SIFT) and speeded-up robust features (SURF), for visually learning to recognize objects. These approaches though cannot apply to transparent objects made of glass or plastic, as such objects take on the visual features of background objects, and the appearance of such objects dramatically varies with changes in scene background. Indeed, in transmitting light, transparent objects have the unique characteristic of distorting the background by refraction. In this paper, we use a single-shot light Aeld image as an input and model the distortion of the light Aeld caused by the refractive property of a transparent object. We propose a new feature, called the light Aeld distortion (LFD) feature, for identifying a transparent object. The proposal incorporates this LFD feature into the bag-of-features approach for recognizing transparent objects. We evaluated its performance in laboratory and real settings.

****

Ensemble Learning for Confidence Measures in Stereo Vision

Ralf Haeusler, Rahul Nair, Daniel Kondermann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 305-312

With the aim to improve accuracy of stereo confidence measures, we apply the random decision forest framework to a large set of diverse stereo confidence measures. Learning and testing sets were drawn from the recently introduced KITTI dataset, which currently poses higher challenges to stereo solvers than other benchmarks with ground truth for stereo evaluation. We experiment with semi global matching stereo (SGM) and a census dataterm, which is the best performing realtime capable stereo method known to date. On KITTI images, SGM still produces a significant amount of error. We obtain consistently improved area under curve values of sparsification measures in comparison to best performing single stereo confidence measures where numbers of stereo errors are large. More specifically, our method performs best in all but one out of 194 frames of the KITTI dataset.

****

Mirror Surface Reconstruction from a Single Image

Miaomiao Liu, Richard Hartley, Mathieu Salzmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 129-136

This paper tackles the problem of reconstructing the shape of a smooth mirror surface from a single image. In particular, we consider the case where the camera is observing the reflection of a static reference target in the unknown mirror. We first study the reconstruction problem given dense correspondences between 3D points on the reference target and image locations. In such conditions, our differential geometry analysis provides a theoretical proof that the shape of the mirror surface can be uniquely recovered if the pose of the reference target is known. We then relax our assumptions by considering the case where only sparse correspondences are available. In this scenario, we formulate reconstruction as an optimization problem, which can be solved using a nonlinear least-squares method. We demonstrate the effectiveness of our method on both synthetic and real images.

****

Handling Noise in Single Image Deblurring Using Directional Filters

Lin Zhong, Sunghyun Cho, Dimitris Metaxas, Sylvain Paris, Jue Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 612-619

State-of-the-art single image deblurring techniques are sensitive to image noise. Even a small amount of noise, which is inevitable in low-light conditions, can degrade the quality of blur kernel estimation dramatically. The recent approach of Tai and Lin [17] tries to iteratively denoise and deblur a blurry and noisy image. However, as we show in this work, directly applying image denoising methods often partially damages the blur information that is extracted from the input image, leading to biased kernel estimation. We propose a new method for handling noise in blind image deconvolution based on new theoretical and practical insights. Our key observation is that applying a directional low-pass filter to the input image greatly reduces the noise level, while preserving the blur information in the orthogonal direction to the filter. Based on this observation, our method applies a series of directional filters at different orientations to the input image, and estimates an accurate Radon transform of the blur kernel from each filtered image. Finally, we reconstruct the blur kernel using inverse Radon transform. Experimental results on synthetic and real data show that our algorithm achieves higher quality results than previous approaches on blurry and noisy images. 1

****

Joint Spectral Correspondence for Disparate Image Matching

Mayank Bansal, Kostas Daniilidis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2802-2809

We address the problem of matching images with disparate appearance arising from factors like dramatic illumination (day vs. night), age (historic vs. new) and rendering style differences. The lack of local intensity or gradient patterns in these images makes the application of pixellevel descriptors like SIFT infeasible. We propose a novel formulation for detecting and matching persistent features between such images by analyzing the eigen-spectrum of the joint image graph constructed from all the pixels in the two images. We show experimental results of our approach on a public dataset of challenging image pairs and demonstrate significant performance improvements over state-of-the-art.

****

From Local Similarity to Global Coding: An Application to Image Classification

Amirreza Shaban, Hamid R. Rabiee, Mehrdad Farajtabar, Marjan Ghazvininejad; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2794-2801

Bag of words models for feature extraction have demonstrated top-notch performance in image classification. These representations are usually accompanied by a coding method. Recently, methods that code a descriptor giving regard to its nearby bases have proved efficacious. These methods take into account the nonlinear structure of descriptors, since local similarities are a good approximation of g lobal similarities. However, they confine their usage of the global similarities to nearby bases. In this paper, we propose a coding scheme that brings into focus the manifold structure of descriptors, and devise a method to compute the global similarities of descriptors to the bases. Given a local similarity measure between bases, a global measure is computed. Exploiting the local similarity of a descriptor and its nearby bases, a global measure of association of a descriptor to all the bases is computed. Unlike the locality-based and sparse coding methods, the proposed coding varies smoothly with respect to the underlying manifold. Experiments on benchmark image classification datasets substantiate the superiority of the proposed method over its locality and sparsity based rivals.

****

Probabilistic Elastic Matching for Pose Variant Face Verification
Haoxiang Li, Gang Hua, Zhe Lin, Jonathan Brandt, Jianchao Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3499-3506

Pose variation remains to be a major challenge for realworld face recognition. We approach this problem through a probabilistic elastic matching method. We take a part based representation by extracting local features (e.g., LBP or SIFT) from densely sampled multi-scale image patches. By augmenting each feature with its location, a Gaussian mixture model (GMM) is trained to capture the spatialappearance distribution of all face images in the training corpus. Each mixture component of the GMM is confined to be a spherical Gaussian to balance the influence of the appearance and the location terms. Each Gaussian component builds correspondence of a pair of features to be matched between two faces/face tracks. For face verification, we train an SVM on the vector concatenating the difference vectors of all the feature pairs to decide if a pair of faces/face tracks is matched or not. We further propose a joint Bayesian adaptation algorithm to adapt the universally trained GMM to better model the pose variations between the target pair of faces/face tracks, which consistently improves face verification accuracy. Our experiments show that our method outperforms the state-ofthe-art in the most restricted protocol on Labeled Face in the Wild (LFW) and the YouTube video face database by a significant margin.

****

Story-Driven Summarization for Egocentric Video
Zheng Lu, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2714-2721

We present a video summarization approach that discovers the story of an egocentric video. Given a long input video, our method selects a short chain of video subshots depicting the essential events. Inspired by work in text analysis that links news articles over time, we define a randomwalk based metric of influence between subshots that reflects how visual objects contribute to the progression of events. Using this influence metric, we define an objective for the optimal k-subshot summary. Whereas traditional methods optimize a summary's diversity or representativeness, ours explicitly accounts for how one sub-event "leads to" another--which, critically, captures event connectivity beyond simple object co-occurrence. As a result, our summaries provide a better sense of story. We apply our approach to over 12 hours of daily activity video taken from 23 unique camera wearers, and systematically evaluate its quality compared to multiple baselines with 34 human subjects.

****

Towards Pose Robust Face Recognition
Dong Yi, Zhen Lei, Stan Z. Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3539-3545

Most existing pose robust methods are too computational complex to meet practical applications and their performance under unconstrained environments are rarely evaluated. In this paper, we propose a novel method for pose robust face recognition towards practical applications, which is fast, pose robust and can work well under unconstrained environments. Firstly, a 3D deformable model is built and a fast 3D model fitting algorithm is proposed to estimate the pose of face image. Secondly, a group of Gabor filters are transformed according to the pose and shape of face image for feature extraction. Finally, PCA is applied on the pose adaptive Gabor features to remove the redundances and Cosine metric is used to evaluate the similarity. The proposed method has three advantages: (1) The pose correction is applied in the filter space rather than image space, which makes our method less affected by the precision of the 3D model; (2) By combining the holistic pose transformation and local Gabor filtering, the final feature is robust to pose and other negative factors in face recognition; (3) The 3D structure and facial symmetry are successfully used to deal with self-occlusion. Extensive experiments on FERET and PIE show the proposed method outperforms state-ofthe-art methods significantly, meanwhile, the method works well on LFW.

****

All About VLAD

Relja Arandjelovic, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1578-1585

The objective of this paper is large scale object instance retrieval, given a query image. A starting point of such systems is feature detection and description, for example using SIFT. The focus of this paper, however, is towards very large scale retrieval where, due to storage requirements, very compact image descriptors are required and no information about the original SIFT descriptors can be accessed directly at run time. We start from VLAD, the state-of-the art compact descriptor introduced by J??gou et al. [8] for this purpose, and make three novel contributions: first, we show that a simple change to the normalization method significantly improves retrieval performance; second, we show that vocabulary adaptation can substantially alleviate problems caused when images are added to the dataset after initial vocabulary learning. These two methods set a new stateof-the-art over all benchmarks investigated here for both mid-dimensional (20k-D to 30k-D) and small (128-D) descriptors. Our third contribution is a multiple spatial VLAD representation, MultiVLAD, that allows the retrieval and localization of objects that only extend over a small part of an image (again without requiring use of the original image SIFT descriptors).

****

Graph-Based Discriminative Learning for Location Recognition

Song Cao, Noah Snavely; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 700-707

Recognizing the location of a query image by matching it to a database is an important problem in computer vision, and one for which the representation of the database is a key issue. We explore new ways for exploiting the structure of a database by representing it as a graph, and show how the rich information embedded in a graph can improve a bagof-words-based location recognition method. In particular, starting from a graph on a set of images based on visual connectivity, we propose a method for selecting a set of subgraphs and learning a local distance function for each using discriminative techniques. For a query image, each database image is ranked according to these local distance functions in order to place the image in the right part of the graph. In addition, we propose a probabilistic method for increasing the diversity of these ranked database images, again based on the structure of the image graph. We demonstrate that our methods improve performance over standard bag-of-words methods on several existing location recognition datasets.

****

Calibrating Photometric Stereo by Holistic Reflectance Symmetry Analysis

Zhe Wu, Ping Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1498-1505

Under unknown directional lighting, the uncalibrated Lambertian photometric stereo algorithm recovers the shape of a smooth surface up to the generalized bas-relief (GBR) ambiguity. We resolve this ambiguity from the halfvector symmetry, which is observed in many isotropic materials. Under this symmetry, a 2D BRDF slice with low-rank structure can be obtained from an image, if the surface normals and light directions are correctly recovered. In general, this structure is destroyed by the GBR ambiguity. As a result, we can resolve the ambiguity by restoring this structure. We develop a simple algorithm of auto-calibration from separa ble homogeneous specular reflection of real images. Compared with previous methods, this method takes a holistic approach to exploiting reflectance symmetry and produces superior results.

****

Relative Hidden Markov Models for Evaluating Motion Skill

Qiang Zhang, Baoxin Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 548-555

This paper is concerned with a novel problem: learning temporal models using only relative information. Such a problem arises naturally in many applications involving motion or video data. Our focus in this paper is on videobased surgical training, in which a key task is to rate the performance of a trainee based on a video capturing his motion. Compared with the conventional method of relying on ratings from senior surgeons, an automatic approach to this problem is desirable for its potential lower cost, better objectiveness, and real-time availability. To this end, we propose a novel formulation termed Relative Hidden Markov Model and develop an algorithm for obtaining a solution under this model. The proposed method utilizes only a relative ranking (based on an attribute of interest) between pairs of the inputs, which is easier to obtain and often more consistent, especially for the chosen application domain. The proposed algorithm effectively learns a model from the training data so that the attribute under consideration is linked to the likelihood of the inputs under the learned model. Hence the model can be used to compare new sequences. Synthetic data is first used to systematically evaluate the model and the algorithm, and then we experiment with real data from a surgical training system. The experimental results suggest that the proposed approach provides a promising solution to the real-world problem of motion skill evaluation from video.

****

Classification of Tumor Histology via Morphometric Context

Hang Chang, Alexander Borowsky, Paul Spellman, Bahram Parvin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2203-2210

Image-based classification of tissue histology, in terms of different components (e.g., normal signature, categories of aberrant signatures), provides a series of indices for tumor composition. Subsequently, aggregation of these indices in each whole slide image (WSI) from a large cohort can provide predictive models of clinical outcome. However, the performance of the existing techniques is hindered as a result of large technical and biological variations that are always present in a large cohort. In this paper, we propose two algorithms for classification of tissue histology based on robust representations of morphometric context, which are built upon nuclear level morphometric features at various locations and scales within the spatial pyramid matching (SPM) framework. These methods have been evaluated on two distinct datasets of different tumor types collected from The Cancer Genome Atlas (TCGA), and the experimental results indicate that our methods are (i) extensible to different tumor types; (ii) robust in the presence of wide technical and biological variations; (iii) invariant to different nuclear segmentation strategies; and (iv) scalable with varying training sample size. In addition, our experiments suggest that enforcing sparsity, during the construction of morphometric context, further improves the performance of the system.

****

Five Shades of Grey for Fast and Reliable Camera Pose Estimation

Adam Herout, Istvan Szentandrasi, Michal Zacharias, Marketa Dubska, Rudolf Kajan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1384-1390

We introduce here an improved design of the Uniform Marker Fields and an algorithm for their fast and reliable detection. Our concept of the marker field is designed so that it can be detected and recognized for camera pose estimation: in various lighting conditions, under a severe perspective, while heavily occluded, and under a strong motion blur. Our marker field detection harnesses the fact that the edges within the marker field meet at two vanishing points and that the projected planar grid of squares can be defined by a detectable mathematical form alism. The modules of the grid are greyscale and the locations within the marker field are defined by the edges between the modules. The assumption that the marker field is planar allows for a very cheap and reliable camera pose estimation in the captured scene. The detection rates and accuracy are slightly better compared to state-of-the-art marker-based solutions. At the same time, and more importantly, our detector of the marker field is several times faster and the reliable real-time detection can be thus achieved on mobile and low-power devices. We show three targeted applications where the planarity is assured and where the presented marker field design and detection algorithm provide a reliable and extremely fast solution.

****

Fast Patch-Based Denoising Using Approximated Patch Geodesic Paths

Xiaogang Chen, Sing Bing Kang, Jie Yang, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1211-1218

Patch-based methods such as Non-Local Means (NLM) and BM3D have become the de facto gold standard for image denoising. The core of these approaches is to use similar patches within the image as cues for denoising. The operation usually requires expensive pair-wise patch comparisons. In this paper, we present a novel fast patch-based denoising technique based on Patch Geodesic Paths (PatchGP). PatchGPs treat image patches as nodes and patch differences as edge weights for computing the shortest (geodesic) paths. The path lengths can then be used as weights of the smoothing/denoising kernel. We first show that, for natural images, PatchGPs can be effectively approximated by minimum hop paths (MHPs) that generally correspond to Euclidean line paths connecting two patch nodes. To construct the denoising kernel, we further discretize the MHP search directions and use only patches along the search directions. Along each MHP, we apply a weight propagation scheme to robustly and efficiently compute the path distance. To handle noise at multiple scales, we conduct wavelet image decomposition and apply PatchGP scheme at each scale. Comprehensive experiments show that our approach achieves comparable quality as the state-of-the-art methods such as NLM and BM3D but is a few orders of magnitude faster.

****

Boosting Binary Keypoint Descriptors

Tomasz Trzcinski, Mario Christoudias, Pascal Fua, Vincent Lepetit; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2874-2881

Binary keypoint descriptors provide an efficient alternative to their floating-point competitors as they enable faster processing while requiring less memory. In this paper, we propose a novel framework to learn an extremely compact binary descriptor we call BinBoost that is very robust to illumination and viewpoint changes. Each bit of our descriptor is computed with a boosted binary hash function, and we show how to efficiently optimize the different hash functions so that they complement each other, which is key to compactness and robustness. The hash functions rely on weak learners that are applied directly to the image patches, which frees us from any intermediate representation and lets us automatically learn the image gradient pooling configuration of the final descriptor. Our resulting descriptor significantly outperforms the state-of-the-art binary descriptors and performs similarly to the best floating-point descriptors at a fraction of the matching time and memory footprint.

****

Structured Face Hallucination

Chih-Yuan Yang, Sifei Liu, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1099-1106

The goal of face hallucination is to generate highresolution images with fidelity from low-resolution ones. In contrast to existing methods based on patch similarity or holistic constraints in the image space, we propose to exploit local image structures for face hallucination. Each face image is represented in terms of facial components, contours and smooth regions. The image structure is maintained via matching gradients in the reconstructed highresolution output. For facial components, we align input images to generate accurate exemplars and transfer the high-frequency details for preserving structural consistency. For contours, we learn statistical priors to generate salient structures in the high-resolution images. A patch matching method is utilized on the smooth regions where the image gradients are preserved. Experimental results demonstrate that the proposed algorithm generates hallucinated face images with favorable quality and adaptability.

****

Adaptive Active Learning for Image Classification

Xin Li, Yuhong Guo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 859-866

Recently active learning has attracted a lot of attention in computer vision field, as it is time and cost consuming to prepare a good set of labeled images for vision data analysis. Most existing active learning approaches employed in computer vision adopt most uncertainty measures as instance selection criteria. Although most uncertainty query selection strategies are very effective in many circumstances, they fail to take information in the large amount of unlabeled instances into account and are prone to querying outliers. In this paper, we present a novel adaptive active learning approach that combines an information density measure and a most uncertainty measure together to select critical instances to label for image classifications. Our experiments on two essential tasks of computer vision, object recognition and scene recognition, demonstrate the efficacy of the proposed approach.

****

Improving an Object Detector and Extracting Regions Using Superpixels

Guang Shu, Afshin Dehghan, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3721-3727

We propose an approach to improve the detection performance of a generic detector when it is applied to a particular video. The performance of offline-trained objects detectors are usually degraded in unconstrained video environments due to variant illuminations, backgrounds and camera viewpoints. Moreover, most object detectors are trained using Haar-like features or gradient features but ignore video specific features like consistent color patterns. In our approach, we apply a Superpixel-based Bag-of-Words (BoW) model to iteratively refine the output of a generic detector. Compared to other related work, our method builds a video-specific detector using superpixels, hence it can handle the problem of appearance variation. Most importantly, using Conditional Random Field (CRF) along with our super pixel-based BoW model, we develop and algorithm to segment the object from the background . Therefore our method generates an output of the exact object regions instead of the bounding boxes generated by most detectors. In general, our method takes detection bounding boxes of a generic detector as input and generates the detection output with higher average precision and precise object regions. The experiments on four recent datasets demonstrate the effectiveness of our approach and significantly improves the state-of-art detector by 5-16% in average precision.

****

HDR Deghosting: How to Deal with Saturation?

Jun Hu, Orazio Gallo, Kari Pulli, Xiaobai Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1163-1170

We present a novel method for aligning images in an HDR (high-dynamic-range) image stack to produce a new exposure stack where all the images are aligned and appear as if they were taken simultaneously, even in the case of highly dynamic scenes. Our method produces plausible results even where the image used as a reference is either too dark or bright to allow for an accurate registration.

****

Transfer Sparse Coding for Robust Image Representation

Mingsheng Long, Guiguang Ding, Jianmin Wang, Jiaguang Sun, Yuchen Guo, Philip S. Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 407-414

Sparse coding learns a set of basis functions such that each input signal can be well approximated by a linear combination of just a few of the bases. It has at tracted increasing interest due to its state-of-the-art performance in BoW based image representation. However, when labeled and unlabeled images are sampled from different distributions, they may be quantized into different visual words of the codebook and encoded with different representations, which may severely degrade classification performance. In this paper, we propose a Transfer Sparse Coding (TSC) approach to construct robust sparse representations for classifying cross-distribution images accurately. Specifically, we aim to minimize the distribution divergence between the labeled and unlabeled images, and incorporate this criterion into the objective function of sparse coding to make the new representations robust to the distribution difference. Experiments show that TSC can significantly outperform state-ofthe-art methods on three types of computer vision datasets.

****

Semi-supervised Learning of Feature Hierarchies for Object Detection in a Video
Yang Yang, Guang Shu, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1650-1657

We propose a novel approach to boost the performance of generic object detectors on videos by learning videospecific features using a deep neural network. The insight behind our proposed approach is that an object appearing in different frames of a video clip should share similar features, which can be learned to build better detectors. Unlike many supervised detector adaptation or detection-bytracking methods, our method does not require any extra annotations or utilize temporal correspondence. We start with the high-confidence detections from a generic detector, then iteratively learn new video-specific features and refine the detection scores. In order to learn discriminative and compact features, we propose a new feature learning method using a deep neural network based on auto encoders. It differs from the existing unsupervised feature learning methods in two ways: first it optimizes both discriminative and generative properties of the features simultaneously, which gives our features better discriminative ability; second, our learned features are more compact, while the unsupervised feature learning methods usually learn a redundant set of over-complete features. Extensive experimental results on person and horse detection show that significant performance improvement can be achieved with our proposed method.

****

Computationally Efficient Regression on a Dependency Graph for Human Pose Estimation
Kota Hara, Rama Chellappa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3390-3397

We present a hierarchical method for human pose estimation from a single still image. In our approach, a dependency graph representing relationships between reference points such as body joints is constructed and the positions of these reference points are sequentially estimated by a successive application of multidimensional output regressions along the dependency paths, starting from the root node. Each regressor takes image features computed from an image patch centered on the current node's position estimated by the previous regressor and is specialized for estimating its child nodes' positions. The use of the dependency graph allows us to decompose a complex pose estimation problem into a set of local pose estimation problems that are less complex. We design a dependency graph for two commonly used human pose estimation datasets, the Buffy Stickmen dataset and the ETHZ PASCAL Stickmen dataset, and demonstrate that our method achieves comparable accuracy to state-of-the-art results on both datasets with significantly lower computation time than existing methods. Furthermore, we propose an importance weighted boosted regression trees method for transductive learning settings and demonstrate the resulting improved performance for pose estimation tasks.

****

In Defense of Sparsity Based Face Recognition
Weihong Deng, Jiani Hu, Jun Guo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 399-406

The success of sparse representation based classification (SRC) has largely boosted the research of sparsity based face recognition in recent years. A prevailin g view is that the sparsity based face recognition performs well only when the training images have been carefully controlled and the number of samples per class is sufficiently large. This paper challenges the prevailing view by proposing a "prototype plus variation" representation model for sparsity based face recognition. Based on the new model, a Superposed SRC (SSRC), in which the dictionary is assembled by the class centroids and the sample-to-centroid differences, leads to a substantial improvement on SRC. The experiments results on AR, FERET and FRGC databases validate that, if the proposed prototype plus variation representation model is applied, sparse coding plays a crucial role in face recognition, and performs well even when the dictionary bases are collected under uncontrolled conditions and only a single sample per classes is available.

****

Image Matting with Local and Nonlocal Smooth Priors

Xiaowu Chen, Dongqing Zou, Steven Zhiying Zhou, Qinping Zhao, Ping Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1902-1907

In this paper we propose a novel alpha matting method with local and nonlocal smooth priors. We observe that the manifold preserving editing propagation [4] essentially introduced a nonlocal smooth prior on the alpha matte. This nonlocal smooth prior and the well known local smooth prior from matting Laplacian complement each other. So we combine them with a simple data term from color sampling in a graph model for nature image matting. Our method has a closed-form solution and can be solved efficiently. Compared with the state-of-the-art methods, our method produces more accurate results according to the evaluation on standard benchmark datasets.

****

Image Tag Completion via Image-Specific and Tag-Specific Linear Sparse Reconstructions

Zijia Lin, Guiguang Ding, Mingqing Hu, Jianmin Wang, Xiaojun Ye; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1618-1625

Though widely utilized for facilitating image management, user-provided image tags are usually incomplete and insufficient to describe the whole semantic content of corresponding images, resulting in performance degradations in tag-dependent applications and thus necessitating effective tag completion methods. In this paper, we propose a novel scheme denoted as LSR for automatic image tag completion via image-specific and tag-specific Linear Sparse Reconstructions. Given an incomplete initial tagging matrix with each row representing an image and each column representing a tag, LSR optimally reconstructs each image (i.e. row) and each tag (i.e. column) with remaining ones under constraints of sparsity, considering imageimage similarity, image-tag association and tag-tag concurrence. Then both image-specific and tag-specific reconstruction values are normalized and merged for selecting missing related tags. Extensive experiments conducted on both benchmark dataset and web images well demonstrate the effectiveness of the proposed LSR.

****

Non-parametric Filtering for Geometric Detail Extraction and Material Representation

Zicheng Liao, Jason Rock, Yang Wang, David Forsyth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 963-970

Geometric detail is a universal phenomenon in real world objects. It is an important component in object modeling, but not accounted for in current intrinsic image works. In this work, we explore using a non-parametric method to separate geometric detail from intrinsic image components. We further decompose an image as albedo * (coarse-scale shading + shading detail). Our decomposition offers quantitative improvement in albedo recovery and material classification.Our method also enables interesting image editing activities, including bump removal, geometric detail smoothing/enhancement and material transfer.

****

Bottom-Up Segmentation for Top-Down Detection

Sanja Fidler, Roozbeh Mottaghi, Alan Yuille, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3294-3301

In this paper we are interested in how semantic segmentation can help object detection. Towards this goal, we propose a novel deformable part-based model which exploits region-based segmentation algorithms that compute candidate object regions by bottom-up clustering followed by ranking of those regions. Our approach allows every detection hypothesis to select a segment (including void), and scores each box in the image using both the traditional HOG filters as well as a set of novel segmentation features. Thus our model "blends" between the detector and segmentation models. Since our features can be computed very efficiently given the segments, we maintain the same complexity as the original DPM [14]. We demonstrate the effectiveness of our approach in PASCAL VOC 2010, and show that when employing only a root filter our approach outperforms Dalal & Triggs detector [12] on all classes, achieving 13% higher average AP. When employing the parts, we outperform the original DPM [14] in 19 out of 20 classes, achieving an improvement of 8% AP. Furthermore, we outperform the previous state-of-the-art on VOC'10 test by 4%.

****

Expanded Parts Model for Human Attribute and Action Recognition in Still Images

Gaurav Sharma, Frederic Jurie, Cordelia Schmid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 652-659

We propose a new model for recognizing human attributes (e.g. wearing a suit, sitting, short hair) and actions (e.g. running, riding a horse) in still images. The proposed model relies on a collection of part templates which are learnt discriminatively to explain specific scale-space locations in the images (in human centric coordinates). It avoids the limitations of highly structured models, which consist of a few (i.e. a mixture of) 'average' templates. To learn our model, we propose an algorithm which automatically mines out parts and learns corresponding discriminative templates with their respective locations from a large number of candidate parts. We validate the method on recent challenging datasets: (i) Willow 7 actions [7], (ii) 27 Human Attributes (HAT) [25], and (iii) Stanford 40 actions [37]. We obtain convincing qualitative and state-of-the-art quantitative results on the three datasets.

****

Inductive Hashing on Manifolds

Fumin Shen, Chunhua Shen, Qinfeng Shi, Anton van den Hengel, Zhenmin Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1562-1569

Learning based hashing methods have attracted considerable attention due to their ability to greatly increase the scale at which existing algorithms may operate. Most of these methods are designed to generate binary codes that preserve the Euclidean distance in the original space. Manifold learning techniques, in contrast, are better able to model the intrinsic structure embedded in the original highdimensional data. The complexity of these models, and the problems with out-of-sample data, have previously rendered them unsuitable for application to large-scale embedding, however. In this work, we consider how to learn compact binary embeddings on their intrinsic manifolds. In order to address the above-mentioned difficulties, we describe an efficient, inductive solution to the out-of-sample data problem, and a process by which non-parametric manifold learning may be used as the basis of a hashing method. Our proposed approach thus allows the development of a range of new hashing techniques exploiting the flexibility of the wide variety of manifold learning approaches available. We particularly show that hashing on the basis of t-SNE [29] outperforms state-of-the-art hashing methods on large-scale benchmark datasets, and is very effective for image classification with very short code lengths.

****

Robust Feature Matching with Alternate Hough and Inverted Hough Transforms

Hsin-Yi Chen, Yen-Yu Lin, Bing-Yu Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2762-2769

We present an algorithm that carries out alternate Hough transform and inverted Hough transform to establish feature correspondences, and enhances the quality of matching in both precision and recall. Inspired by the fact that nearby features on the same object share coherent homographies in matching, we cast the task of feature matching as a density estimation problem in the Hough space spanned by the hypotheses of homographies. Specifically, we project all the correspondences into the Hough space, and determine the correctness of the correspondences by their respective densities. In this way, mutual verification of relevant correspondences is activated, and the precision of matching is boosted. On the other hand, we infer the concerted homographies propagated from the locally grouped features, and enrich the correspondence candidates for each feature. The recall is hence increased. The two processes are tightly coupled. Through iterative optimization, plausible enrichments are gradually revealed while more correct correspondences are detected. Promising experimental results on three benchmark datasets manifest the effectiveness of the proposed approach.
****
Fast Object Detection with Entropy-Driven Evaluation
Raphael Sznitman, Carlos Becker, Francois Fleuret, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3270-3277
Cascade-style approaches to implementing ensemble classifiers can deliver significant speed-ups at test time. While highly effective, they remain challenging to tune and their overall performance depends on the availability of large validation sets to estimate rejection thresholds. These characteristics are often prohibitive and thus limit their applicability. We introduce an alternative approach to speeding-up classifier evaluation which overcomes these limitations. It involves maintaining a probability estimate of the class label at each intermediary response and stopping when the corresponding uncertainty becomes small enough. As a result, the evaluation terminates early based on the sequence of responses observed. Furthermore, it does so independently of the type of ensemble classifier used or the way it was trained. We show through extensive experimentation that our method provides 2 to 10 fold speed-ups, over existing state-of-the-art methods, at almost no loss in accuracy on a number of object classification tasks.
****
Event Recognition in Videos by Learning from Heterogeneous Web Sources
Lin Chen, Lixin Duan, Dong Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2666-2673
In this work, we propose to leverage a large number of loosely labeled web videos (e.g., from YouTube) and web images (e.g., from Google/Bing image search) for visual event recognition in consumer videos without requiring any labeled consumer videos. We formulate this task as a new multi-domain adaptation problem with heterogeneous sources, in which the samples from different source domains can be represented by different types of features with different dimensions (e.g., the SIFT features from web images and space-time (ST) features from web videos) while the target domain samples have all types of features. To effectively cope with the heterogeneous sources where some source domains are more relevant to the target domain, we propose a new method called Multi-domain Adaptation with Heterogeneous Sources (MDA-HS) to learn an optimal target classifier, in which we simultaneously seek the optimal weights for different source domains with different types of features as well as infer the labels of unlabeled target domain data based on multiple types of features. We solve our optimization problem by using the cutting-plane algorithm based on group-based multiple kernel learning. Comprehensive experiments on two datasets demonstrate the effectiveness of MDA-HS for event recognition in consumer videos.
****
Discriminative Color Descriptors
Rahat Khan, Joost van de Weijer, Fahad Shahbaz Khan, Damien Muselet, Christophe Ducottet, Cecile Barat; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2866-2873
Color description is a challenging task because of large variations in RGB value

s which occur due to scene accidental events, such as shadows, shading, specularities, illuminant color changes, and changes in viewing geometry. Traditionally, this challenge has been addressed by capturing the variations in physics-based models, and deriving invariants for the undesired variations. The drawback of this approach is that sets of distinguishable colors in the original color space are mapped to the same value in the photometric invariant space. This results in a drop of discriminative power of the color description. In this paper we take an information theoretic approach to color description. We cluster color values together based on their discriminative power in a classification problem. The clustering has the explicit objective to minimize the drop of mutual information of the final representation. We show that such a color description automatically learns a certain degree of photometric invariance. We also show that a universal color representation, which is based on other data sets than the one at hand, can obtain competing performance. Experiments show that the proposed descriptor outperforms existing photometric invariants. Furthermore, we show that combined with shape description these color descriptors obtain excellent results on four challenging datasets, namely, PASCAL VOC 2007, Flowers-102, Stanford dogs-120 and Birds-200.

****

Optical Flow Estimation Using Laplacian Mesh Energy

Wenbin Li, Darren Cosker, Matthew Brown, Rui Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2435-2442

In this paper we present a novel non-rigid optical flow algorithm for dense image correspondence and non-rigid registration. The algorithm uses a unique Laplacian Mesh Energy term to encourage local smoothness whilst simultaneously preserving non-rigid deformation. Laplacian deformation approaches have become popular in graphics research as they enable mesh deformations to preserve local surface shape. In this work we propose a novel Laplacian Mesh Energy formula to ensure such sensible local deformations between image pairs. We express this wholly within the optical flow optimization, and show its application in a novel coarse-to-fine pyramidal approach. Our algorithm achieves the state-of-the-art performance in all trials on the Garg et al. dataset, and top tier performance on the Middlebury evaluation.

****

Constrained Clustering and Its Application to Face Clustering in Videos

Baoyuan Wu, Yifan Zhang, Bao-Gang Hu, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3507-3514

In this paper, we focus on face clustering in videos. Given the detected faces from real-world videos, we partition all faces into K disjoint clusters. Different from clustering on a collection of facial images, the faces from videos are organized as face tracks and the frame index of each face is also provided. As a result, many pairwise constraints between faces can be easily obtained from the temporal and spatial knowledge of the face tracks. These constraints can be effectively incorporated into a generative clustering model based on the Hidden Markov Random Fields (HMRFs). Within the HMRF model, the pairwise constraints are augmented by label-level and constraint-level local smoothness to guide the clustering process. The parameters for both the unary and the pairwise potential functions are learned by the simulated field algorithm, and the weights of constraints can be easily adjusted. We further introduce an efficient clustering framework specially for face clustering in videos, considering that faces in adjacent frames of the same face track are very similar. The framework is applicable to other clustering algorithms to significantly reduce the computational cost. Experiments on two face data sets from real-world videos demonstrate the significantly improved performance of our algorithm over state-of-theart algorithms.

****

Subcategory-Aware Object Classification

Jian Dong, Wei Xia, Qiang Chen, Jianshi Feng, Zhongyang Huang, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 827-834

In this paper, we introduce a subcategory-aware object classification framework to boost category level object classification performance. Motivated by the observation of considerable intra-class diversities and inter-class ambiguities in many current object classification datasets, we explicitly split data into subcategories by ambiguity guided subcategory mining. We then train an individual model for each subcategory rather than attempt to represent an object category with a monolithic model. More specifically, we build the instance affinity graph by combining both intraclass similarity and inter-class ambiguity. Visual subcategories, which correspond to the dense subgraphs, are detected by the graph shift algorithm and seamlessly integrated into the state-of-the-art detection assisted classification framework. Finally the responses from subcategory models are aggregated by subcategory-aware kernel regression. The extensive experiments over the PASCAL VOC 2007 and PASCAL VOC 2010 databases show the state-ofthe-art performance from our framework.

****

Face Recognition in Movie Trailers via Mean Sequence Sparse Representation-Based Classification

Enrique G. Ortiz, Alan Wright, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3531-3538

This paper presents an end-to-end video face recognition system, addressing the difficult problem of identifying a video face track using a large dictionary of still face images of a few hundred people, while rejecting unknown individuals. A straightforward application of the popular n-minimization for face recognition on a frame-by-frame basis is prohibitively expensive, so we propose a novel algorithm Mean Sequence SRC (MSSRC) that performs video face recognition using a joint optimization leveraging all of the available video data and the knowledge that the face track frames belong to the same individual. By adding a strict temporal constraint to the ii-minimization that forces individual frames in a face track to all reconstruct a single identity, we show the optimization reduces to a single minimization over the mean of the face track. We also introduce a new Movie Trailer Face Dataset collected from 101 movie trailers on YouTube. Finally, we show that our method matches or outperforms the state-of-the-art on three existing datasets (YouTube Celebrities, YouTube Faces, and Buffy) and our unconstrained Movie Trailer Face Dataset. More importantly, our method excels at rejecting unknown identities by at least 8% in average precision.

****

Multi-attribute Queries: To Merge or Not to Merge?

Mohammad Rastegari, Ali Diba, Devi Parikh, Ali Farhadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3310-3317

Users often have very specific visual content in mind that they are searching for. The most natural way to communicate this content to an image search engine is to use keywords that specify various properties or attributes of the content. A naive way of dealing with such multi-attribute queries is the following: train a classifier for each attribute independently, and then combine their scores on images to judge their fit to the query. We argue that this may not be the most effective or efficient approach. Conjunctions of attribute often correspond to very characteristic appearances. It would thus be beneficial to train classifiers that detect these conjunctions as a whole. But not all conjunctions result in such tight appearance clusters. So given a multi-attribute query, which conjunctions should we model? An exhaustive evaluation of all possible conjunctions would be time consuming. Hence we propose an optimization approach that identifies beneficial conjunctions without explicitly training the corresponding classifier. It reasons about geometric quantities that capture notions similar to intraand inter-class variances. We exploit a discriminative binary space to compute these geometric quantities efficiently. Experimental results on two challenging datasets of objects and birds show that our proposed approach can improve performance significantly over several strong baselines, while being an order of magnitude faster than exhaustively searching through all possible conjunctions.

****

Towards Efficient and Exact MAP-Inference for Large Scale Discrete Computer Visi on Problems via Combinatorial Optimization

Jorg Hendrik Kappes, Markus Speth, Gerhard Reinelt, Christoph Schnorr; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1752-1758

Discrete graphical models (also known as discrete Markov random fields) are a major conceptual tool to model the structure of optimization problems in computer vision. While in the last decade research has focused on fast approximative methods, algorithms that provide globally optimal solutions have come more into the research focus in the last years. However, large scale computer vision problems seemed to be out of reach for such methods. In this paper we introduce a promising way to bridge this gap based on partial optimality and structural properties of the underlying problem factorization. Combining these preprocessing steps, we are able to solve grids of size 2048 x 2048 in less than 90 seconds. On the hitherto unsolvable Chinese character dataset of Nowozin et al. we obtain provably optimal results in 56% of the instances and achieve competitive runtimes on other recent benchmark problems. While in the present work only generalized Potts models are considered, an extension to general graphical models seems to be feasible.

****

Plane-Based Content Preserving Warps for Video Stabilization

Zihan Zhou, Hailin Jin, Yi Ma; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2299-2306

Recently, a new image deformation technique called content-preserving warping (CPW) has been successfully employed to produce the state-of-the-art video stabilization results in many challenging cases. The key insight of CPW is that the true image deformation due to viewpoint change can be well approximated by a carefully constructed warp using a set of sparsely constructed 3D points only. However, since CPW solely relies on the tracked feature points to guide the warping, it works poorly in large textureless regions, such as ground and building interiors. To overcome this limitation, in this paper we present a hybrid approach for novel view synthesis, observing that the textureless regions often correspond to large planar surfaces in the scene. Particularly, given a jittery video, we first segment each frame into piecewise planar regions as well as regions labeled as non-planar using Markov random fields. Then, a new warp is computed by estimating a single homography for regions belong to the same plane, while inheriting results from CPW in the non-planar regions. We demonstrate how the segmentation information can be efficiently obtained and seamlessly integrated into the stabilization framework. Experimental results on a variety of real video sequences verify the effectiveness of our method.

****

Three-Dimensional Bilateral Symmetry Plane Estimation in the Phase Domain

Ramakrishna Kakarala, Prabhu Kaliamoorthi, Vittal Premachandran; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 249-256

We show that bilateral symmetry plane estimation for three-dimensional (3-D) shapes may be carried out accurately, and efficiently, in the spherical harmonic domain. Our methods are valuable for applications where spherical harmonic expansion is already employed, such as 3-D shape registration, morphometry, and retrieval. We show that the presence of bilateral symmetry in the 3-D shape is equivalent to a linear phase structure in the corresponding spherical harmonic coefficients, and provide algorithms for estimating the orientation of the symmetry plane. The benefit of using spherical harmonic phase is that symmetry estimation reduces to matching a compact set of descriptors, without the need to solve a correspondence problem. Our methods work on point clouds as well as large-scale mesh models of 3-D shapes.

****

Watching Unlabeled Video Helps Learn New Human Actions from Very Few Labeled Snapshots

Chao-Yeh Chen, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 572-579

We propose an approach to learn action categories from static images that leverages prior observations of generic human motion to augment its training process. Using unlabeled video containing various human activities, the system first learns how body pose tends to change locally in time. Then, given a small number of labeled static images, it uses that model to extrapolate beyond the given exemplars and generate "synthetic" training examples--poses that could link the observed images and/or immediately precede or follow them in time. In this way, we expand the training set without requiring additional manually labeled examples. We explore both example-based and manifold-based methods to implement our idea. Applying our approach to recognize actions in both images and video, we show it enhances a state-of-the-art technique when very few labeled training examples are available.

****

Long-Term Occupancy Analysis Using Graph-Based Optimisation in Thermal Imagery

Rikke Gade, Anders Jorgensen, Thomas B. Moeslund; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3698-3705

This paper presents a robust occupancy analysis system for thermal imaging. Reliable detection of people is very hard in crowded scenes, due to occlusions and segmentation problems. We therefore propose a framework that optimises the occupancy analysis over long periods by including information on the transition in occupancy, when people enter or leave the monitored area. In stable periods, with no activity close to the borders, people are detected and counted which contributes to a weighted histogram. When activity close to the border is detected, local tracking is applied in order to identify a crossing. After a full sequence, the number of people during all periods are estimated using a probabilistic graph search optimisation. The system is tested on a total of 51,000 frames, captured in sports arenas. The mean error for a 30-minute period containing 3-13 people is 4.44 %, which is a half of the error percentage optained by detection only, and better than the results of comparable work. The framework is also tested on a public available dataset from an outdoor scene, which proves the generality of the method.

****

Topical Video Object Discovery from Key Frames by Modeling Word Co-occurrence Prior

Gangqiang Zhao, Junsong Yuan, Gang Hua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1602-1609

A topical video object refers to an object that is frequently highlighted in a video. It could be, e.g., the product logo and the leading actor/actress in a TV commercial. We propose a topic model that incorporates a word co-occurrence prior for efficient discovery of topical video objects from a set of key frames. Previous work using topic models, such as Latent Dirichelet Allocation (LDA), for video object discovery often takes a bag-of-visual-words representation, which ignored important co-occurrence information among the local features. We show that such data driven co-occurrence information from bottom-up can conveniently be incorporated in LDA with a Gaussian Markov prior, which combines top down probabilistic topic modeling with bottom up priors in a unified model. Our experiments on challenging videos demonstrate that the proposed approach can discover different types of topical objects despite variations in scale, view-point, color and lighting changes, or even partial occlusions. The efficacy of the co-occurrence prior is clearly demonstrated when comparing with topic models without such priors.

****

Keypoints from Symmetries by Wave Propagation

Samuele Salti, Alessandro Lanza, Luigi Di Stefano; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2898-2905

The paper conjectures and demonstrates that repeatable keypoints based on salient symmetries at different scales can be detected by a novel analysis grounded on the wave equation rather than the heat equation underlying traditional Gaussian scale-space theory. While the image structures found by most state-of-the-art detectors, such as blobs and corners, occur typically on planar highly textured s

urfaces, salient symmetries are widespread in diverse kinds of images, including those related to untextured objects, which are hardly dealt with by current feature-based recognition pipelines. We provide experimental results on standard datasets and also contribute with a new dataset focused on untextured objects. Based on the positive experimental results, we hope to foster further research on the promising topic of scale invariant analysis through the wave equation.
****

Robust Real-Time Tracking of Multiple Objects by Volumetric Mass Densities

Horst Possegger, Sabine Sternig, Thomas Mauthner, Peter M. Roth, Horst Bischof; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2395-2402

Combining foreground images from multiple views by projecting them onto a common ground-plane has been recently applied within many multi-object tracking approaches. These planar projections introduce severe artifacts and constrain most approaches to objects moving on a common 2D ground-plane. To overcome these limitations, we introduce the concept of an occupancy volume exploiting the full geometry and the objects' center of mass and develop an efficient algorithm for 3D object tracking. Individual objects are tracked using the local mass density scores within a particle filter based approach, constrained by a Voronoi partitioning between nearby trackers. Our method benefits from the geometric knowledge given by the occupancy volume to robustly extract features and train classifiers on-demand, when volumetric information becomes unreliable. We evaluate our approach on several challenging real-world scenarios including the public APIDIS dataset. Experimental evaluations demonstrate significant improvements compared to state-of-theart methods, while achieving real-time performance.
****

A Divide-and-Conquer Method for Scalable Low-Rank Latent Matrix Pursuit

Yan Pan, Hanjiang Lai, Cong Liu, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 524-531

Data fusion, which effectively fuses multiple prediction lists from different kinds of features to obtain an accurate model, is a crucial component in various computer vision applications. Robust late fusion (RLF) is a recent proposed method that fuses multiple output score lists from different models via pursuing a shared low-rank latent matrix. Despite showing promising performance, the repeated full Singular Value Decomposition operations in RLF's optimization algorithm limits its scalability in real world vision datasets which usually have large number of test examples. To address this issue, we provide a scalable solution for large-scale low-rank latent matrix pursuit by a divide-andconquer method. The proposed method divides the original low-rank latent matrix learning problem into two sizereduced subproblems, which may be solved via any base algorithm, and combines the results from the subproblems to obtain the final solution. Our theoretical analysis shows that with fixed probability, the proposed divide-and-conquer method has recovery guarantees comparable to those of its base algorithm. Moreover, we develop an efficient base algorithm for the corresponding subproblems by factorizing a large matrix into the product of two size-reduced matrices. We also provide high probability recovery guarantees of the base algorithm. The proposed method is evaluated on various fusion problems in object categorization and video event detection. Under comparable accuracy, the proposed method performs more than 180 times faster than the stateof-the-art baselines on the CCV dataset with about 4,500 test examples for video event detection.
****

PDM-ENLOR: Learning Ensemble of Local PDM-Based Regressions

Yen H. Le, Uday Kurkure, Ioannis A. Kakadiaris; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1878-1885

Statistical shape models, such as Active Shape Models (ASMs), suffer from their inability to represent a large range of variations of a complex shape and to account for the large errors in detection of model points. We propose a novel method (dubbed PDM-ENLOR) that overcomes these limitations by locating each shape model point individually using an ensemble of local regression models and appearance cues from selected model points. Our method first detects a set of reference p oints which were selected based on their saliency during training. For each model point, an ensemble of regressors is built. From the locations of the detected reference points, each regressor infers a candidate location for that model point using local geometric constraints, encoded by a point distribution model (PDM). The final location of that point is determined as a weighted linear combination, whose coefficients are learnt from the training data, of candidates proposed from its ensemble's component regressors. We use different subsets of reference points as explanatory variables for the component regressors to provide varying degrees of locality for the models in each ensemble. This helps our ensemble model to capture a larger range of shape variations as compared to a single PDM. We demonstrate the advantages of our method on the challenging problem of segmenting gene expression images of mouse brain.

****

Beta Process Joint Dictionary Learning for Coupled Feature Spaces with Application to Single Image Super-Resolution

Li He, Hairong Qi, Russell Zaretzki; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 345-352

This paper addresses the problem of learning overcomplete dictionaries for the coupled feature spaces, where the learned dictionaries also reflect the relationship between the two spaces. A Bayesian method using a beta process prior is applied to learn the over-complete dictionaries. Compared to previous couple feature spaces dictionary learning algorithms, our algorithm not only provides dictionaries that customized to each feature space, but also adds more consistent and accurate mapping between the two feature spaces. This is due to the unique property of the beta process model that the sparse representation can be decomposed to values and dictionary atom indicators. The proposed algorithm is able to learn sparse representations that correspond to the same dictionary atoms with the same sparsity but different values in coupled feature spaces, thus bringing consistent and accurate mapping between coupled feature spaces. Another advantage of the proposed method is that the number of dictionary atoms and their relative importance may be inferred non-parametrically. We compare the proposed approach to several state-of-the-art dictionary learning methods by applying this method to single image super-resolution. The experimental results show that dictionaries learned by our method produces the best superresolution results compared to other state-of-the-art methods.

****

Fast Trust Region for Segmentation

Lena Gorelick, Frank R. Schmidt, Yuri Boykov; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1714-1721

Trust region is a well-known general iterative approach to optimization which offers many advantages over standard gradient descent techniques. In particular, it allows more accurate nonlinear approximation models. In each iteration this approach computes a global optimum of a suitable approximation model within a fixed radius around the current solution, a.k.a. trust region. In general, this approach can be used only when some efficient constrained optimization algorithm is available for the selected nonlinear (more accurate) approximation model. In this paper we propose a Fast Trust Region (FTR) approach for optimization of segmentation energies with nonlinear regional terms, which are known to be challenging for existing algorithms. These energies include, but are not limited to, KL divergence and Bhattacharyya distance between the observed and the target appearance distributions, volume constraint on segment size, and shape prior constraint in a form of L 2 distance from target shape moments. Our method is 1-2 orders of magnitude faster than the existing state-of-the-art methods while converging to comparable or better solutions.

****

Area Preserving Brain Mapping

Zhengyu Su, Wei Zeng, Rui Shi, Yalin Wang, Jian Sun, Xianfeng Gu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2235-2242

Brain mapping transforms the brain cortical surface to canonical planar domains, which plays a fundamental role in morphological study. Most existing brain mapping methods are based on angle preserving maps, which may introduce large area distortions. This work proposes an area preserving brain mapping method based on MongeBrenier theory. The brain mapping is intrinsic to the Riemannian metric, unique, and diffeomorphic. The computation is equivalent to convex energy minimization and power Voronoi diagram construction. Comparing to the existing approaches based on Monge-Kantorovich theory, the proposed one greatly reduces the complexity (from n ,tunknowns to n), and improves the simplicity and efficiency. Experimental results on caudate nucleus surface mapping and cortical surface mapping demonstrate the efficacy and efficiency of the proposed method. Conventional methods for caudate nucleus surface mapping may suffer from numerical instability; in contrast, current method produces diffeomorpic mappings stably. In the study of cortical surface classification for recognition of Alzheimer's Disease, the proposed method outperforms some other morphometry features.

****

Multi-level Discriminative Dictionary Learning towards Hierarchical Visual Categorization

Li Shen, Shuhui Wang, Gang Sun, Shuqiang Jiang, Qingming Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 383-390

For the task of visual categorization, the learning model is expected to be endowed with discriminative visual feature representation and flexibilities in processing many categories. Many existing approaches are designed based on a flat category structure, or rely on a set of pre-computed visual features, hence may not be appreciated for dealing with large numbers of categories. In this paper, we propose a novel dictionary learning method by taking advantage of hierarchical category correlation. For each internode of the hierarchical category structure, a discriminative dictionary and a set of classification models are learnt for visual categorization, and the dictionaries in different layers are learnt to exploit the discriminative visual properties of different granularity. Moreover, the dictionaries in lower levels also inherit the dictionary of ancestor nodes, so that categories in lower levels are described with multi-scale visual information using our dictionary learning approach. Experiments on ImageNet object data subset and SUN397 scene dataset demonstrate that our approach achieves promising performance on data with large numbers of classes compared with some state-of-the-art methods, and is more efficient in processing large numbers of categories.

****

BFO Meets HOG: Feature Extraction Based on Histograms of Oriented p.d.f. Gradients for Image Classification

Takumi Kobayashi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 747-754

Image classification methods have been significantly developed in the last decade. Most methods stem from bagof-features (BoF) approach and it is recently extended to a vector aggregation model, such as using Fisher kernels. In this paper, we propose a novel feature extraction method for image classification. Following the BoF approach, a plenty of local descriptors are first extracted in an image and the proposed method is built upon the probability density function (p.d.f) formed by those descriptors. Since the p.d.f essentially represents the image, we extract the features from the p.d.f by means of the gradients on the p.d.f. The gradients, especially their orientations, effectively characterize the shape of the p.d.f from the geometrical viewpoint. We construct the features by the histogram of the oriented p.d.f gradients via orientation coding followed by aggregation of the orientation codes. The proposed image features, imposing no specific assumption on the targets, are so general as to be applicable to any kinds of tasks regarding image classifications. In the experiments on object recognition and scene classification using various datasets, the proposed method exhibits superior performances compared to the other existing methods.

****

Single-Sample Face Recognition with Image Corruption and Misalignment via Sparse Illumination Transfer

Liansheng Zhuang, Allen Y. Yang, Zihan Zhou, S. Shankar Sastry, Yi Ma; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3546-3553

Single-sample face recognition is one of the most challenging problems in face recognition. We propose a novel face recognition algorithm to address this problem based on a sparse representation based classification (SRC) framework. The new algorithm is robust to image misalignment and pixel corruption, and is able to reduce required training images to one sample per class. To compensate the missing illumination information typically provided by multiple training images, a sparse illumination transfer (SIT) technique is introduced. The SIT algorithms seek additional illumination examples of face images from one or more additional subject classes, and form an illumination dictionary. By enforcing a sparse representation of the query image, the method can recover and transfer the pose and illumination information from the alignment stage to the recognition stage. Our extensive experiments have demonstrated that the new algorithms significantly outperform the existing algorithms in the single-sample regime and with less restrictions. In particular, the face alignment accuracy is comparable to that of the well-known Deformable SRC algorithm using multiple training images; and the face recognition accuracy exceeds those of the SRC and Extended SRC algorithms using hand labeled alignment initialization.

****

GeoF: Geodesic Forests for Learning Coupled Predictors

Peter Kontschieder, Pushmeet Kohli, Jamie Shotton, Antonio Criminisi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 65-72

Conventional decision forest based methods for image labelling tasks like object segmentation make predictions for each variable (pixel) independently [3, 5, 8]. This prevents them from enforcing dependencies between variables and translates into locally inconsistent pixel labellings. Random field models, instead, encourage spatial consistency of labels at increased computational expense. This paper presents a new and efficient forest based model that achieves spatially consistent semantic image segmentation by encoding variable dependencies directly in the feature space the forests operate on. Such correlations are captured via new long-range, soft connectivity features, computed via generalized geodesic distance transforms. Our model can be thought of as a generalization of the successful Semantic Texton Forest, Auto-Context, and Entangled Forest models. A second contribution is to show the connection between the typical Conditional Random Field (CRF) energy and the forest training objective. This analysis yields a new objective for training decision forests that encourages more accurate structured prediction. Our GeoF model is validated quantitatively on the task of semantic image segmentation, on four challenging and very diverse image datasets. GeoF outperforms both stateof-the-art forest models and the conventional pairwise CRF.

****

Improving Image Matting Using Comprehensive Sampling Sets

Ehsan Shahrian, Deepu Rajan, Brian Price, Scott Cohen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 636-643

In this paper, we present a new image matting algorithm that achieves state-of-the-art performance on a benchmark dataset of images. This is achieved by solving two major problems encountered by current sampling based algorithms. The first is that the range in which the foreground and background are sampled is often limited to such an extent that the true foreground and background colors are not present. Here, we describe a method by which a more comprehensive and representative set of samples is collected so as not to miss out on the true samples. This is accomplished by expanding the sampling range for pixels farther from the foreground or background boundary and ensuring that samples from each color distribution are included. The second problem is the overlap in color distributions of foreground and background regions. This causes sampling based methods to fail to pick the correct samples for foreground and background. Our design of an objective function forces those foreground and background samples to be picked that are generated from well-separated distributions. Comparison on the dataset at and e valuation by www.alphamatting.com shows that the proposed method ranks first in terms of error measures used in the website.

****

Sketch Tokens: A Learned Mid-level Representation for Contour and Object Detection

Joseph J. Lim, C. L. Zitnick, Piotr Dollar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3158-3165

We propose a novel approach to both learning and detecting local contour-based representations for mid-level features. Our features, called sketch tokens, are learned using supervised mid-level information in the form of hand drawn contours in images. Patches of human generated contours are clustered to form sketch token classes and a random forest classifier is used for efficient detection in novel images. We demonstrate our approach on both topdown and bottom-up tasks. We show state-of-the-art results on the top-down task of contour detection while being over 200x faster than competing methods. We also achieve large improvements in detection accuracy for the bottom-up tasks of pedestrian and object detection as measured on INRIA [5] and PASCAL [10], respectively. These gains are due to the complementary information provided by sketch tokens to low-level features such as gradient histograms.

****

Subspace Interpolation via Dictionary Learning for Unsupervised Domain Adaptation

Jie Ni, Qiang Qiu, Rama Chellappa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 692-699

Domain adaptation addresses the problem where data instances of a source domain have different distributions from that of a target domain, which occurs frequently in many real life scenarios. This work focuses on unsupervised domain adaptation, where labeled data are only available in the source domain. We propose to interpolate subspaces through dictionary learning to link the source and target domains. These subspaces are able to capture the intrinsic domain shift and form a shared feature representation for cross domain recognition. Further, we introduce a quantitative measure to characterize the shift between two domains, which enables us to select the optimal domain to adapt to the given multiple source domains. We present experiments on face recognition across pose, illumination and blur variations, cross dataset object recognition, and report improved performance over the state of the art.

****

Probabilistic Graphlet Cut: Exploiting Spatial Structure Cue for Weakly Supervised Image Segmentation

Luming Zhang, Mingli Song, Zicheng Liu, Xiao Liu, Jiajun Bu, Chun Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1908-1915

Weakly supervised image segmentation is a challenging problem in computer vision field. In this paper, we present a new weakly supervised image segmentation algorithm by learning the distribution of spatially structured superpixel sets from image-level labels. Specifically, we first extract graphlets from each image where a graphlet is a smallsized graph consisting of superpixels as its nodes and it encapsulates the spatial structure of those superpixels. Then, a manifold embedding algorithm is proposed to transform graphlets of different sizes into equal-length feature vectors. Thereafter, we use GMM to learn the distribution of the post-embedding graphlets. Finally, we propose a novel image segmentation algorithm, called graphlet cut, that leverages the learned graphlet distribution in measuring the homogeneity of a set of spatially structured superpixels. Experimental results show that the proposed approach outperforms state-of-the-art weakly supervised image segmentation methods, and its performance is comparable to those of the fully supervised segmentation models.

****

Fast Energy Minimization Using Learned State Filters

Matthieu Guillaumin, Luc Van Gool, Vittorio Ferrari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1682-1689

Pairwise discrete energies defined over graphs are ubiquitous in computer vision. Many algorithms have been proposed to minimize such energies, often concentrating on sparse graph topologies or specialized classes of pairwise potentials. However, when the graph is fully connected and the pairwise potentials are arbitrary, the complexity of even approximate minimization algorithms such as TRW-S grows quadratically both in the number of nodes and in the number of states a node can take. Moreover, recent applications are using more and more computationally expensive pairwise potentials. These factors make it very hard to employ fully connected models. In this paper we propose a novel, generic algorithm to approximately minimize any discrete pairwise energy function. Our method exploits tractable sub-energies to filter the domain of the function. The parameters of the filter are learnt from instances of the same class of energies with good candidate solutions. Compared to existing methods, it efficiently handles fully connected graphs, with many states per node, and arbitrary pairwise potentials, which might be expensive to compute. We demonstrate experimentally on two applications that our algorithm is much more efficient than other generic minimization algorithms such as TRW-S, while returning essentially identical solutions.
****

Learning Binary Codes for High-Dimensional Data Using Bilinear Projections
Yunchao Gong, Sanjiv Kumar, Henry A. Rowley, Svetlana Lazebnik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 484-491

Recent advances in visual recognition indicate that to achieve good retrieval and classification accuracy on largescale datasets like ImageNet, extremely high-dimensional visual descriptors, e.g., Fisher Vectors, are needed. We present a novel method for converting such descriptors to compact similarity-preserving binary codes that exploits their natural matrix structure to reduce their dimensionality using compact bilinear projections instead of a single large projection matrix. This method achieves comparable retrieval and classification accuracy to the original descriptors and to the state-of-the-art Product Quantization approach while having orders of magnitude faster code generation time and smaller memory footprint.
****

Multi-scale Curve Detection on Surfaces
Michael Kolomenkin, Ilan Shimshoni, Ayellet Tal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 225-232

This paper extends to surfaces the multi-scale approach of edge detection on images. The common practice for detecting curves on surfaces requires the user to first select the scale of the features, apply an appropriate smoothing, and detect the edges on the smoothed surface. This approach suffers from two drawbacks. First, it relies on a hidden assumption that all the features on the surface are of the same scale. Second, manual user intervention is required. In this paper, we propose a general framework for automatically detecting the optimal scale for each point on the surface. We smooth the surface at each point according to this optimal scale and run the curve detection algorithm on the resulting surface. Our multi-scale algorithm solves the two disadvantages of the single-scale approach mentioned above. We demonstrate how to realize our approach on two commonly-used special cases: ridges & valleys and relief edges. In each case, the optimal scale is found in accordance with the mathematical definition of the curve.
****

Saliency Aggregation: A Data-Driven Approach
Long Mai, Yuzhen Niu, Feng Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1131-1138

A variety of methods have been developed for visual saliency analysis. These methods often complement each other. This paper addresses the problem of aggregating various saliency analysis methods such that the aggregation result outperforms each individual one. We have two major observations. First, different methods perform differently in saliency analysis. Second, the performance of a saliency analysis method varies with individual images. Our idea is to use data-driven approaches to saliency aggregation that appropriately consider the performance gaps among individual methods and the performance dependence of each method on individual images. This paper discusses various data-driven approaches and finds that the image-dependent aggregation method works best. Specifically, our method uses a Conditional Random Field (CRF) framework for saliency aggregation that not only models the contribution from individual saliency map but also the interaction between neighboring pixels. To account for the dependence of aggregation on an individual image, our approach selects a subset of images similar to the input image from a training data set and trains the CRF aggregation model only using this subset instead of the whole training set. Our experiments on public saliency benchmarks show that our aggregation method outperforms each individual saliency method and is robust with the selection of aggregated methods.

****

Crossing the Line: Crowd Counting by Integer Programming with Local Features

Zheng Ma, Antoni B. Chan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2539-2546

We propose an integer programming method for estimating the instantaneous count of pedestrians crossing a line of interest in a video sequence. Through a line sampling process, the video is first converted into a temporal slice image. Next, the number of people is estimated in a set of overlapping sliding windows on the temporal slice image, using a regression function that maps from local features to a count. Given that count in a sliding window is the sum of the instantaneous counts in the corresponding time interval, an integer programming method is proposed to recover the number of pedestrians crossing the line of interest in each frame. Integrating over a specific time interval yields the cumulative count of pedestrian crossing the line. Compared with current methods for line counting, our proposed approach achieves state-of-the-art performance on several challenging crowd video datasets.

****

Discriminative Subspace Clustering

Vasileios Zografos, Liam Ellis, Rudolf Mester; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2107-2114

We present a novel method for clustering data drawn from a union of arbitrary dimensional subspaces, called Discriminative Subspace Clustering (DiSC). DiSC solves the subspace clustering problem by using a quadratic classifier trained from unlabeled data (clustering by classification). We generate labels by exploiting the locality of points from the same subspace and a basic affinity criterion. A number of classifiers are then diversely trained from different partitions of the data, and their results are combined together in an ensemble, in order to obtain the final clustering result. We have tested our method with 4 challenging datasets and compared against 8 state-of-the-art methods from literature. Our results show that DiSC is a very strong performer in both accuracy and robustness, and also of low computational complexity.

****

Measuring Crowd Collectiveness

Bolei Zhou, Xiaoou Tang, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3049-3056

Collective motions are common in crowd systems and have attracted a great deal of attention in a variety of multidisciplinary fields. Collectiveness, which indicates the degree of individuals acting as a union in collective motion, is a fundamental and universal measurement for various crowd systems. By integrating path similarities among crowds on collective manifold, this paper proposes a descriptor of collectiveness and an efficient computation for the crowd and its constituent individuals. The algorithm of the Collective Merging is then proposed to detect collective motions from random motions. We validate the effectiveness and robustness of the proposed collectiveness descriptor on the system of self-driven particles. We then compare the collectiveness descriptor to human perception for collective motion and show high consistency. Our experiments regarding the detection of collective motions and the measurement of collectiveness in videos of pedestrian crowds and bacteria colony demonstrate a wide range of applications of the collectiveness descriptor 1 .

****

MKPLS: Manifold Kernel Partial Least Squares for Lipreading and Speaker Identification

Amr Bakry, Ahmed Elgammal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 684-691

Visual speech recognition is a challenging problem, due to confusion between visual speech features. The speaker identification problem is usually coupled with speech recognition. Moreover, speaker identification is important to several applications, such as automatic access control, biometrics, authentication, and personal privacy issues. In this paper, we propose a novel approach for lipreading and speaker identification. We propose a new approach for manifold parameterization in a low-dimensional latent space, where each manifold is represented as a point in that space. We initially parameterize each instance manifold using a nonlinear mapping from a unified manifold representation. We then factorize the parameter space using Kernel Partial Least Squares (KPLS) to achieve a low-dimension manifold latent space. We use two-way projections to achieve two manifold latent spaces, one for the speech content and one for the speaker. We apply our approach on two public databases: AVLetters and OuluVS. We show the results for three different settings of lipreading: speaker independent, speaker dependent, and speaker semi-dependent. Our approach outperforms for the speaker semi-dependent setting by at least 15% of the baseline, and competes in the other two settings.

****

Whitened Expectation Propagation: Non-Lambertian Shape from Shading and Shadow

Brian Potetz, Mohammadreza Hajiarbabi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1674-1681

For problems over continuous random variables, MRFs with large cliques pose a challenge in probabilistic inference. Difficulties in performing optimization efficiently have limited the probabilistic models explored in computer vision and other fields. One inference technique that handles large cliques well is Expectation Propagation. EP offers run times independent of clique size, which instead depend only on the rank, or intrinsic dimensionality, of potentials. This property would be highly advantageous in computer vision. Unfortunately, for grid-shaped models common in vision, traditional Gaussian EP requires quadratic space and cubic time in the number of pixels. Here, we propose a variation of EP that exploits regularities in natural scene statistics to achieve run times that are linear in both number of pixels and clique size. We test these methods on shape from shading, and we demonstrate strong performance not only for Lambertian surfaces, but also on arbitrary surface reflectance and lighting arrangements, which requires highly non-Gaussian potentials. Finally, we use large, non-local cliques to exploit cast shadow, which is traditionally ignored in shape from shading.

****

Multi-class Video Co-segmentation with a Generative Multi-video Model

Wei-Chen Chiu, Mario Fritz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 321-328

Video data provides a rich source of information that is available to us today in large quantities e.g. from online resources. Tasks like segmentation benefit greatly from the analysis of spatio-temporal motion patterns in videos and recent advances in video segmentation has shown great progress in exploiting these addition cues. However, observing a single video is often not enough to predict meaningful segmentations and inference across videos becomes necessary in order to predict segmentations that are consistent with objects classes. Therefore the task of video cosegmentation is being proposed, that aims at inferring segmentation from multiple videos. But current approaches are limited to only considering binary foreground/background segmentation and multiple videos of the same object. This is a clear mismatch to the challenges that we are facing with videos from online resources or consumer videos. We propose to study multi-class video co-segmentation where the number of object classes is unknown as well as the number of instances in each frame and video. We achieve this by formulating a non-parametric bayesian model across videos sequences that is based on a new videos segmentation prior as well as a global appearance model that links segments of the sam e class. We present the first multi-class video co-segmentation evaluation. We show that our method is applicable to real video data from online resources and outperforms state-of-the-art video segmentation and image co-segmentation baselines.

****

Lp-Norm IDF for Large Scale Image Search
Liang Zheng, Shengjin Wang, Ziqiong Liu, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1626-1633

The Inverse Document Frequency (IDF) is prevalently utilized in the Bag-of-Words based image search. The basic idea is to assign less weight to terms with high frequency, and vice versa. However, the estimation of visual word frequency is coarse and heuristic. Therefore, the effectiveness of the conventional IDF routine is marginal, and far from optimal. To tackle this problem, this paper introduces a novel IDF expression by the use of L p -norm pooling technique. Carefully designed, the proposed IDF takes into account the term frequency, document frequency, the complexity of images, as well as the codebook information. Optimizing the IDF function towards optimal balancing between TF and pIDF weights yields the so-called L p -norm IDF (pIDF). We show that the conventional IDF is a special case of our generalized version, and two novel IDFs, i.e. the average IDF and the max IDF, can also be derived from our formula. Further, by counting for the term-frequency in each image, the proposed L p -norm IDF helps to alleviate the visual word burstiness phenomenon. Our method is evaluated through extensive experiments on three benchmark datasets (Oxford 5K, Paris 6K and Flickr 1M). We report a performance improvement of as large as 27.1% over the baseline approach. Moreover, since the L p -norm IDF is computed offline, no extra computation or memory cost is introduced to the system at all.

****

Saliency Detection via Graph-Based Manifold Ranking
Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3166-3173

Most existing bottom-up methods measure the foreground saliency of a pixel or region based on its contrast within a local context or the entire image, whereas a few methods focus on segmenting out background regions and thereby salient objects. Instead of considering the contrast between the salient objects and their surrounding regions, we consider both foreground and background cues in a different way. We rank the similarity of the image elements (pixels or regions) with foreground cues or background cues via graph-based manifold ranking. The saliency of the image elements is defined based on their relevances to the given seeds or queries. We represent the image as a close-loop graph with superpixels as nodes. These nodes are ranked based on the similarity to background and foreground queries, based on affinity matrices. Saliency detection is carried out in a two-stage scheme to extract background regions and foreground salient objects efficiently. Experimental results on two large benchmark databases demonstrate the proposed method performs well when against the state-of-the-art methods in terms of accuracy and speed. We also create a more difficult benchmark database containing 5,172 images to test the proposed saliency model and make this database publicly available with this paper for further studies in the saliency field.

****

Online Object Tracking: A Benchmark
Yi Wu, Jongwoo Lim, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2411-2418

Object tracking is one of the most important components in numerous applications of computer vision. While much progress has been made in recent years with efforts on sharing code and datasets, it is of great importance to develop a library and benchmark to gauge the state of the art. After briefly reviewing recent advances of online object tracking, we carry out large scale experiments with various evaluation criteria to understand how these algorithms perform. The test image sequences are annotated with different attributes for performance evaluation and analysis. By analyzing quantitative results, we identify effective approaches for robust tracking and provide potential future research directions in this field.

****

Tracking Sports Players with Context-Conditioned Motion Models

Jingchen Liu, Peter Carr, Robert T. Collins, Yanxi Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1830-1837

We employ hierarchical data association to track players in team sports. Player movements are often complex and highly correlated with both nearby and distant players. A single model would require many degrees of freedom to represent the full motion diversity and could be difficult to use in practice. Instead, we introduce a set of Game Context Features extracted from noisy detections to describe the current state of the match, such as how the players are spatially distributed. Our assumption is that players react to the current situation in only a finite number of ways. As a result, we are able to select an appropriate simplified affinity model for each player and time instant using a random decision forest based on current track and game context features. Our context-conditioned motion models implicitly incorporate complex inter-object correlations while remaining tractable. We demonstrate significant performance improvements over existing multi-target tracking algorithms on basketball and field hockey sequences several minutes in duration and containing 10 and 20 players respectively.

****

Physically Plausible 3D Scene Tracking: The Single Actor Hypothesis

Nikolaos Kyriazis, Antonis Argyros; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 9-16

In several hand-object(s) interaction scenarios, the change in the objects' state is a direct consequence of the hand's motion. This has a straightforward representation in Newtonian dynamics. We present the first approach that exploits this observation to perform model-based 3D tracking of a table-top scene comprising passive objects and an active hand. Our forward modelling of 3D hand-object(s) interaction regards both the appearance and the physical state of the scene and is parameterized over the hand motion (26 DoFs) between two successive instants in time. We demonstrate that our approach manages to track the 3D pose of all objects and the 3D pose and articulation of the hand by only searching for the parameters of the hand motion. In the proposed framework, covert scene state is inferred by connecting it to the overt state, through the incorporation of physics. Thus, our tracking approach treats a variety of challenging observability issues in a principled manner, without the need to resort to heuristics.

****

Improved Image Set Classification via Joint Sparse Approximated Nearest Subspaces

Shaokang Chen, Conrad Sanderson, Mehrtash T. Harandi, Brian C. Lovell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 452-459

Existing multi-model approaches for image set classification extract local models by clustering each image set individually only once, with fixed clusters used for matching with other image sets. However, this may result in the two closest clusters to represent different characteristics of an object, due to different undesirable environmental conditions (such as variations in illumination and pose). To address this problem, we propose to constrain the clustering of each query image set by forcing the clusters to have resemblance to the clusters in the gallery image sets. We first define a Frobenius norm distance between subspaces over Grassmann manifolds based on reconstruction error. We then extract local linear subspaces from a gallery image set via sparse representation. For each local linear subspace, we adaptively construct the corresponding closest subspace from the samples of a probe image set by joint sparse representation. We show that by minimising the sparse representation reconstruction error, we approach the nearest point on a Grassmann manifold. Experiments on Honda, ETH-80 and Cambridge-Gesture datasets show that the proposed method consistently outperforms several other recent techniques, such as Affine Hull based Image Set Distance (AHISD), Sp arse Approximated Nearest Points (SANP) and Manifold Discriminant Analysis (MDA).

****

Underwater Camera Calibration Using Wavelength Triangulation

Timothy Yau, Minglun Gong, Yee-Hong Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2499-2506

In underwater imagery, the image formation process includes refractions that occur when light passes from water into the camera housing, typically through a flat glass port. We extend the existing work on physical refraction models by considering the dispersion of light, and derive new constraints on the model parameters for use in calibration. This leads to a novel calibration method that achieves improved accuracy compared to existing work. We describe how to construct a novel calibration device for our method and evaluate the accuracy of the method through synthetic and real experiments.

****

Expressive Visual Text-to-Speech Using Active Appearance Models

Robert Anderson, Bjorn Stenger, Vincent Wan, Roberto Cipolla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3382-3389

This paper presents a complete system for expressive visual text-to-speech (VTTS), which is capable of producing expressive output, in the form of a 'talking head', given an input text and a set of continuous expression weights. The face is modeled using an active appearance model (AAM), and several extensions are proposed which make it more applicable to the task of VTTS. The model allows for normalization with respect to both pose and blink state which significantly reduces artifacts in the resulting synthesized sequences. We demonstrate quantitative improvements in terms of reconstruction error over a million frames, as well as in large-scale user studies, comparing the output of different systems.

****

Joint Sparsity-Based Representation and Analysis of Unconstrained Activities

Raghuraman Gopalan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2738-2745

While the notion of joint sparsity in understanding common and innovative components of a multi-receiver signal ensemble has been well studied, we investigate the utility of such joint sparse models in representing information contained in a single video signal. By decomposing the content of a video sequence into that observed by multiple spatially and/or temporally distributed receivers, we first recover a collection of common and innovative components pertaining to individual videos. We then present modeling strategies based on subspace-driven manifold metrics to characterize patterns among these components, across other videos in the system, to perform subsequent video analysis. We demonstrate the efficacy of our approach for activity classification and clustering by reporting competitive results on standard datasets such as, HMDB, UCF-50, Olympic Sports and KTH.

****

Discriminative Brain Effective Connectivity Analysis for Alzheimer's Disease: A Kernel Learning Approach upon Sparse Gaussian Bayesian Network

Luping Zhou, Lei Wang, Lingqiao Liu, Philip Ogunbona, Dinggang Shen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2243-2250

Analyzing brain networks from neuroimages is becoming a promising approach in identifying novel connectivitybased biomarkers for the Alzheimer's disease (AD). In this regard, brain "effective connectivity" analysis, which studies the causal relationship among brain regions, is highly challenging and of many research opportunities. Most of the existing works in this field use generative methods. Despite their success in data representation and other important merits, generative methods are not necessarily discriminative, which may cause the ignorance of subtle but critical disease-induced changes. In this paper, we propose a learning-based approach that integrates the benefits of generative and discriminative methods to recover effective connectivity. In particular, we employ Fisher kernel to bridge the generative models of sparse Bayesian networks (SBN) and the discri minative classifiers of SVMs, and convert the SBN parameter learning to Fisher kernel learning via minimizing a generalization error bound of SVMs. Our method is able to simultaneously boost the discriminative power of both the generative SBN models and the SBN-induced SVM classifiers via Fisher kernel. The proposed method is tested on analyzing brain effective connectivity for AD from ADNI data, and demonstrates significant improvements over the state-of-the-art work.

****

Robust Monocular Epipolar Flow Estimation

Koichiro Yamaguchi, David McAllester, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1862-1869

We consider the problem of computing optical flow in monocular video taken from a moving vehicle. In this setting, the vast majority of image flow is due to the vehicle's ego-motion. We propose to take advantage of this fact and estimate flow along the epipolar lines of the egomotion. Towards this goal, we derive a slanted-plane MRF model which explicitly reasons about the ordering of planes and their physical validity at junctions. Furthermore, we present a bottom-up grouping algorithm which produces over-segmentations that respect flow boundaries. We demonstrate the effectiveness of our approach in the challenging KITTI flow benchmark [11] achieving half the error of the best competing general flow algorithm and one third of the error of the best epipolar flow algorithm.

****

Heterogeneous Visual Features Fusion via Sparse Multimodal Machine

Hua Wang, Feiping Nie, Heng Huang, Chris Ding; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3097-3102

To better understand, search, and classify image and video information, many visual feature descriptors have been proposed to describe elementary visual characteristics, such as the shape, the color, the texture, etc. How to integrate these heterogeneous visual features and identify the important ones from them for specific vision tasks has become an increasingly critical problem. In this paper, We propose a novel Sparse Multimodal Learning (SMML) approach to integrate such heterogeneous features by using the joint structured sparsity regularizations to learn the feature importance of for the vision tasks from both group-wise and individual point of views. A new optimization algorithm is also introduced to solve the non-smooth objective with rigorously proved global convergence. We applied our SMML method to five broadly used object categorization and scene understanding image data sets for both singlelabel and multi-label image classification tasks. For each data set we integrate six different types of popularly used image features. Compared to existing scene and object categorization methods using either single modality or multimodalities of features, our approach always achieves better performances measured.

****

A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching

Pradipto Das, Chenliang Xu, Richard F. Doell, Jason J. Corso; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2634-2641

The problem of describing images through natural language has gained importance in the computer vision community. Solutions to image description have either focused on a top-down approach of generating language through combinations of object detections and language models or bottom-up propagation of keyword tags from training images to test images through probabilistic or nearest neighbor techniques. In contrast, describing videos with natural language is a less studied problem. In this paper, we combine ideas from the bottom-up and top-down approaches to image description and propose a method for video description that captures the most relevant contents of a video in a natural language description. We propose a hybrid system consisting of a low level multimodal latent topic model for initial keyword annotation, a middle level of concept detectors and a high level module to produce final lingual descriptions. We compare the results of our system to human descriptions in both short and long forms on two datasets, and demonstrate that final system output has greater agreement with the human descriptions than any single level.

****

Online Dominant and Anomalous Behavior Detection in Videos

Mehrsan Javan Roshtkhari, Martin D. Levine; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2611-2618

We present a novel approach for video parsing and simultaneous online learning of dominant and anomalous behaviors in surveillance videos. Dominant behaviors are those occurring frequently in videos and hence, usually do not attract much attention. They can be characterized by different complexities in space and time, ranging from a scene background to human activities. In contrast, an anomalous behavior is defined as having a low likelihood of occurrence. We do not employ any models of the entities in the scene in order to detect these two kinds of behaviors. In this paper, video events are learnt at each pixel without supervision using densely constructed spatio-temporal video volumes. Furthermore, the volumes are organized into large contextual graphs. These compositions are employed to construct a hierarchical codebook model for the dominant behaviors. By decomposing spatio-temporal contextual information into unique spatial and temporal contexts, the proposed framework learns the models of the dominant spatial and temporal events. Thus, it is ultimately capable of simultaneously modeling high-level behaviors as well as low-level spatial, temporal and spatio-temporal pixel level changes.

****

Learning Class-to-Image Distance with Object Matchings

Guang-Tong Zhou, Tian Lan, Weilong Yang, Greg Mori; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 795-802

We conduct image classification by learning a class-toimage distance function that matches objects. The set of objects in training images for an image class are treated as a collage. When presented with a test image, the best matching between this collage of training image objects and those in the test image is found. We validate the efficacy of the proposed model on the PASCAL 07 and SUN 09 datasets, showing that our model is effective for object classification and scene classification tasks. State-of-the-art image classification results are obtained, and qualitative results demonstrate that objects can be accurately matched.

****

Spectral Modeling and Relighting of Reflective-Fluorescent Scenes

Antony Lam, Imari Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1452-1459

Hyperspectral reflectance data allows for highly accurate spectral relighting under arbitrary illumination, which is invaluable to applications ranging from archiving cultural e-heritage to consumer product design. Past methods for capturing the spectral reflectance of scenes has proven successful in relighting but they all share a common assumption. All the methods do not consider the effects of fluorescence despite fluorescence being found in many everyday objects. In this paper, we describe the very different ways that reflectance and fluorescence interact with illuminants and show the need to explicitly consider fluorescence in the relighting problem. We then propose a robust method based on well established theories of reflectance and fluorescence for imaging each of these components. Finally, we show that we can relight real scenes of reflective-fluorescent surfaces with much higher accuracy in comparison to only considering the reflective component.

****

Is There a Procedural Logic to Architecture?

Julien Weissenberg, Hayko Riemenschneider, Mukta Prasad, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 185-192

Urban models are key to navigation, architecture and entertainment. Apart from visualizing fac,ades, a number of tedious tasks remain largely manual (e.g. compression, generating new fac,ade designs and structurally comparing fac,ades for classification, retrieval and clustering). We propose a novel procedural modelling method to automatically learn a grammar from a set of fac,ades, generate new f ac,ade instances and compare fac,ades. To deal with the difficulty of grammatical inference, we reformulate the problem. Instead of inferring a compromising, onesize-fits-all, single grammar for all tasks, we infer a model whose successive refinements are production rules tailored for each task. We demonstrate our automatic rule inference on datasets of two different architectural styles. Our method supercedes manual expert work and cuts the time required to build a procedural model of a fac,ade from several days to a few milliseconds.

****

Motion Estimation for Self-Driving Cars with a Generalized Camera

Gim Hee Lee, Friedrich Faundorfer, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2746-2753

In this paper, we present a visual ego-motion estimation algorithm for a self-driving car equipped with a closeto-market multi-camera system. By modeling the multicamera system as a generalized camera and applying the non-holonomic motion constraint of a car, we show that this leads to a novel 2-point minimal solution for the generalized essential matrix where the full relative motion including metric scale can be obtained. We provide the analytical solutions for the general case with at least one inter-camera correspondence and a special case with only intra-camera correspondences. We show that up to a maximum of 6 solutions exist for both cases. We identify the existence of degeneracy when the car undergoes straight motion in the special case with only intra-camera correspondences where the scale becomes unobservable and provide a practical alternative solution. Our formulation can be efficiently implemented within RANSAC for robust estimation. We verify the validity of our assumptions on the motion model by comparing our results on a large real-world dataset collected by a car equipped with 4 cameras with minimal overlapping field-of-views against the GPS/INS ground truth.

****

Histograms of Sparse Codes for Object Detection

Xiaofeng Ren, Deva Ramanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3246-3253

Object detection has seen huge progress in recent years, much thanks to the heavily-engineered Histograms of Oriented Gradients (HOG) features. Can we go beyond gradients and do better than HOG? We provide an affirmative answer by proposing and investigating a sparse representation for object detection, Histograms of Sparse Codes (HSC). We compute sparse codes with dictionaries learned from data using K-SVD, and aggregate per-pixel sparse codes to form local histograms. We intentionally keep true to the sliding window framework (with mixtures and parts) and only change the underlying features. To keep training (and testing) efficient, we apply dimension reduction by computing SVD on learned models, and adopt supervised training where latent positions of roots and parts are given externally e.g. from a HOG-based detector. By learning and using local representations that are much more expressive than gradients, we demonstrate large improvements over the state of the art on the PASCAL benchmark for both rootonly and part-based models.

****

Video Object Segmentation through Spatially Accurate and Temporally Dense Extraction of Primary Object Regions

Dong Zhang, Omar Javed, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 628-635

In this paper, we propose a novel approach to extract primary object segments in videos in the 'object proposal' domain. The extracted primary object regions are then used to build object models for optimized video segmentation. The proposed approach has several contributions: First, a novel layered Directed Acyclic Graph (DAG) based framework is presented for detection and segmentation of the primary object in video. We exploit the fact that, in general, objects are spatially cohesive and characterized by locally smooth motion trajectories, to extract the primary object from the set of all available proposals based on motion, appearance and predicted-shape similarity across frames. Second, the DAG is initialized with an enhanced object proposal set where motion based proposal predictions (from adjacent frames) are used to expand the set of object proposals for a part icular frame. Last, the paper presents a motion scoring function for selection of object proposals that emphasizes high optical flow gradients at proposal boundaries to discriminate between moving objects and the background. The proposed approach is evaluated using several challenging benchmark videos and it outperforms both unsupervised and supervised state-of-the-art methods.

****

Capturing Complex Spatio-temporal Relations among Facial Muscles for Facial Expression Recognition

Ziheng Wang, Shangfei Wang, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3422-3429

Spatial-temporal relations among facial muscles carry crucial information about facial expressions yet have not been thoroughly exploited. One contributing factor for this is the limited ability of the current dynamic models in capturing complex spatial and temporal relations. Existing dynamic models can only capture simple local temporal relations among sequential events, or lack the ability for incorporating uncertainties. To overcome these limitations and take full advantage of the spatio-temporal information, we propose to model the facial expression as a complex activity that consists of temporally overlapping or sequential primitive facial events. We further propose the Interval Temporal Bayesian Network to capture these complex temporal relations among primitive facial events for facial expression modeling and recognition. Experimental results on benchmark databases demonstrate the feasibility of the proposed approach in recognizing facial expressions based purely on spatio-temporal relations among facial muscles, as well as its advantage over the existing methods.

****

Bayesian Depth-from-Defocus with Shading Constraints

Chen Li, Shuochen Su, Yasuyuki Matsushita, Kun Zhou, Stephen Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 217-224

We present a method that enhances the performance of depth-from-defocus (DFD) through the use of shading information. DFD suffers from important limitations namely coarse shape reconstruction and poor accuracy on textureless surfaces that can be overcome with the help of shading. We integrate both forms of data within a Bayesian framework that capitalizes on their relative strengths. Shading data, however, is challenging to recover accurately from surfaces that contain texture. To address this issue, we propose an iterative technique that utilizes depth information to improve shading estimation, which in turn is used to elevate depth estimation in the presence of textures. With this approach, we demonstrate improvements over existing DFD techniques, as well as effective shape reconstruction of textureless surfaces.

****

Sparse Output Coding for Large-Scale Visual Recognition

Bin Zhao, Eric P. Xing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3350-3357

Many vision tasks require a multi-class classifier to discriminate multiple categories, on the order of hundreds or thousands. In this paper, we propose sparse output coding, a principled way for large-scale multi-class classification, by turning high-cardinality multi-class categorization into a bit-by-bit decoding problem. Specifically, sparse output coding is composed of two steps: efficient coding matrix learning with scalability to thousands of classes, and probabilistic decoding. Empirical results on object recognition and scene classification demonstrate the effectiveness of our proposed approach.

****

Boundary Cues for 3D Object Shape Recovery

Kevin Karsch, Zicheng Liao, Jason Rock, Jonathan T. Barron, Derek Hoiem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2163-2170

Early work in computer vision considered a host of geometric cues for both shape reconstruction [11] and recognition [14]. However, since then, the vision community has focused heavily on shading cues for reconstruction [1], and moved towar ds data-driven approaches for recognition [6]. In this paper, we reconsider these perhaps overlooked "boundary" cues (such as self occlusions and folds in a surface), as well as many other established constraints for shape reconstruction. In a variety of user studies and quantitative tasks, we evaluate how well these cues inform shape reconstruction (relative to each other) in terms of both shape quality and shape recognition. Our findings suggest many new directions for future research in shape reconstruction, such as automatic boundary cue detection and relaxing assumptions in shape from shading (e.g. orthographic projection, Lambertian surfaces).

****

Image Segmentation by Cascaded Region Agglomeration

Zhile Ren, Gregory Shakhnarovich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2011-2018

We propose a hierarchical segmentation algorithm that starts with a very fine oversegmentation and gradually merges regions using a cascade of boundary classifiers. This approach allows the weights of region and boundary features to adapt to the segmentation scale at which they are applied. The stages of the cascade are trained sequentially, with asymetric loss to maximize boundary recall. On six segmentation data sets, our algorithm achieves best performance under most region-quality measures, and does it with fewer segments than the prior work. Our algorithm is also highly competitive in a dense oversegmentation (superpixel) regime under boundary-based measures.

****

Spatial Inference Machines

Roman Shapovalov, Dmitry Vetrov, Pushmeet Kohli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2985-2992

This paper addresses the problem of semantic segmentation of 3D point clouds. We extend the inference machines framework of Ross et al. by adding spatial factors that model mid-range and long-range dependencies inherent in the data. The new model is able to account for semantic spatial context. During training, our method automatically isolates and retains factors modelling spatial dependencies between variables that are relevant for achieving higher prediction accuracy. We evaluate the proposed method by using it to predict 17-category semantic segmentations on sets of stitched Kinect scans. Experimental results show that the spatial dependencies learned by our method significantly improve the accuracy of segmentation. They also show that our method outperforms the existing segmentation technique of Koppula et al.

****

Can a Fully Unconstrained Imaging Model Be Applied Effectively to Central Cameras?

Filippo Bergamasco, Andrea Albarelli, Emanuele Rodola, Andrea Torsello; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1391-1398

Traditional camera models are often the result of a compromise between the ability to account for non-linearities in the image formation model and the need for a feasible number of degrees of freedom in the estimation process. These considerations led to the definition of several ad hoc models that best adapt to different imaging devices, ranging from pinhole cameras with no radial distortion to the more complex catadioptric or polydioptric optics. In this paper we propose the use of an unconstrained model even in standard central camera settings dominated by the pinhole model, and introduce a novel calibration approach that can deal effectively with the huge number of free parameters associated with it, resulting in a higher precision calibration than what is possible with the standard pinhole model with correction for radial distortion. This effectively extends the use of general models to settings that traditionally have been ruled by parametric approaches out of practical considerations. The benefit of such an unconstrained model to quasipinhole central cameras is supported by an extensive experimental validation.

****

Learning Compact Binary Codes for Visual Tracking

Xi Li, Chunhua Shen, Anthony Dick, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2419-2426

A key problem in visual tracking is to represent the appearance of an object in a way that is robust to visual changes. To attain this robustness, increasingly complex models are used to capture appearance variations. However, such models can be difficult to maintain accurately and efficiently. In this paper, we propose a visual tracker in which objects are represented by compact and discriminative binary codes. This representation can be processed very efficiently, and is capable of effectively fusing information from multiple cues. An incremental discriminative learner is then used to construct an appearance model that optimally separates the object from its surrounds. Furthermore, we design a hypergraph propagation method to capture the contextual information on samples, which further improves the tracking accuracy. Experimental results on challenging videos demonstrate the effectiveness and robustness of the proposed tracker.

****

Efficient Maximum Appearance Search for Large-Scale Object Detection

Qiang Chen, Zheng Song, Rogerio Feris, Ankur Datta, Liangliang Cao, Zhongyang Huang, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3190-3197

In recent years, efficiency of large-scale object detection has arisen as an important topic due to the exponential growth in the size of benchmark object detection datasets. Most current object detection methods focus on improving accuracy of large-scale object detection with efficiency being an afterthought. In this paper, we present the Efficient Maximum Appearance Search (EMAS) model which is an order of magnitude faster than the existing state-of-the-art large-scale object detection approaches, while maintaining comparable accuracy. Our EMAS model consists of representing an image as an ensemble of densely sampled feature points with the proposed Pointwise Fisher Vector encoding method, so that the learnt discriminative scoring function can be applied locally. Consequently, the object detection problem is transformed into searching an image sub-area for maximum local appearance probability, thereby making EMAS an order of magnitude faster than the traditional detection methods. In addition, the proposed model is also suitable for incorporating global context at a negligible extra computational cost. EMAS can also incorporate fusion of multiple features, which greatly improves its performance in detecting multiple object categories. Our experiments show that the proposed algorithm can perform detection of 1000 object classes in less than one minute per image on the Image Net ILSVRC2012 dataset and for 107 object classes in less than 5 seconds per image for the SUN09 dataset using a single CPU.

****

A New Perspective on Uncalibrated Photometric Stereo

Thoma Papadhimitri, Paolo Favaro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1474-1481

We investigate the problem of reconstructing normals, albedo and lights of Lambertian surfaces in uncalibrated photometric stereo under the perspective projection model. Our analysis is based on establishing the integrability constraint. In the orthographic projection case, it is well-known that when such constraint is imposed, a solution can be identified only up to 3 parameters, the so-called generalized bas-relief (GBR) ambiguity. We show that in the perspective projection case the solution is unique. We also propose a closed-form solution which is simple, efficient and robust. We test our algorithm on synthetic data and publicly available real data. Our quantitative tests show that our method outperforms all prior work of uncalibrated photometric stereo under orthographic projection.

****

A Joint Model for 2D and 3D Pose Estimation from a Single Image

Edgar Simo-Serra, Ariadna Quattoni, Carme Torras, Francesc Moreno-Noguer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3634-3641

We introduce a novel approach to automatically recover 3D human pose from a sing le image. Most previous work follows a pipelined approach: initially, a set of 2D features such as edges, joints or silhouettes are detected in the image, and then these observations are used to infer the 3D pose. Solving these two problems separately may lead to erroneous 3D poses when the feature detector has performed poorly. In this paper, we address this issue by jointly solving both the 2D detection and the 3D inference problems. For this purpose, we propose a Bayesian framework that integrates a generative model based on latent variables and discriminative 2D part detectors based on HOGs, and perform inference using evolutionary algorithms. Real experimentation demonstrates competitive results, and the ability of our methodology to provide accurate 2D and 3D pose estimations even when the 2D detectors are inaccurate.

****

A Statistical Model for Recreational Trails in Aerial Images

Andrew Predoehl, Scott Morris, Kobus Barnard; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 337-344

We present a statistical model of aerial images of recreational trails, and a method to infer trail routes in such images. We learn a set of textons describing the images, and use them to divide the image into super-pixels represented by their texton. We then learn, for each texton, the frequency of generating on-trail and off-trail pixels, and the direction of trail through on-trail pixels. From these, we derive an image likelihood function. We combine that with a prior model of trail length and smoothness, yielding a posterior distribution for trails, given an image. We search for good values of this posterior using a novel stochastic variation of Dijkstra's algorithm. Our experiments, on trail images and groundtruth collected in the western continental USA, show substantial improvement over those of the previous best trail-finding method.

****

Learning Video Saliency from Human Gaze Using Candidate Selection

Dmitry Rudoy, Dan B. Goldman, Eli Shechtman, Lihi Zelnik-Manor; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1147-1154

During recent years remarkable progress has been made in visual saliency modeling. Our interest is in video saliency. Since videos are fundamentally different from still images, they are viewed differently by human observers. For example, the time each video frame is observed is a fraction of a second, while a still image can be viewed leisurely. Therefore, video saliency estimation methods should differ substantially from image saliency methods. In this paper we propose a novel method for video saliency estimation, which is inspired by the way people watch videos. We explicitly model the continuity of the video by predicting the saliency map of a given frame, conditioned on the map from the previous frame. Furthermore, accuracy and computation speed are improved by restricting the salient locations to a carefully selected candidate set. We validate our method using two gaze-tracked video datasets and show we outperform the state-of-the-art.

****

Designing Category-Level Attributes for Discriminative Visual Recognition

Felix X. Yu, Liangliang Cao, Rogerio S. Feris, John R. Smith, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 771-778

Attribute-based representation has shown great promises for visual recognition due to its intuitive interpretation and cross-category generalization property. However, human efforts are usually involved in the attribute designing process, making the representation costly to obtain. In this paper, we propose a novel formulation to automatically design discriminative "category-level attributes", which can be efficiently encoded by a compact category-attribute matrix. The formulation allows us to achieve intuitive and critical design criteria (category-separability, learnability) in a principled way. The designed attributes can be used for tasks of cross-category knowledge transfer, achieving superior performance over well-known attribute dataset Animals with Attributes (AwA) and a large-scale ILSVRC2010 dataset (1.2M images). This approach also leads to state-ofthe-art performance on the zero-shot learning task on AwA.

****

Dense Segmentation-Aware Descriptors

Eduard Trulls, Iasonas Kokkinos, Alberto Sanfeliu, Francesc Moreno-Noguer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2890-2897

In this work we exploit segmentation to construct appearance descriptors that can robustly deal with occlusion and background changes. For this, we downplay measurements coming from areas that are unlikely to belong to the same region as the descriptor's center, as suggested by soft segmentation masks. Our treatment is applicable to any image point, i.e. dense, and its computational overhead is in the order of a few seconds. We integrate this idea with Dense SIFT, and also with Dense Scale and Rotation Invariant Descriptors (SID), delivering descriptors that are densely computable, invariant to scaling and rotation, and robust to background changes. We apply our approach to standard benchmarks on large displacement motion estimation using SIFT-flow and widebaseline stereo, systematically demonstrating that the introduction of segmentation yields clear improvements.

****

Modeling Mutual Visibility Relationship in Pedestrian Detection

Wanli Ouyang, Xingyu Zeng, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3222-3229

Detecting pedestrians in cluttered scenes is a challenging problem in computer vision. The difficulty is added when several pedestrians overlap in images and occlude each other. We observe, however, that the occlusion/visibility statuses of overlapping pedestrians provide useful mutual relationship for visibility estimation the visibility estimation of one pedestrian facilitates the visibility estimation of another. In this paper, we propose a mutual visibility deep model that jointly estimates the visibility statuses of overlapping pedestrians. The visibility relationship among pedestrians is learned from the deep model for recognizing co-existing pedestrians. Experimental results show that the mutual visibility deep model effectively improves the pedestrian detection results. Compared with existing image-based pedestrian detection approaches, our approach has the lowest average miss rate on the CaltechTrain dataset, the Caltech-Test dataset and the ETH dataset. Including mutual visibility leads to 4% 8% improvements on multiple benchmark datasets.

****

Discriminatively Trained And-Or Tree Models for Object Detection

Xi Song, Tianfu Wu, Yunde Jia, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3278-3285

This paper presents a method of learning reconfigurable And-Or Tree (AOT) models discriminatively from weakly annotated data for object detection. To explore the appearance and geometry space of latent structures effectively, we first quantize the image lattice using an overcomplete set of shape primitives, and then organize them into a directed acyclic And-Or Graph (AOG) by exploiting their compositional relations. We allow overlaps between child nodes when combining them into a parent node, which is equivalent to introducing an appearance Or-node implicitly for the overlapped portion. The learning of an AOT model consists of three components: (i) Unsupervised sub-category learning (i.e., branches of an object Or-node) with the latent structures in AOG being integrated out. (ii) Weaklysupervised part configuration learning (i.e., seeking the globally optimal parse trees in AOG for each sub-category). To search the globally optimal parse tree in AOG efficiently, we propose a dynamic programming (DP) algorithm. (iii) Joint appearance and structural parameters training under latent structural SVM framework. In experiments, our method is tested on PASCAL VOC 2007 and 2010 detection benchmarks of 20 object classes and outperforms comparable state-of-the-art methods.

****

Intrinsic Scene Properties from a Single RGB-D Image

Jonathan T. Barron, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 17-24

In this paper we extend the "shape, illumination and reflectance from shading" (

SIRFS) model [3, 4], which recovers intrinsic scene properties from a single image. Though SIRFS performs well on images of segmented objects, it performs poorly on images of natural scenes, which contain occlusion and spatially-varying illumination. We therefore present Scene-SIRFS, a generalization of SIRFS in which we have a mixture of shapes and a mixture of illuminations, and those mixture components are embedded in a "soft" segmentation of the input image. We additionally use the noisy depth maps provided by RGB-D sensors (in this case, the Kinect) to improve shape estimation. Our model takes as input a single RGB-D image and produces as output an improved depth map, a set of surface normals, a reflectance image, a shading image, and a spatially varying model of illumination. The output of our model can be used for graphics applications, or for any application involving RGB-D images.

****

Cross-View Image Geolocalization

Tsung-Yi Lin, Serge Belongie, James Hays; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 891-898

The recent availability of large amounts of geotagged imagery has inspired a number of data driven solutions to the image geolocalization problem. Existing approaches predict the location of a query image by matching it to a database of georeferenced photographs. While there are many geotagged images available on photo sharing and street view sites, most are clustered around landmarks and urban areas. The vast majority of the Earth's land area has no ground level reference photos available, which limits the applicability of all existing image geolocalization methods. On the other hand, there is no shortage of visual and geographic data that densely covers the Earth we examine overhead imagery and land cover survey data but the relationship between this data and ground level query photographs is complex. In this paper, we introduce a cross-view feature translation approach to greatly extend the reach of image geolocalization methods. We can often localize a query even if it has no corresponding groundlevel images in the database. A key idea is to learn the relationship between ground level appearance and overhead appearance and land cover attributes from sparsely available geotagged ground-level images. We perform experiments over a 1600 km d-region containing a variety of scenes and land cover types. For each query, our algorithm produces a probability density over the region of interest.

****

Learning Cross-Domain Information Transfer for Location Recognition and Clustering

Raghuraman Gopalan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 731-738

Estimating geographic location from images is a challenging problem that is receiving recent attention. In contrast to many existing methods that primarily model discriminative information corresponding to different locations, we propose joint learning of information that images across locations share and vary upon. Starting with generative and discriminative subspaces pertaining to domains, which are obtained by a hierarchical grouping of images from adjacent locations, we present a top-down approach that first models cross-domain information transfer by utilizing the geometry of these subspaces, and then encodes the model results onto individual images to infer their location. We report competitive results for location recognition and clustering on two public datasets, im2GPS and San Francisco, and empirically validate the utility of various design choices involved in the approach.

****

Statistical Textural Distinctiveness for Salient Region Detection in Natural Images

Christian Scharfenberger, Alexander Wong, Khalil Fergani, John S. Zelek, David A. Clausi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 979-986

A novel statistical textural distinctiveness approach for robustly detecting salient regions in natural images is proposed. Rotational-invariant neighborhood-based textural representations are extracted and used to learn a set of representa tive texture atoms for defining a sparse texture model for the image. Based on the learnt sparse texture model, a weighted graphical model is constructed to characterize the statistical textural distinctiveness between all representative texture atom pairs. Finally, the saliency of each pixel in the image is computed based on the probability of occurrence of the representative texture atoms, their respective statistical textural distinctiveness based on the constructed graphical model, and general visual attentive constraints. Experimental results using a public natural image dataset and a variety of performance evaluation metrics show that the proposed approach provides interesting and promising results when compared to existing saliency detection methods.

****

Robust Multi-resolution Pedestrian Detection in Traffic Scenes

Junjie Yan, Xucong Zhang, Zhen Lei, Shengcai Liao, Stan Z. Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3033-3040

The serious performance decline with decreasing resolution is the major bottleneck for current pedestrian detection techniques [14, 23]. In this paper, we take pedestrian detection in different resolutions as different but related problems, and propose a Multi-Task model to jointly consider their commonness and differences. The model contains resolution aware transformations to map pedestrians in different resolutions to a common space, where a shared detector is constructed to distinguish pedestrians from background. For model learning, we present a coordinate descent procedure to learn the resolution aware transformations and deformable part model (DPM) based detector iteratively. In traffic scenes, there are many false positives located around vehicles, therefore, we further build a context model to suppress them according to the pedestrian-vehicle relationship. The context model can be learned automatically even when the vehicle annotations are not available. Our method reduces the mean miss rate to 60% for pedestrians taller than 30 pixels on the Caltech Pedestrian Benchmark, which noticeably outperforms previous state-of-the-art (71%).

****

Hypergraphs for Joint Multi-view Reconstruction and Multi-object Tracking

Martin Hofmann, Daniel Wolf, Gerhard Rigoll; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3650-3657

We generalize the network flow formulation for multiobject tracking to multi-camera setups. In the past, reconstruction of multi-camera data was done as a separate extension. In this work, we present a combined maximum a posteriori (MAP) formulation, which jointly models multicamera reconstruction as well as global temporal data association. A flow graph is constructed, which tracks objects in 3D world space. The multi-camera reconstruction can be efficiently incorporated as additional constraints on the flow graph without making the graph unnecessarily large. The final graph is efficiently solved using binary linear programming. On the PETS 2009 dataset we achieve results that significantly exceed the current state of the art.

****

Recognizing Activities via Bag of Words for Attribute Dynamics

Weixin Li, Qian Yu, Harpreet Sawhney, Nuno Vasconcelos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2587-2594

In this work, we propose a novel video representation for activity recognition that models video dynamics with attributes of activities. A video sequence is decomposed into short-term segments, which are characterized by the dynamics of their attributes. These segments are modeled by a dictionary of attribute dynamics templates, which are implemented by a recently introduced generative model, the binary dynamic system (BDS). We propose methods for learning a dictionary of BDSs from a training corpus, and for quantizing attribute sequences extracted from videos into these BDS codewords. This procedure produces a representation of the video as a histogram of BDS codewords, which is denoted the bag-of-words for attribute dynamics (BoWAD). An extensive experimental evaluation reveals that this representation outperforms other state-of-the-art approaches in temporal struct ure modeling for complex activity recognition.

****

Towards Fast and Accurate Segmentation

Camillo J. Taylor; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1916-1922

In this paper we explore approaches to accelerating segmentation and edge detection algorithms based on the gPb framework. The paper characterizes the performance of a simple but effective edge detection scheme which can be computed rapidly and offers performance that is competitive with the pB detector. The paper also describes an approach for computing a reduced order normalized cut that captures the essential features of the original problem but can be computed in less than half a second on a standard computing platform.

****

Fast, Accurate Detection of 100,000 Object Classes on a Single Machine

Thomas Dean, Mark A. Ruzon, Mark Segal, Jonathon Shlens, Sudheendra Vijayanarasimhan, Jay Yagnik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1814-1821

Many object detection systems are constrained by the time required to convolve a target image with a bank of filters that code for different aspects of an object's appearance, such as the presence of component parts. We exploit locality-sensitive hashing to replace the dot-product kernel operator in the convolution with a fixed number of hash-table probes that effectively sample all of the filter responses in time independent of the size of the filter bank. To show the effectiveness of the technique, we apply it to evaluate 100,000 deformable-part models requiring over a million (part) filters on multiple scales of a target image in less than 20 seconds using a single multi-core processor with 20GB of RAM. This represents a speed-up of approximately 20,000 times-four orders of magnitude-when compared with performing the convolutions explicitly on the same hardware. While mean average precision over the full set of 100,000 object classes is around 0.16 due in large part to the challenges in gathering training data and collecting ground truth for so many classes, we achieve a mAP of at least 0.20 on a third of the classes and 0.30 or better on about 20% of the classes.

****

Robust Object Co-detection

Xin Guo, Dong Liu, Brendan Jou, Mojun Zhu, Anni Cai, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3206-3213

Object co-detection aims at simultaneous detection of objects of the same category from a pool of related images by exploiting consistent visual patterns present in candidate objects in the images. The related image set may contain a mixture of annotated objects and candidate objects generated by automatic detectors. Co-detection differs from the conventional object detection paradigm in which detection over each test image is determined one-by-one independently without taking advantage of common patterns in the data pool. In this paper, we propose a novel, robust approach to dramatically enhance co-detection by extracting a shared low-rank representation of the object instances in multiple feature spaces. The idea is analogous to that of the well-known Robust PCA [28], but has not been explored in object co-detection so far. The representation is based on a linear reconstruction over the entire data set and the low-rank approach enables effective removal of noisy and outlier samples. The extracted low-rank representation can be used to detect the target objects by spectral clustering. Extensive experiments over diverse benchmark datasets demonstrate consistent and significant performance gains of the proposed method over the state-of-the-art object codetection method and the generic object detection methods without co-detection formulations.

****

Supervised Kernel Descriptors for Visual Recognition

Peng Wang, Jingdong Wang, Gang Zeng, Weiwei Xu, Hongbin Zha, Shipeng Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2858-2865

In visual recognition tasks, the design of low level image feature representation is fundamental. The advent of local patch features from pixel attributes such as SIFT and LBP, has precipitated dramatic progresses. Recently, a kernel view of these features, called kernel descriptors (KDES) [1], generalizes the feature design in an unsupervised fashion and yields impressive results. In this paper, we present a supervised framework to embed the image level label information into the design of patch level kernel descriptors, which we call supervised kernel descriptors (SKDES). Specifically, we adopt the broadly applied bag-of-words (BOW) image classification pipeline and a large margin criterion to learn the lowlevel patch representation, which makes the patch features much more compact and achieve better discriminative ability than KDES. With this method, we achieve competitive results over several public datasets comparing with stateof-the-art methods.

****

Shape from Silhouette Probability Maps: Reconstruction of Thin Objects in the Presence of Silhouette Extraction and Calibration Error
Amy Tabb; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 161-168

This paper considers the problem of reconstructing the shape of thin, texture-less objects such as leafless trees when there is noise or deterministic error in the silhouette extraction step or there are small errors in camera calibration. Traditional intersection-based techniques such as the visual hull are not robust to error because they penalize false negative and false positive error unequally. We provide a voxel-based formalism that penalizes false negative and positive error equally, by casting the reconstruction problem as a pseudo-Boolean minimization problem, where voxels are the variables of a pseudo-Boolean function and are labeled occupied or empty. Since the pseudo-Boolean minimization problem is NP-Hard for nonsubmodular functions, we developed an algorithm for an approximate solution using local minimum search. Our algorithm treats input binary probability maps (in other words, silhouettes) or continuously-valued probability maps identically, and places no constraints on camera placement. The algorithm was tested on three different leafless trees and one metal object where the number of voxels is 54.4 million (voxel sides measure 3.6 mm). Results show that our approach reconstructs the complicated branching structure of thin, texture-less objects in the presence of error where intersection-based approaches currently fail. 1

****

Measures and Meta-Measures for the Supervised Evaluation of Image Segmentation
Jordi Pont-Tuset, Ferran Marques; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2131-2138

This paper tackles the supervised evaluation of image segmentation algorithms. First, it surveys and structures the measures used to compare the segmentation results with a ground truth database; and proposes a new measure: the precision-recall for objects and parts. To compare the goodness of these measures, it defines three quantitative meta-measures involving six state of the art segmentation methods. The meta-measures consist in assuming some plausible hypotheses about the results and assessing how well each measure reflects these hypotheses. As a conclusion, this paper proposes the precision-recall curves for boundaries and for objects-and-parts as the tool of choice for the supervised evaluation of image segmentation. We make the datasets and code of all the measures publicly available.

****

A Fast Approximate AIB Algorithm for Distributional Word Clustering
Lei Wang, Jianjia Zhang, Luping Zhou, Wanqing Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 556-563

Distributional word clustering merges the words having similar probability distributions to attain reliable parameter estimation, compact classification models and even better classification performance. Agglomerative Information Bottleneck (AIB) is one of the typical word clustering algorithms and has been applied to both traditional text classification and recent image recognition. Although enjo

ying theoretical elegance, AIB has one main issue on its computational efficiency, especially when clustering a large number of words. Different from existing solutions to this issue, we analyze the characteristics of its objective function -the loss of mutual information, and show that by merely using the ratio of word-class joint probabilities of each word, good candidate word pairs for merging can be easily identified. Based on this finding, we propose a fast approximate AIB algorithm and show that it can significantly improve the computational efficiency of AIB while well maintaining or even slightly increasing its classification performance. Experimental study on both text and image classification benchmark data sets shows that our algorithm can achieve more than 100 times speedup on large real data sets over the state-of-the-art method.
****
Separable Dictionary Learning
Simon Hawe, Matthias Seibert, Martin Kleinsteuber; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 438-445
Many techniques in computer vision, machine learning, and statistics rely on the fact that a signal of interest admits a sparse representation over some dictionary. Dictionaries are either available analytically, or can be learned from a suitable training set. While analytic dictionaries permit to capture the global structure of a signal and allow a fast implementation, learned dictionaries often perform better in applications as they are more adapted to the considered class of signals. In imagery, unfortunately, the numerical burden for (i) learning a dictionary and for (ii) employing the dictionary for reconstruction tasks only allows to deal with relatively small image patches that only capture local image information. The approach presented in this paper aims at overcoming these drawbacks by allowing a separable structure on the dictionary throughout the learning process. On the one hand, this permits larger patch-sizes for the learning phase, on the other hand, the dictionary is applied efficiently in reconstruction tasks. The learning procedure is based on optimizing over a product of spheres which updates the dictionary as a whole, thus enforces basic dictionary properties such as mutual coherence explicitly during the learning procedure. In the special case where no separable structure is enforced, our method competes with state-of-the-art dictionary learning methods like K-SVD.
****
Representing and Discovering Adversarial Team Behaviors Using Player Roles
Patrick Lucey, Alina Bialkowski, Peter Carr, Stuart Morgan, Iain Matthews, Yaser Sheikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2706-2713
In this paper, we describe a method to represent and discover adversarial group behavior in a continuous domain. In comparison to other types of behavior, adversarial behavior is heavily structured as the location of a player (or agent) is dependent both on their teammates and adversaries, in addition to the tactics or strategies of the team. We present a method which can exploit this relationship through the use of a spatiotemporal basis model. As players constantly change roles during a match, we show that employing a "role-based" representation instead of one based on player "identity" can best exploit the playing structure. As vision-based systems currently do not provide perfect detection/tracking (e.g. missed or false detections), we show that our compact representation can effectively "denoise" erroneous detections as well as enabling temporal analysis, which was previously prohibitive due to the dimensionality of the signal. To evaluate our approach, we used a fully instrumented field-hockey pitch with 8 fixed highdefinition (HD) cameras and evaluated our approach on approximately 200,000 frames of data from a state-of-theart real-time player detector and compare it to manually labelled data.
****
Object-Centric Anomaly Detection by Attribute-Based Reasoning
Babak Saleh, Ali Farhadi, Ahmed Elgammal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 787-794
When describing images, humans tend not to talk about the obvious, but rather mention what they find interesting. We argue that abnormalities and deviations fro

m typicalities are among the most important components that form what is worth mentioning. In this paper we introduce the abnormality detection as a recognition problem and show how to model typicalities and, consequently, meaningful deviations from prototypical properties of categories. Our model can recognize abnormalities and report the main reasons of any recognized abnormality. We also show that abnormality predictions can help image categorization. We introduce the abnormality detection dataset and show interesting results on how to reason about abnormalities.

****

Cartesian K-Means
Mohammad Norouzi, David J. Fleet; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3017-3024

A fundamental limitation of quantization techniques like the k-means clustering algorithm is the storage and runtime cost associated with the large numbers of clusters required to keep quantization errors small and model fidelity high. We develop new models with a compositional parameterization of cluster centers, so representational capacity increases super-linearly in the number of parameters. This allows one to effectively quantize data using billions or trillions of centers. We formulate two such models, Orthogonal k-means and Cartesian k-means. They are closely related to one another, to k-means, to methods for binary hash function optimization like ITQ [5], and to Product Quantization for vector quantization [7]. The models are tested on largescale ANN retrieval tasks (1M GIST, 1B SIFT features), and on codebook learning for object recognition (CIFAR-10).

****

Optimal Geometric Fitting under the Truncated L2-Norm
Erik Ask, Olof Enqvist, Fredrik Kahl; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1722-1729

This paper is concerned with model fitting in the presence of noise and outliers. Previously it has been shown that the number of outliers can be minimized with polynomial complexity in the number of measurements. This paper improves on these results in two ways. First, it is shown that for a large class of problems, the statistically more desirable truncated L 2 -norm can be optimized with the same complexity. Then, with the same methodology, it is shown how to transform multi-model fitting into a purely combinatorial problem--with worst-case complexity that is polynomial in the number of measurements, though exponential in the number of models. We apply our framework to a series of hard registration and stitching problems demonstrating that the approach is not only of theoretical interest. It gives a practical method for simultaneously dealing with measurement noise and large amounts of outliers for fitting problems with lowdimensional models.

****

Pedestrian Detection with Unsupervised Multi-stage Feature Learning
Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, Yann Lecun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3626-3633

Pedestrian detection is a problem of considerable practical interest. Adding to the list of successful applications of deep learning methods to vision, we report state-of-theart and competitive results on all major pedestrian datasets with a convolutional network model. The model uses a few new twists, such as multi-stage features, connections that skip layers to integrate global shape information with local distinctive motif information, and an unsupervised method based on convolutional sparse coding to pre-train the filters at each stage.

****

Integrating Grammar and Segmentation for Human Pose Estimation
Brandon Rothrock, Seyoung Park, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3214-3221

In this paper we present a compositional and-or graph grammar model for human pose estimation. Our model has three distinguishing features: (i) large appearance differences between people are handled compositionally by allowing parts or collections of parts to be substituted with alternative variants, (ii) each variant is a sub-model that can define its own articulated geometry and context-sensiti ve compatibility with neighboring part variants, and (iii) background region segmentation is incorporated into the part appearance models to better estimate the contrast of a part region from its surroundings, and improve resilience to background clutter. The resulting integrated framework is trained discriminatively in a max-margin framework using an efficient and exact inference algorithm. We present experimental evaluation of our model on two popular datasets, and show performance improvements over the state-of-art on both benchmarks.

****

Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images
Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, Andrew Fitzgibbon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2930-2937
We address the problem of inferring the pose of an RGB-D camera relative to a known 3D scene, given only a single acquired image. Our approach employs a regression forest that is capable of inferring an estimate of each pixel's correspondence to 3D points in the scene's world coordinate frame. The forest uses only simple depth and RGB pixel comparison features, and does not require the computation of feature descriptors. The forest is trained to be capable of predicting correspondences at any pixel, so no interest point detectors are required. The camera pose is inferred using a robust optimization scheme. This starts with an initial set of hypothesized camera poses, constructed by applying the forest at a small fraction of image pixels. Preemptive RANSAC then iterates sampling more pixels at which to evaluate the forest, counting inliers, and refining the hypothesized poses. We evaluate on several varied scenes captured with an RGB-D camera and observe that the proposed technique achieves highly accurate relocalization and substantially out-performs two state of the art baselines.

****

Joint Detection, Tracking and Mapping by Semantic Bundle Adjustment
Nicola Fioraio, Luigi Di Stefano; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1538-1545
In this paper we propose a novel Semantic Bundle Adjustment framework whereby known rigid stationary objects are detected while tracking the camera and mapping the environment. The system builds on established tracking and mapping techniques to exploit incremental 3D reconstruction in order to validate hypotheses on the presence and pose of sought objects. Then, detected objects are explicitly taken into account for a global semantic optimization of both camera and object poses. Thus, unlike all systems proposed so far, our approach allows for solving jointly the detection and SLAM problems, so as to achieve object detection together with improved SLAM accuracy.

****

Robust Region Grouping via Internal Patch Statistics
Xiaobai Liu, Liang Lin, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1931-1938
In this work, we present an efficient multi-scale low-rank representation for image segmentation. Our method begins with partitioning the input images into a set of superpixels, followed by seeking the optimal superpixel-pair affinity matrix, both of which are performed at multiple scales of the input images. Since low-level superpixel features are usually corrupted by image noises, we propose to infer the low-rank refined affinity matrix. The inference is guided by two observations on natural images. First, looking into a single image, local small-size image patterns tend to recur frequently within the same semantic region, but may not appear in semantically different regions. We call this internal image statistics as replication prior, and quantitatively justify it on real image databases. Second, the affinity matrices at different scales should be consistently solved, which leads to the cross-scale consistency constraint. We formulate these two purposes with one unified formulation and develop an efficient optimization procedure. Our experiments demonstrate the presented method can substantially improve segmentation accuracy.

****

Boundary Detection Benchmarking: Beyond F-Measures

Xiaodi Hou, Alan Yuille, Christof Koch; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2123-2130

For an ill-posed problem like boundary detection, human labeled datasets play a critical role. Compared with the active research on finding a better boundary detector to refresh the performance record, there is surprisingly little discussion on the boundary detection benchmark itself. The goal of this paper is to identify the potential pitfalls of today's most popular boundary benchmark, BSDS 300. In the paper, we first introduce a psychophysical experiment to show that many of the "weak" boundary labels are unreliable and may contaminate the benchmark. Then we analyze the computation of f-measure and point out that the current benchmarking protocol encourages an algorithm to bias towards those problematic "weak" boundary labels. With this evidence, we focus on a new problem of detecting strong boundaries as one alternative. Finally, we assess the performances of 9 major algorithms on different ways of utilizing the dataset, suggesting new directions for improvements.

****

Manhattan Junction Catalogue for Spatial Reasoning of Indoor Scenes

Srikumar Ramalingam, Jaishanker K. Pillai, Arpit Jain, Yuichi Taguchi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3065-3072

Junctions are strong cues for understanding the geometry of a scene. In this paper, we consider the problem of detecting junctions and using them for recovering the spatial layout of an indoor scene. Junction detection has always been challenging due to missing and spurious lines. We work in a constrained Manhattan world setting where the junctions are formed by only line segments along the three principal orthogonal directions. Junctions can be classified into several categories based on the number and orientations of the incident line segments. We provide a simple and efficient voting scheme to detect and classify these junctions in real images. Indoor scenes are typically modeled as cuboids and we formulate the problem of the cuboid layout estimation as an inference problem in a conditional random field. Our formulation allows the incorporation of junction features and the training is done using structured prediction techniques. We outperform other single view geometry estimation methods on standard datasets.

****

What Makes a Patch Distinct?

Ran Margolin, Ayellet Tal, Lihi Zelnik-Manor; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1139-1146

What makes an object salient? Most previous work assert that distinctness is the dominating factor. The difference between the various algorithms is in the way they compute distinctness. Some focus on the patterns, others on the colors, and several add high-level cues and priors. We propose a simple, yet powerful, algorithm that integrates these three factors. Our key contribution is a novel and fast approach to compute pattern distinctness. We rely on the inner statistics of the patches in the image for identifying unique patterns. We provide an extensive evaluation and show that our approach outperforms all state-of-the-art methods on the five most commonly-used datasets.

****

Detection- and Trajectory-Level Exclusion in Multiple Object Tracking

Anton Milan, Konrad Schindler, Stefan Roth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3682-3689

When tracking multiple targets in crowded scenarios, modeling mutual exclusion between distinct targets becomes important at two levels: (1) in data association, each target observation should support at most one trajectory and each trajectory should be assigned at most one observation per frame; (2) in trajectory estimation, two trajectories should remain spatially separated at all times to avoid collisions. Yet, existing trackers often sidestep these important constraints. We address this using a mixed discrete-continuous conditional random field (CRF) that explicitly models both types of constraints: Exclusion between conflicting observations with supermodular pairwise terms, and exclusion between trajectories by generalizing global label costs to suppress the co-occurrence of incompati ble labels (trajectories). We develop an expansion move-based MAP estimation scheme that handles both non-submodular constraints and pairwise global label costs. Furthermore, we perform a statistical analysis of ground-truth trajectories to derive appropriate CRF potentials for modeling data fidelity, target dynamics, and inter-target occlusion.

****

Real-Time Model-Based Rigid Object Pose Estimation and Tracking Combining Dense and Sparse Visual Cues

Karl Pauwels, Leonardo Rubio, Javier Diaz, Eduardo Ros; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2347-2354

We propose a novel model-based method for estimating and tracking the six-degrees-of-freedom (6DOF) pose of rigid objects of arbitrary shapes in real-time. By combining dense motion and stereo cues with sparse keypoint correspondences, and by feeding back information from the model to the cue extraction level, the method is both highly accurate and robust to noise and occlusions. A tight integration of the graphical and computational capability of Graphics Processing Units (GPUs) results in pose updates at framerates exceeding 60 Hz. Since a benchmark dataset that enables the evaluation of stereo-vision-based pose estimators in complex scenarios is currently missing in the literature, we have introduced a novel synthetic benchmark dataset with varying objects, background motion, noise and occlusions. Using this dataset and a novel evaluation methodology, we show that the proposed method greatly outperforms state-of-the-art methods. Finally, we demonstrate excellent performance on challenging real-world sequences involving object manipulation.

****

Unnatural L0 Sparse Representation for Natural Image Deblurring

Li Xu, Shicheng Zheng, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1107-1114

We show in this paper that the success of previous maximum a posterior (MAP) based blur removal methods partly stems from their respective intermediate steps, which implicitly or explicitly create an unnatural representation containing salient image structures. We propose a generalized and mathematically sound L 0 sparse expression, together with a new effective method, for motion deblurring. Our system does not require extra filtering during optimization and demonstrates fast energy decreasing, making a small number of iterations enough for convergence. It also provides a unified framework for both uniform and non-uniform motion deblurring. We extensively validate our method and show comparison with other approaches with respect to convergence speed, running time, and result quality.

****

Decoding Children's Social Behavior

J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim, H. Rao, J. Kim, L. Lo Presti, J. Zhang, D. Lantsman, J. Bidwell, Z. Ye; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3414-3421

We introduce a new problem domain for activity recognition: the analysis of children's social and communicative behaviors based on video and audio data. We specifically target interactions between children aged 1-2 years and an adult. Such interactions arise naturally in the diagnosis and treatment of developmental disorders such as autism. We introduce a new publicly-available dataset containing over 160 sessions of a 3-5 minute child-adult interaction. In each session, the adult examiner followed a semistructured play interaction protocol which was designed to elicit a broad range of social behaviors. We identify the key technical challenges in analyzing these behaviors, and describe methods for decoding the interactions. We present experimental results that demonstrate the potential of the dataset to drive interesting research questions, and show preliminary results for multi-modal activity recognition.

****

Finding Group Interactions in Social Clutter

Ruonan Li, Parker Porfilio, Todd Zickler; Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2722-2729

We consider the problem of finding distinctive social interactions involving groups of agents embedded in larger social gatherings. Given a pre-defined gallery of short exemplar interaction videos, and a long input video of a large gathering (with approximately-tracked agents), we identify within the gathering small sub-groups of agents exhibiting social interactions that resemble those in the exemplars. The participants of each detected group interaction are localized in space; the extent of their interaction is localized in time; and when the gallery of exemplars is annotated with group-interaction categories, each detected interaction is classified into one of the pre-defined categories. Our approach represents group behaviors by dichotomous collections of descriptors for (a) individual actions, and (b) pairwise interactions; and it includes efficient algorithms for optimally distinguishing participants from by-standers in every temporal unit and for temporally localizing the extent of the group interaction. Most importantly, the method is generic and can be applied whenever numerous interacting agents can be approximately tracked over time. We evaluate the approach using three different video collections, two that involve humans and one that involves mice.

****

Least Soft-Threshold Squares Tracking

Dong Wang, Huchuan Lu, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2371-2378

In this paper, we propose a generative tracking method based on a novel robust linear regression algorithm. In contrast to existing methods, the proposed Least Soft-thresold Squares (LSS) algorithm models the error term with the Gaussian-Laplacian distribution, which can be solved efficiently. Based on maximum joint likelihood of parameters, we derive a LSS distance to measure the difference between an observation sample and the dictionary. Compared with the distance derived from ordinary least squares methods, the proposed metric is more effective in dealing with outliers. In addition, we present an update scheme to capture the appearance change of the tracked target and ensure that the model is properly updated. Experimental results on several challenging image sequences demonstrate that the proposed tracker achieves more favorable performance than the state-of-the-art methods.

****

Online Robust Dictionary Learning

Cewu Lu, Jiaping Shi, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 415-422

Online dictionary learning is particularly useful for processing large-scale and dynamic data in computer vision. It, however, faces the major difficulty to incorporate robust functions, rather than the square data fitting term, to handle outliers in training data. In this paper, we propose a new online framework enabling the use of ersparse data fitting term in robust dictionary learning, notably enhancing the usability and practicality of this important technique. Extensive experiments have been carried out to validate our new framework.

****

Learning the Change for Automatic Image Cropping

Jianzhou Yan, Stephen Lin, Sing Bing Kang, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 971-978

Image cropping is a common operation used to improve the visual quality of photographs. In this paper, we present an automatic cropping technique that accounts for the two primary considerations of people when they crop: removal of distracting content, and enhancement of overall composition. Our approach utilizes a large training set consisting of photos before and after cropping by expert photographers to learn how to evaluate these two factors in a crop. In contrast to the many methods that exist for general assessment of image quality, ours specifically examines differences between the original and cropped photo in solving for the crop parameters. To this end, several novel image features are proposed to model the changes in image content and composition when a crop is applied. Our experiments demonstrate improvements of our method over recent cropping algorithms on a broad range of images.

****

Multi-resolution Shape Analysis via Non-Euclidean Wavelets: Applications to Mesh Segmentation and Surface Alignment Problems

Won Hwa Kim, Moo K. Chung, Vikas Singh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2139-2146

The analysis of 3-D shape meshes is a fundamental problem in computer vision, graphics, and medical imaging. Frequently, the needs of the application require that our analysis take a multi-resolution view of the shape's local and global topology, and that the solution is consistent across multiple scales. Unfortunately, the preferred mathematical construct which offers this behavior in classical image/signal processing, Wavelets, is no longer applicable in this general setting (data with non-uniform topology). In particular, the traditional definition does not allow writing out an expansion for graphs that do not correspond to the uniformly sampled lattice (e.g., images). In this paper, we adapt recent results in harmonic analysis, to derive NonEuclidean Wavelets based algorithms for a range of shape analysis problems in vision and medical imaging. We show how descriptors derived from the dual domain representation offer native multi-resolution behavior for characterizing local/global topology around vertices. With only minor modifications, the framework yields a method for extracting interest/key points from shapes, a surprisingly simple algorithm for 3-D shape segmentation (competitive with state of the art), and a method for surface alignment (without landmarks). We give an extensive set of comparison results on a large shape segmentation benchmark and derive a uniqueness theorem for the surface alignment problem.

****

Stochastic Deconvolution

James Gregson, Felix Heide, Matthias B. Hullin, Mushfiqur Rouf, Wolfgang Heidrich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1043-1050

We present a novel stochastic framework for non-blind deconvolution based on point samples obtained from random walks. Unlike previous methods that must be tailored to specific regularization strategies, the new Stochastic Deconvolution method allows arbitrary priors, including nonconvex and data-dependent regularizers, to be introduced and tested with little effort. Stochastic Deconvolution is straightforward to implement, produces state-of-the-art results and directly leads to a natural boundary condition for image boundaries and saturated pixels.

****

Nonparametric Scene Parsing with Adaptive Feature Relevance and Semantic Context

Gautam Singh, Jana Kosecka; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3151-3157

This paper presents a nonparametric approach to semantic parsing using small patches and simple gradient, color and location features. We learn the relevance of individual feature channels at test time using a locally adaptive distance metric. To further improve the accuracy of the nonparametric approach, we examine the importance of the retrieval set used to compute the nearest neighbours using a novel semantic descriptor to retrieve better candidates. The approach is validated by experiments on several datasets used for semantic parsing demonstrating the superiority of the method compared to the state of art approaches.

****

Social Role Discovery in Human Events

Vignesh Ramanathan, Bangpeng Yao, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2475-2482

We deal with the problem of recognizing social roles played by people in an event. Social roles are governed by human interactions, and form a fundamental component of human event description. We focus on a weakly supervised setting, where we are provided different videos belonging to an event class, without training role labels. Since social roles are described by the interaction between people in an event, we propose a Conditional Random Field to model the inter-role interactions, along with person specific social descriptors. We develop tractable variational inference to simultaneously infer model weights, as well as role assignment to all people in the videos. We also present a novel YouTube social roles da taset with ground truth role annotations, and introduce annotations on a subset of videos from the TRECVID-MED11 [1] event kits for evaluation purposes. The performance of the model is compared against different baseline methods on these datasets.

****

Learning to Estimate and Remove Non-uniform Image Blur

Florent Couzinie-Devy, Jian Sun, Karteek Alahari, Jean Ponce; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1075-1082

This paper addresses the problem of restoring images subjected to unknown and spatially varying blur caused by defocus or linear (say, horizontal) motion. The estimation of the global (non-uniform) image blur is cast as a multilabel energy minimization problem. The energy is the sum of unary terms corresponding to learned local blur estimators, and binary ones corresponding to blur smoothness. Its global minimum is found using Ishikawa's method by exploiting the natural order of discretized blur values for linear motions and defocus. Once the blur has been estimated, the image is restored using a robust (non-uniform) deblurring algorithm based on sparse regularization with global image statistics. The proposed algorithm outputs both a segmentation of the image into uniform-blur layers and an estimate of the corresponding sharp image. We present qualitative results on real images, and use synthetic data to quantitatively compare our approach to the publicly available implementation of Chakrabarti et al. [5].

****

Scene Parsing by Integrating Function, Geometry and Appearance Models

Yibiao Zhao, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3119-3126

Indoor functional objects exhibit large view and appearance variations, thus are difficult to be recognized by the traditional appearance-based classification paradigm. In this paper, we present an algorithm to parse indoor images based on two observations: i) The functionality is the most essential property to define an indoor object, e.g. "a chair to sit on"; ii) The geometry (3D shape) of an object is designed to serve its function. We formulate the nature of the object function into a stochastic grammar model. This model characterizes a joint distribution over the function-geometryappearance (FGA) hierarchy. The hierarchical structure includes a scene category, functional groups, functional objects, functional parts and 3D geometric shapes. We use a simulated annealing MCMC algorithm to find the maximum a posteriori (MAP) solution, i.e. a parse tree. We design four data-driven steps to accelerate the search in the FGA space: i) group the line segments into 3D primitive shapes, ii) assign functional labels to these 3D primitive shapes, iii) fill in missing objects/parts according to the functional labels, and iv) synthesize 2D segmentation maps and verify the current parse tree by the Metropolis-Hastings acceptance probability. The experimental results on several challenging indoor datasets demonstrate the proposed approach not only significantly widens the scope of indoor scene parsing algorithm from the segmentation and the 3D recovery to the functional object recognition, but also yields improved overall performance.

****

Efficient Detector Adaptation for Object Detection in a Video

Pramod Sharma, Ram Nevatia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3254-3261

In this work, we present a novel and efficient detector adaptation method which improves the performance of an offline trained classifier (baseline classifier) by adapting it to new test datasets. We address two critical aspects of adaptation methods: generalizability and computational efficiency. We propose an adaptation method, which can be applied to various baseline classifiers and is computationally efficient also. For a given test video, we collect online samples in an unsupervised manner and train a random fern adaptive classifier . The adaptive classifier improves precision of the baseline classifier by validating the obtained detection responses from baseline classifier as correct detections or false alarms. Experiments demonstrate generalizability, computational efficiency and ef fectiveness of our method, as we compare our method with state of the art approaches for the problem of human detection and show good performance with high computational efficiency on two different baseline classifiers.

****

Evaluation of Color STIPs for Human Action Recognition

Ivo Everts, Jan C. van Gemert, Theo Gevers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2850-2857

This paper is concerned with recognizing realistic human actions in videos based on spatio-temporal interest points (STIPs). Existing STIP-based action recognition approaches operate on intensity representations of the image data. Because of this, these approaches are sensitive to disturbing photometric phenomena such as highlights and shadows. Moreover, valuable information is neglected by discarding chromaticity from the photometric representation. These issues are addressed by Color STIPs. Color STIPs are multi-channel reformulations of existing intensity-based STIP detectors and descriptors, for which we consider a number of chromatic representations derived from the opponent color space. This enhanced modeling of appearance improves the quality of subsequent STIP detection and description. Color STIPs are shown to substantially outperform their intensity-based counterparts on the challenging UCF sports, UCF11 and UCF50 action recognition benchmarks. Moreover, the results show that color STIPs are currently the single best low-level feature choice for STIP-based approaches to human action recognition.

****

A Global Approach for the Detection of Vanishing Points and Mutually Orthogonal Vanishing Directions

Michel Antunes, Joao P. Barreto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1336-1343

This article presents a new global approach for detecting vanishing points and groups of mutually orthogonal vanishing directions using lines detected in images of man-made environments. These two multi-model fitting problems are respectively cast as Uncapacited Facility Location (UFL) and Hierarchical Facility Location (HFL) instances that are efficiently solved using a message passing inference algorithm. We also propose new functions for measuring the consistency between an edge and a putative vanishing point, and for computing the vanishing point defined by a subset of edges. Extensive experiments in both synthetic and real images show that our algorithms outperform the state-ofthe-art methods while keeping computation tractable. In addition, we show for the first time results in simultaneously detecting multiple Manhattan-world configurations that can either share one vanishing direction (Atlanta world) or be completely independent.

****

Poselet Conditioned Pictorial Structures

Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 588-595

In this paper we consider the challenging problem of articulated human pose estimation in still images. We observe that despite high variability of the body articulations, human motions and activities often simultaneously constrain the positions of multiple body parts. Modelling such higher order part dependencies seemingly comes at a cost of more expensive inference, which resulted in their limited use in state-of-the-art methods. In this paper we propose a model that incorporates higher order part dependencies while remaining efficient. We achieve this by defining a conditional model in which all body parts are connected a-priori, but which becomes a tractable tree-structured pictorial structures model once the image observations are available. In order to derive a set of conditioning variables we rely on the poselet-based features that have been shown to be effective for people detection but have so far found limited application for articulated human pose estimation. We demonstrate the effectiveness of our approach on three publicly available pose estimation benchmarks improving or being on-par with state of the art in each case.

****

Enriching Texture Analysis with Semantic Data

Tim Matthews, Mark S. Nixon, Mahesan Niranjan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1248-1255

We argue for the importance of explicit semantic modelling in human-centred texture analysis tasks such as retrieval, annotation, synthesis, and zero-shot learning. To this end, low-level attributes are selected and used to define a semantic space for texture. 319 texture classes varying in illumination and rotation are positioned within this semantic space using a pairwise relative comparison procedure. Low-level visual features used by existing texture descriptors are then assessed in terms of their correspondence to the semantic space. Textures with strong presence of attributes connoting randomness and complexity are shown to be poorly modelled by existing descriptors. In a retrieval experiment semantic descriptors are shown to outperform visual descriptors. Semantic modelling of texture is thus shown to provide considerable value in both feature selection and in analysis tasks.

****

Scene Text Recognition Using Part-Based Tree-Structured Character Detection

Cunzhao Shi, Chunheng Wang, Baihua Xiao, Yang Zhang, Song Gao, Zhong Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2961-2968

Scene text recognition has inspired great interests from the computer vision community in recent years. In this paper, we propose a novel scene text recognition method using part-based tree-structured character detection. Different from conventional multi-scale sliding window character detection strategy, which does not make use of the character-specific structure information, we use part-based tree-structure to model each type of character so as to detect and recognize the characters at the same time. While for word recognition, we build a Conditional Random Field model on the potential character locations to incorporate the detection scores, spatial constraints and linguistic knowledge into one framework. The final word recognition result is obtained by minimizing the cost function defined on the random field. Experimental results on a range of challenging public datasets (ICDAR 2003, ICDAR 2011, SVT) demonstrate that the proposed method outperforms stateof-the-art methods significantly both for character detection and word recognition.

****

MODEC: Multimodal Decomposable Models for Human Pose Estimation

Ben Sapp, Ben Taskar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3674-3681

We propose a multimodal, decomposable model for articulated human pose estimation in monocular images. A typical approach to this problem is to use a linear structured model, which struggles to capture the wide range of appearance present in realistic, unconstrained images. In this paper, we instead propose a model of human pose that explicitly captures a variety of pose modes. Unlike other multimodal models, our approach includes both global and local pose cues and uses a convex objective and joint training for mode selection and pose estimation. We also employ a cascaded mode selection step which controls the trade-off between speed and accuracy, yielding a 5x speedup in inference and learning. Our model outperforms state-of-theart approaches across the accuracy-speed trade-off curve for several pose datasets. This includes our newly-collected dataset of people in movies, FLIC, which contains an order of magnitude more labeled data for training and testing than existing datasets. The new dataset and code are available online. 1

****

Multi-task Sparse Learning with Beta Process Prior for Action Recognition

Chunfeng Yuan, Weiming Hu, Guodong Tian, Shuang Yang, Haoran Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 423-429

In this paper, we formulate human action recognition as a novel Multi-Task Sparse Learning(MTSL) framework which aims to construct a test sample with multiple features from as few bases as possible. Learning the sparse representation under each feature modality is considered as a single task in MTSL. Since the tasks are generated from multiple features associated with the same visual input, they are not independent but inter-related. We introduce a Beta process(BP) prior to the hierarchical MTSL model, which efficiently learns a compact dictionary and infers the sparse structure shared across all the tasks. The MTSL model enforces the robustness in coefficient estimation compared with performing each task independently. Besides, the sparseness is achieved via the Beta process formulation rather than the computationally expensive l 1 norm penalty. In terms of non-informative gamma hyper-priors, the sparsity level is totally decided by the data. Finally, the learning problem is solved by Gibbs sampling inference which estimates the full posterior on the model parameters. Experimental results on the KTH and UCF sports datasets demonstrate the effectiveness of the proposed MTSL approach for action recognition.

****

Decoding, Calibration and Rectification for Lenselet-Based Plenoptic Cameras

Donald G. Dansereau, Oscar Pizarro, Stefan B. Williams; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1027-1034

Plenoptic cameras are gaining attention for their unique light gathering and post-capture processing capabilities. We describe a decoding, calibration and rectification procedure for lenselet-based plenoptic cameras appropriate for a range of computer vision applications. We derive a novel physically based 4D intrinsic matrix relating each recorded pixel to its corresponding ray in 3D space. We further propose a radial distortion model and a practical objective function based on ray reprojection. Our 15-parameter camera model is of much lower dimensionality than camera array models, and more closely represents the physics of lenselet-based cameras. Results include calibration of a commercially available camera using three calibration grid sizes over five datasets. Typical RMS ray reprojection errors are 0.0628, 0.105 and 0.363 mm for 3.61, 7.22 and 35.1 mm calibration grids, respectively. Rectification examples include calibration targets and real-world imagery.

****

Hierarchical Video Representation with Trajectory Binary Partition Tree

Guillem Palou, Philippe Salembier; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2099-2106

As early stage of video processing, we introduce an iterative trajectory merging algorithm that produces a regionbased and hierarchical representation of the video sequence, called the Trajectory Binary Partition Tree (BPT). From this representation, many analysis and graph cut techniques can be used to extract partitions or objects that are useful in the context of specific applications. In order to define trajectories and to create a precise merging algorithm, color and motion cues have to be used. Both types of informations are very useful to characterize objects but present strong differences of behavior in the spatial and the temporal dimensions. On the one hand, scenes and objects are rich in their spatial color distributions, but these distributions are rather stable over time. Object motion, on the other hand, presents simple structures and low spatial variability but may change from frame to frame. The proposed algorithm takes into account this key difference and relies on different models and associated metrics to deal with color and motion information. We show that the proposed algorithm outperforms existing hierarchical video segmentation algorithms and provides more stable and precise regions.

****

Cloud Motion as a Calibration Cue

Nathan Jacobs, Mohammad T. Islam, Scott Workman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1344-1351

We propose cloud motion as a natural scene cue that enables geometric calibration of static outdoor cameras. This work introduces several new methods that use observations of an outdoor scene over days and weeks to estimate radial distortion, focal length and geo-orientation. Cloud-based cues provide strong constraints and are an important alternative to methods that require specific forms of stat ic scene geometry or clear sky conditions. Our method makes simple assumptions about cloud motion and builds upon previous work on motion-based and line-based calibration. We show results on real scenes that highlight the effectiveness of our proposed methods.

****

FasT-Match: Fast Affine Template Matching

Simon Korman, Daniel Reichman, Gilad Tsur, Shai Avidan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2331-2338

Fast-Match is a fast algorithm for approximate template matching under 2D affine transformations that minimizes the Sum-of-Absolute-Differences (SAD) error measure. There is a huge number of transformations to consider but we prove that they can be sampled using a density that depends on the smoothness of the image. For each potential transformation, we approximate the SAD error using a sublinear algorithm that randomly examines only a small number of pixels. We further accelerate the algorithm using a branch-and-bound scheme. As images are known to be piecewise smooth, the result is a practical affine template matching algorithm with approximation guarantees, that takes a few seconds to run on a standard machine. We perform several experiments on three different datasets, and report very good results. To the best of our knowledge, this is the first template matching algorithm which is guaranteed to handle arbitrary 2D affine transformations.

****

Dense Non-rigid Point-Matching Using Random Projections

Raffay Hamid, Dennis Decoste, Chih-Jen Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2914-2921

We present a robust and efficient technique for matching dense sets of points undergoing non-rigid spatial transformations. Our main intuition is that the subset of points that can be matched with high confidence should be used to guide the matching procedure for the rest. We propose a novel algorithm that incorporates these high-confidence matches as a spatial prior to learn a discriminative subspace that simultaneously encodes both the feature similarity as well as their spatial arrangement. Conventional subspace learning usually requires spectral decomposition of the pair-wise distance matrix across the point-sets, which can become inefficient even for moderately sized problems. To this end, we propose the use of random projections for approximate subspace learning, which can provide significant time improvements at the cost of minimal precision loss. This efficiency gain allows us to iteratively find and remove high-confidence matches from the point sets, resulting in high recall. To show the effectiveness of our approach, we present a systematic set of experiments and results for the problem of dense non-rigid image-feature matching.

****

Large Displacement Optical Flow from Nearest Neighbor Fields

Zhuoyuan Chen, Hailin Jin, Zhe Lin, Scott Cohen, Ying Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2443-2450

We present an optical flow algorithm for large displacement motions. Most existing optical flow methods use the standard coarse-to-fine framework to deal with large displacement motions which has intrinsic limitations. Instead, we formulate the motion estimation problem as a motion segmentation problem. We use approximate nearest neighbor fields to compute an initial motion field and use a robust algorithm to compute a set of similarity transformations as the motion candidates for segmentation. To account for deviations from similarity transformations, we add local deformations in the segmentation process. We also observe that small objects can be better recovered using translations as the motion candidates. We fuse the motion results obtained under similarity transformations and under translations together before a final refinement. Experimental validation shows that our method can successfully handle large displacement motions. Although we particularly focus on large displacement motions in this work, we make no sacrifice in terms of overall performance. In particular, our method ranks at the top of the Middlebury benchmark.

****

Hallucinated Humans as the Hidden Context for Labeling 3D Scenes

Yun Jiang, Hema Koppula, Ashutosh Saxena; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2993-3000

For scene understanding, one popular approach has been to model the object-object relationships. In this paper, we hypothesize that such relationships are only an artifact of certain hidden factors, such as humans. For example, the objects, monitor and keyboard, are strongly spatially correlated only because a human types on the keyboard while watching the monitor. Our goal is to learn this hidden human context (i.e., the human-object relationships), and also use it as a cue for labeling the scenes. We present Infinite Factored Topic Model (IFTM), where we consider a scene as being generated from two types of topics: human configurations and human-object relationships. This enables our algorithm to hallucinate the possible configurations of the humans in the scene parsimoniously. Given only a dataset of scenes containing objects but not humans, we show that our algorithm can recover the human object relationships. We then test our algorithm on the task of attribute and object labeling in 3D scenes and show consistent improvements over the state-of-the-art.

****

Discrete MRF Inference of Marginal Densities for Non-uniformly Discretized Variable Space

Masaki Saito, Takayuki Okatani, Koichiro Deguchi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 57-64

This paper is concerned with the inference of marginal densities based on MRF models. The optimization algorithms for continuous variables are only applicable to a limited number of problems, whereas those for discrete variables are versatile. Thus, it is quite common to convert the continuous variables into discrete ones for the problems that ideally should be solved in the continuous domain, such as stereo matching and optical flow estimation. In this paper, we show a novel formulation for this continuous-discrete conversion. The key idea is to estimate the marginal densities in the continuous domain by approximating them with mixtures of rectangular densities. Based on this formulation, we derive a mean field (MF) algorithm and a belief propagation (BP) algorithm. These algorithms can correctly handle the case where the variable space is discretized in a non-uniform manner. By intentionally using such a non-uniform discretization, a higher balance between computational efficiency and accuracy of marginal density estimates could be achieved. We present a method for actually doing this, which dynamically discretizes the variable space in a coarse-to-fine manner in the course of the computation. Experimental results show the effectiveness of our approach.

****

Efficient Large-Scale Structured Learning

Steve Branson, Oscar Beijbom, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1806-1813

We introduce an algorithm, SVM-IS, for structured SVM learning that is computationally scalable to very large datasets and complex structural representations. We show that structured learning is at least as fast-and often much faster-than methods based on binary classification for problems such as deformable part models, object detection, and multiclass classification, while achieving accuracies that are at least as good. Our method allows problem-specific structural knowledge to exploited for faster optimization by integrating with a user-defined importance sampling function. We demonstrate fast train times on two challenging large scale datasets for two very different problems: ImageNet for multiclass classification and CUB-200-2011 for deformable part model training. Our method is shown to be 10-50 times faster than SVMstruct for cost-sensitive multiclass classification while being about as fast as the fastest 1-vs-all methods for multiclass classification. For deformable part model training, it is shown to be 50-1000 times faster than methods based on SVMstruct, mining hard negatives, and Pegasos-style stochastic gradient descent. Source code of our method is publicly available.

****

Relative Volume Constraints for Single View 3D Reconstruction

Eno Toppe, Claudia Nieuwenhuis, Daniel Cremers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 177-184

We introduce the concept of relative volume constraints in order to account for insufficient information in the reconstruction of 3D objects from a single image. The key idea is to formulate a variational reconstruction approach with shape priors in form of relative depth profiles or volume ratios relating object parts. Such shape priors can easily be derived either from a user sketch or from the object's shading profile in the image. They can handle textured or shadowed object regions by propagating information. We propose a convex relaxation of the constrained optimization problem which can be solved optimally in a few seconds on graphics hardware. In contrast to existing single view reconstruction algorithms, the proposed algorithm provides substantially more flexibility to recover shape details such as self-occlusions, dents and holes, which are not visible in the object silhouette.

****

Uncalibrated Photometric Stereo for Unknown Isotropic Reflectances

Feng Lu, Yasuyuki Matsushita, Imari Sato, Takahiro Okabe, Yoichi Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1490-1497

We propose an uncalibrated photometric stereo method that works with general and unknown isotropic reflectances. Our method uses a pixel intensity profile, which is a sequence of radiance intensities recorded at a pixel across multi-illuminance images. We show that for general isotropic materials, the geodesic distance between intensity profiles is linearly related to the angular difference of their surface normals, and that the intensity distribution of an intensity profile conveys information about the reflectance properties, when the intensity profile is obtained under uniformly distributed directional lightings. Based on these observations, we show that surface normals can be estimated up to a convex/concave ambiguity. A solution method based on matrix decomposition with missing data is developed for a reliable estimation. Quantitative and qualitative evaluations of our method are performed using both synthetic and real-world scenes.

****

Image Understanding from Experts' Eyes by Modeling Perceptual Skill of Diagnostic Reasoning Processes

Rui Li, Pengcheng Shi, Anne R. Haake; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2187-2194

Eliciting and representing experts' remarkable perceptual capability of locating, identifying and categorizing objects in images specific to their domains of expertise will benefit image understanding in terms of transferring human domain knowledge and perceptual expertise into image-based computational procedures. In this paper, we present a hierarchical probabilistic framework to summarize the stereotypical and idiosyncratic eye movement patterns shared within 11 board-certified dermatologists while they are examining and diagnosing medical images. Each inferred eye movement pattern characterizes the similar temporal and spatial properties of its corresponding segments of the experts' eye movement sequences. We further discover a subset of distinctive eye movement patterns which are commonly exhibited across multiple images. Based on the combinations of the exhibitions of these eye movement patterns, we are able to categorize the images from the perspective of experts' viewing strategies. In each category, images share similar lesion distributions and configurations. The performance of our approach shows that modeling physicians' diagnostic viewing behaviors informs about medical images' understanding to correct diagnosis.

****

Robust Discriminative Response Map Fitting with Constrained Local Models

Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, Maja Pantic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3444-3451

We present a novel discriminative regression based approach for the Constrained Local Models (CLMs) framework, referred to as the Discriminative Response Map Fitting (DRMF) method, which shows impressive performance in the generic face fitt ing scenario. The motivation behind this approach is that, unlike the holistic texture based features used in the discriminative AAM approaches, the response map can be represented by a small set of parameters and these parameters can be very efficiently used for reconstructing unseen response maps. Furthermore, we show that by adopting very simple off-the-shelf regression techniques, it is possible to learn robust functions from response maps to the shape parameters updates. The experiments, conducted on Multi-PIE, XM2VTS and LFPW database, show that the proposed DRMF method outperforms stateof-the-art algorithms for the task of generic face fitting. Moreover, the DRMF method is computationally very efficient and is real-time capable. The current MATLAB implementation takes 1 second per image. To facilitate future comparisons, we release the MATLAB code acand the pretrained models for research purposes.

****

Unconstrained Monocular 3D Human Pose Estimation by Action Detection and Cross-Modality Regression Forest

Tsz-Ho Yu, Tae-Kyun Kim, Roberto Cipolla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3642-3649

This work addresses the challenging problem of unconstrained 3D human pose estimation (HPE) from a novel perspective. Existing approaches struggle to operate in realistic applications, mainly due to their scene-dependent priors, such as background segmentation and multi-camera network, which restrict their use in unconstrained environments. We therfore present a framework which applies action detection and 2D pose estimation techniques to infer 3D poses in an unconstrained video. Action detection offers spatiotemporal priors to 3D human pose estimation by both recognising and localising actions in space-time. Instead of holistic features, e.g. silhouettes, we leverage the flexibility of deformable part model to detect 2D body parts as a feature to estimate 3D poses. A new unconstrained pose dataset has been collected to justify the feasibility of our method, which demonstrated promising results, significantly outperforming the relevant state-of-the-arts.

****

Revisiting Depth Layers from Occlusions

Adarsh Kowdle, Andrew Gallagher, Tsuhan Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2091-2098

In this work, we consider images of a scene with a moving object captured by a static camera. As the object (human or otherwise) moves about the scene, it reveals pairwise depth-ordering or occlusion cues. The goal of this work is to use these sparse occlusion cues along with monocular depth occlusion cues to densely segment the scene into depth layers. We cast the problem of depth-layer segmentation as a discrete labeling problem on a spatiotemporal Markov Random Field (MRF) that uses the motion occlusion cues along with monocular cues and a smooth motion prior for the moving object. We quantitatively show that depth ordering produced by the proposed combination of the depth cues from object motion and monocular occlusion cues are superior to using either feature independently, and using a na??ve combination of the features.

****

Efficient Object Detection and Segmentation for Fine-Grained Recognition

Anelia Angelova, Shenghuo Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 811-818

We propose a detection and segmentation algorithm for the purposes of fine-grained recognition. The algorithm first detects low-level regions that could potentially belong to the object and then performs a full-object segmentation through propagation. Apart from segmenting the object, we can also 'zoom in' on the object, i.e. center it, normalize it for scale, and thus discount the effects of the background. We then show that combining this with a state-of-the-art classification algorithm leads to significant improvements in performance especially for datasets which are considered particularly hard for recognition, e.g. birds species. The proposed algorithm is much more efficient than other known methods in similar scenarios [4, 21]. Our method is also simpler and we apply it here to different classes of objects, e.g. birds, flowers, cats and dogs. We tested the algor ithm on a number of benchmark datasets for fine-grained categorization. It outperforms all the known state-of-the-art methods on these datasets, sometimes by as much as 11%. It improves the performance of our baseline algorithm by 3-4%, consistently on all datasets. We also observed more than a 4% improvement in the recognition performance on a challenging largescale flower dataset, containing 578 species of flowers and 250,000 images.

****

Fast Rigid Motion Segmentation via Incrementally-Complex Local Models

Fernando Flores-Mangas, Allan D. Jepson; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2259-2266

The problem of rigid motion segmentation of trajectory data under orthography has been long solved for nondegenerate motions in the absence of noise. But because real trajectory data often incorporates noise, outliers, motion degeneracies and motion dependencies, recently proposed motion segmentation methods resort to non-trivial representations to achieve state of the art segmentation accuracies, at the expense of a large computational cost. This paper proposes a method that dramatically reduces this cost (by two or three orders of magnitude) with minimal accuracy loss (from 98.8% achieved by the state of the art, to 96.2% achieved by our method on the standard Hopkins 155 dataset). Computational efficiency comes from the use of a simple but powerful representation of motion that explicitly incorporates mechanisms to deal with noise, outliers and motion degeneracies. Subsets of motion models with the best balance between prediction accuracy and model complexity are chosen from a pool of candidates, which are then used for segmentation.

****

A Lazy Man's Approach to Benchmarking: Semisupervised Classifier Evaluation and Recalibration

Peter Welinder, Max Welling, Pietro Perona; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3262-3269

How many labeled examples are needed to estimate a classifier's performance on a new dataset? We study the case where data is plentiful, but labels are expensive. We show that by making a few reasonable assumptions on the structure of the data, it is possible to estimate performance curves, with confidence bounds, using a small number of ground truth labels. Our approach, which we call Semisupervised Performance Evaluation (SPE), is based on a generative model for the classifier's confidence scores. In addition to estimating the performance of classifiers on new datasets, SPE can be used to recalibrate a classifier by reestimating the class-conditional confidence distributions.

****

Motionlets: Mid-level 3D Parts for Human Motion Recognition

LiMin Wang, Yu Qiao, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2674-2681

This paper proposes motionlet, a mid-level and spatiotemporal part, for human motion recognition. Motionlet can be seen as a tight cluster in motion and appearance space, corresponding to the moving process of different body parts. We postulate three key properties of motionlet for action recognition: high motion saliency, multiple scale representation, and representative-discriminative ability. Towards this goal, we develop a data-driven approach to learn motionlets from training videos. First, we extract 3D regions with high motion saliency. Then we cluster these regions and preserve the centers as candidate templates for motionlet. Finally, we examine the representative and discriminative power of the candidates, and introduce a greedy method to select effective candidates. With motionlets, we present a mid-level representation for video, called motionlet activation vector. We conduct experiments on three datasets, KTH, HMDB51, and UCF50. The results show that the proposed methods significantly outperform state-of-the-art methods.

****

Understanding Indoor Scenes Using 3D Geometric Phrases

Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3

3-40

Visual scene understanding is a difficult problem interleaving object detection, geometric reasoning and scene classification. We present a hierarchical scene model for learning and reasoning about complex indoor scenes which is computationally tractable, can be learned from a reasonable amount of training data, and avoids oversimplification. At the core of this approach is the 3D Geometric Phrase Model which captures the semantic and geometric relationships between objects which frequently co-occur in the same 3D spatial configuration. Experiments show that this model effectively explains scene semantics, geometry and object groupings from a single image, while also improving individual object detections.

****

Intrinsic Characterization of Dynamic Surfaces

Tony Tung, Takashi Matsuyama; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 233-240

This paper presents a novel approach to characterize deformable surface using intrinsic property dynamics. 3D dynamic surfaces representing humans in motion can be obtained using multiple view stereo reconstruction methods or depth cameras. Nowadays these technologies have become capable to capture surface variations in real-time, and give details such as clothing wrinkles and deformations. Assuming repetitive patterns in the deformations, we propose to model complex surface variations using sets of linear dynamical systems (LDS) where observations across time are given by surface intrinsic properties such as local curvatures. We introduce an approach based on bags of dynamical systems, where each surface feature to be represented in the codebook is modeled by a set of LDS equipped with timing structure. Experiments are performed on datasets of real-world dynamical surfaces and show compelling results for description, classification and segmentation.

****

Detecting Changes in 3D Structure of a Scene from Multi-view Images Captured by a Vehicle-Mounted Camera

Ken Sakurada, Takayuki Okatani, Koichiro Deguchi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 137-144

This paper proposes a method for detecting temporal changes of the three-dimensional structure of an outdoor scene from its multi-view images captured at two separate times. For the images, we consider those captured by a camera mounted on a vehicle running in a city street. The method estimates scene structures probabilistically, not deterministically, and based on their estimates, it evaluates the probability of structural changes in the scene, where the inputs are the similarity of the local image patches among the multi-view images. The aim of the probabilistic treatment is to maximize the accuracy of change detection, behind which there is our conjecture that although it is difficult to estimate the scene structures deterministically, it should be easier to detect their changes. The proposed method is compared with the methods that use multi-view stereo (MVS) to reconstruct the scene structures of the two time points and then differentiate them to detect changes. The experimental results show that the proposed method outperforms such MVS-based methods.

****

Part-Based Visual Tracking with Online Latent Structural Learning

Rui Yao, Qinfeng Shi, Chunhua Shen, Yanning Zhang, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2363-2370

Despite many advances made in the area, deformable targets and partial occlusions continue to represent key problems in visual tracking. Structured learning has shown good results when applied to tracking whole targets, but applying this approach to a part-based target model is complicated by the need to model the relationships between parts, and to avoid lengthy initialisation processes. We thus propose a method which models the unknown parts using latent variables. In doing so we extend the online algorithm pegasos to the structured prediction case (i.e., predicting the location of the bounding boxes) with latent part variables. To better estimate the parts, and to avoid over-fitting caused by the extra model complexity/capacity introduced by the parts, we propose a two-stage training process, based on the primal rather than the dual form. We then show that the method outperforms the state-of-the-art (linear and non-linear kernel) trackers.

****

A Higher-Order CRF Model for Road Network Extraction

Jan D. Wegner, Javier A. Montoya-Zegarra, Konrad Schindler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1698-1705

The aim of this work is to extract the road network from aerial images. What makes the problem challenging is the complex structure of the prior: roads form a connected network of smooth, thin segments which meet at junctions and crossings. This type of a-priori knowledge is more difficult to turn into a tractable model than standard smoothness or co-occurrence assumptions. We develop a novel CRF formulation for road labeling, in which the prior is represented by higher-order cliques that connect sets of superpixels along straight line segments. These long-range cliques have asymmetric P es-potentials, which express a preference to assign all rather than just some of their constituent superpixels to the road class. Thus, the road likelihood is amplified for thin chains of superpixels, while the CRF is still amenable to optimization with graph cuts. Since the number of such cliques of arbitrary length is huge, we furthermore propose a sampling scheme which concentrates on those cliques which are most relevant for the optimization. In experiments on two different databases the model significantly improves both the per-pixel accuracy and the topological correctness of the extracted roads, and outperforms both a simple smoothness prior and heuristic rulebased road completion.

****

Fully-Connected CRFs with Non-Parametric Pairwise Potential

Neill D.F. Campbell, Kartic Subr, Jan Kautz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1658-1665

Conditional Random Fields (CRFs) are used for diverse tasks, ranging from image denoising to object recognition. For images, they are commonly defined as a graph with nodes corresponding to individual pixels and pairwise links that connect nodes to their immediate neighbors. Recent work has shown that fully-connected CRFs, where each node is connected to every other node, can be solved efficiently under the restriction that the pairwise term is a Gaussian kernel over a Euclidean feature space. In this paper, we generalize the pairwise terms to a non-linear dissimilarity measure that is not required to be a distance metric. To this end, we propose a density estimation technique to derive conditional pairwise potentials in a nonparametric manner. We then use an efficient embedding technique to estimate an approximate Euclidean feature space for these potentials, in which the pairwise term can still be expressed as a Gaussian kernel. We demonstrate that the use of non-parametric models for the pairwise interactions, conditioned on the input data, greatly increases expressive power whilst maintaining efficient inference.

****

Hierarchical Saliency Detection

Qiong Yan, Li Xu, Jianping Shi, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1155-1162

When dealing with objects with complex structures, saliency detection confronts a critical problem namely that detection accuracy could be adversely affected if salient foreground or background in an image contains small-scale high-contrast patterns. This issue is common in natural images and forms a fundamental challenge for prior methods. We tackle it from a scale point of view and propose a multi-layer approach to analyze saliency cues. The final saliency map is produced in a hierarchical model. Different from varying patch sizes or downsizing images, our scale-based region handling is by finding saliency values optimally in a tree model. Our approach improves saliency detection on many images that cannot be handled well traditionally. A new dataset is also constructed.

****

Depth Acquisition from Density Modulated Binary Patterns

Zhe Yang, Zhiwei Xiong, Yueyi Zhang, Jiao Wang, Feng Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 25-32

This paper proposes novel density modulated binary patterns for depth acquisition. Similar to Kinect, the illumination patterns do not need a projector for generation and can be emitted by infrared lasers and diffraction gratings. Our key idea is to use the density of light spots in the patterns to carry phase information. Two technical problems are addressed here. First, we propose an algorithm to design the patterns to carry more phase information without compromising the depth reconstruction from a single captured image as with Kinect. Second, since the carried phase is not strictly sinusoidal, the depth reconstructed from the phase contains a systematic error. We further propose a pixelbased phase matching algorithm to reduce the error. Experimental results show that the depth quality can be greatly improved using the phase carried by the density of light spots. Furthermore, our scheme can achieve 20 fps depth reconstruction with GPU assistance.

****

Pose from Flow and Flow from Pose

Katerina Fragkiadaki, Han Hu, Jianbo Shi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2059-2066

Human pose detectors, although successful in localising faces and torsos of people, often fail with lower arms. Motion estimation is often inaccurate under fast movements of body parts. We build a segmentation-detection algorithm that mediates the information between body parts recognition, and multi-frame motion grouping to improve both pose detection and tracking. Motion of body parts, though not accurate, is often sufficient to segment them from their backgrounds. Such segmentations are crucial for extracting hard to detect body parts out of their interior body clutter. By matching these segments to exemplars we obtain pose labeled body segments. The pose labeled segments and corresponding articulated joints are used to improve the motion flow fields by proposing kinematically constrained affine displacements on body parts. The pose-based articulated motion model is shown to handle large limb rotations and displacements. Our algorithm can detect people under rare poses, frequently missed by pose detectors, showing the benefits of jointly reasoning about pose, segmentation and motion in videos.

****

Composite Statistical Inference for Semantic Segmentation

Fuxin Li, Joao Carreira, Guy Lebanon, Cristian Sminchisescu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3302-3309

In this paper we present an inference procedure for the semantic segmentation of images. Different from many CRF approaches that rely on dependencies modeled with unary and pairwise pixel or superpixel potentials, our method is entirely based on estimates of the overlap between each of a set of mid-level object segmentation proposals and the objects present in the image. We define continuous latent variables on superpixels obtained by multiple intersections of segments, then output the optimal segments from the inferred superpixel statistics. The algorithm is capable of recombine and refine initial mid-level proposals, as well as handle multiple interacting objects, even from the same class, all in a consistent joint inference framework by maximizing the composite likelihood of the underlying statistical model using an EM algorithm. In the PASCAL VOC segmentation challenge, the proposed approach obtains high accuracy and successfully handles images of complex object interactions.

****

The Variational Structure of Disparity and Regularization of 4D Light Fields

Bastian Goldluecke, Sven Wanner; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1003-1010

Unlike traditional images which do not offer information for different directions of incident light, a light field is defined on ray space, and implicitly encodes scene geometry data in a rich structure which becomes visible on its epipolar plane images. In this work, we analyze regularization of light fields in variational frameworks and show that their variational structure is induced by dispari ty, which is in this context best understood as a vector field on epipolar plane image space. We derive differential constraints on this vector field to enable consistent disparity map regularization. Furthermore, we show how the disparity field is related to the regularization of more general vector-valued functions on the 4D ray space of the light field. This way, we derive an efficient variational framework with convex priors, which can serve as a fundament for a large class of inverse problems on ray space.

****

Gauging Association Patterns of Chromosome Territories via Chromatic Median
Hu Ding, Branislav Stojkovic, Ronald Berezney, Jinhui Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1296-1303

Computing accurate and robust organizational patterns of chromosome territories inside the cell nucleus is critical for understanding several fundamental genomic processes, such as co-regulation of gene activation, gene silencing, X chromosome inactivation, and abnormal chromosome rearrangement in cancer cells. The usage of advanced fluorescence labeling and image processing techniques has enabled researchers to investigate interactions of chromosome territories at large spatial resolution. The resulting high volume of generated data demands for high-throughput and automated image analysis methods. In this paper, we introduce a novel algorithmic tool for investigating association patterns of chromosome territories in a population of cells. Our method takes as input a set of graphs, one for each cell, containing information about spatial interaction of chromosome territories, and yields a single graph that contains essential information for the whole population and stands as its structural representative. We formulate this combinatorial problem as a semi-definite programming and present novel techniques to efficiently solve it. We validate our approach on both artificial and real biological data; the experimental results suggest that our approach yields a nearoptimal solution, and can handle large-size datasets, which are significant improvements over existing techniques.

****

Detecting Pulse from Head Motions in Video
Guha Balakrishnan, Fredo Durand, John Guttag; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3430-3437

We extract heart rate and beat lengths from videos by measuring subtle head motion caused by the Newtonian reaction to the influx of blood at each beat. Our method tracks features on the head and performs principal component analysis (PCA) to decompose their trajectories into a set of component motions. It then chooses the component that best corresponds to heartbeats based on its temporal frequency spectrum. Finally, we analyze the motion projected to this component and identify peaks of the trajectories, which correspond to heartbeats. When evaluated on 18 subjects, our approach reported heart rates nearly identical to an electrocardiogram device. Additionally we were able to capture clinically relevant information about heart rate variability.

****

Articulated Pose Estimation Using Discriminative Armlet Classifiers
Georgia Gkioxari, Pablo Arbelaez, Lubomir Bourdev, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3342-3349

We propose a novel approach for human pose estimation in real-world cluttered scenes, and focus on the challenging problem of predicting the pose of both arms for each person in the image. For this purpose, we build on the notion of poselets [4] and train highly discriminative classifiers to differentiate among arm configurations, which we call armlets. We propose a rich representation which, in addition to standard HOG features, integrates the information of strong contours, skin color and contextual cues in a principled manner. Unlike existing methods, we evaluate our approach on a large subset of images from the PASCAL VOC detection dataset, where critical visual phenomena, such as occlusion, truncation, multiple instances and clutter are the norm. Our approach outperforms Yang and Ramanan [26], the state-of-the-art technique, with an improvement from 29.0% to 37.5

% PCP accuracy on the arm keypoint prediction task, on this new pose estimation dataset.

****

Salient Object Detection: A Discriminative Regional Feature Integration Approach

Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, Shipeng Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2083-2090

Salient object detection has been attracting a lot of interest, and recently various heuristic computational models have been designed. In this paper, we regard saliency map computation as a regression problem. Our method, which is based on multi-level image segmentation, uses the supervised learning approach to map the regional feature vector to a saliency score, and finally fuses the saliency scores across multiple levels, yielding the saliency map. The contributions lie in two-fold. One is that we show our approach, which integrates the regional contrast, regional property and regional backgroundness descriptors together to form the master saliency map, is able to produce superior saliency maps to existing algorithms most of which combine saliency maps heuristically computed from different types of features. The other is that we introduce a new regional feature vector, backgroundness, to characterize the background, which can be regarded as a counterpart of the objectness descriptor [2]. The performance evaluation on several popular benchmark data sets validates that our approach outperforms existing state-of-the-arts.

****

Learning Locally-Adaptive Decision Functions for Person Verification

Zhen Li, Shiyu Chang, Feng Liang, Thomas S. Huang, Liangliang Cao, John R. Smith; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3610-3617

This paper considers the person verification problem in modern surveillance and video retrieval systems. The problem is to identify whether a pair of face or human body images is about the same person, even if the person is not seen before. Traditional methods usually look for a distance (or similarity) measure between images (e.g., by metric learning algorithms), and make decisions based on a fixed threshold. We show that this is nevertheless insufficient and sub-optimal for the verification problem. This paper proposes to learn a decision function for verification that can be viewed as a joint model of a distance metric and a locally adaptive thresholding rule. We further formulate the inference on our decision function as a second-order large-margin regularization problem, and provide an efficient algorithm in its dual from. We evaluate our algorithm on both human body verification and face verification problems. Our method outperforms not only the classical metric learning algorithm including LMNN and ITML, but also the state-of-the-art in the computer vision community.

****

BRDF Slices: Accurate Adaptive Anisotropic Appearance Acquisition

Jiri Filip, Radomir Vavra, Michal Haindl, Pavel Zid, Mikulas Krupika, Vlastimil Havran; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1468-1473

In this paper we introduce unique publicly available dense anisotropic BRDF data measurements. We use this dense data as a reference for performance evaluation of the proposed BRDF sparse angular sampling and interpolation approach. The method is based on sampling of BRDF subspaces at fixed elevations by means of several adaptively-represented, uniformly distributed, perpendicular slices. Although this proposed method requires only a sparse sampling of material, the interpolation provides a very accurate reconstruction, visually and computationally comparable to densely measured reference. Due to the simple slices measurement and method's robustness it allows for a highly accurate acquisition of BRDFs. This in comparison with standard uniform angular sampling, is considerably faster yet uses far less samples.

****

Explicit Occlusion Modeling for 3D Object Class Representations

M. Zeeshan Zia, Michael Stark, Konrad Schindler; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3326-3333

Despite the success of current state-of-the-art object class detectors, severe occlusion remains a major challenge. This is particularly true for more geometrically expressive 3D object class representations. While these representations have attracted renewed interest for precise object pose estimation, the focus has mostly been on rather clean datasets, where occlusion is not an issue. In this paper, we tackle the challenge of modeling occlusion in the context of a 3D geometric object class model that is capable of fine-grained, part-level 3D object reconstruction. Following the intuition that 3D modeling should facilitate occlusion reasoning, we design an explicit representation of likely geometric occlusion patterns. Robustness is achieved by pooling image evidence from of a set of fixed part detectors as well as a non-parametric representation of part configurations in the spirit of poselets. We confirm the potential of our method on cars in a newly collected data set of inner-city street scenes with varying levels of occlusion, and demonstrate superior performance in occlusion estimation and part localization, compared to baselines that are unaware of occlusions.

****

Tag Taxonomy Aware Dictionary Learning for Region Tagging

Jingjing Zheng, Zhuolin Jiang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 369-376

Tags of image regions are often arranged in a hierarchical taxonomy based on their semantic meanings. In this paper, using the given tag taxonomy, we propose to jointly learn multi-layer hierarchical dictionaries and corresponding linear classifiers for region tagging. Specifically, we generate a node-specific dictionary for each tag node in the taxonomy, and then concatenate the node-specific dictionaries from each level to construct a level-specific dictionary. The hierarchical semantic structure among tags is preserved in the relationship among node-dictionaries. Simultaneously, the sparse codes obtained using the levelspecific dictionaries are summed up as the final feature representation to design a linear classifier. Our approach not only makes use of sparse codes obtained from higher levels to help learn the classifiers for lower levels, but also encourages the tag nodes from lower levels that have the same parent tag node to implicitly share sparse codes obtained from higher levels. Experimental results using three benchmark datasets show that the proposed approach yields the best performance over recently proposed methods.

****

A Fast Semidefinite Approach to Solving Binary Quadratic Problems

Peng Wang, Chunhua Shen, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1312-1319

Many computer vision problems can be formulated as binary quadratic programs (BQPs). Two classic relaxation methods are widely used for solving BQPs, namely, spectral methods and semidefinite programming (SDP), each with their own advantages and disadvantages. Spectral relaxation is simple and easy to implement, but its bound is loose. Semidefinite relaxation has a tighter bound, but its computational complexity is high for large scale problems. We present a new SDP formulation for BQPs, with two desirable properties. First, it has a similar relaxation bound to conventional SDP formulations. Second, compared with conventional SDP methods, the new SDP formulation leads to a significantly more efficient and scalable dual optimization approach, which has the same degree of complexity as spectral methods. Extensive experiments on various applications including clustering, image segmentation, co-segmentation and registration demonstrate the usefulness of our SDP formulation for solving large-scale BQPs.

****

Learning without Human Scores for Blind Image Quality Assessment

Wufeng Xue, Lei Zhang, Xuanqin Mou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 995-1002

General purpose blind image quality assessment (BIQA) has been recently attracting significant attention in the fields of image processing, vision and machine learning. Stateof-the-art BIQA methods usually learn to evaluate the image quality by regression from human subjective scores of the training samples. However, t hese methods need a large number of human scored images for training, and lack an explicit explanation of how the image quality is affected by image local features. An interesting question is then: can we learn for effective BIQA without using human scored images? This paper makes a good effort to answer this question. We partition the distorted images into overlapped patches, and use a percentile pooling strategy to estimate the local quality of each patch. Then a quality-aware clustering (QAC) method is proposed to learn a set of centroids on each quality level. These centroids are then used as a codebook to infer the quality of each patch in a given image, and subsequently a perceptual quality score of the whole image can be obtained. The proposed QAC based BIQA method is simple yet effective. It not only has comparable accuracy to those methods using human scored images in learning, but also has merits such as high linearity to human perception of image quality, real-time implementation and availability of image local quality map.

****

Hollywood 3D: Recognizing Actions in 3D Natural Scenes

Simon Hadfield, Richard Bowden; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3398-3405

Action recognition in unconstrained situations is a difficult task, suffering from massive intra-class variations. It is made even more challenging when complex 3D actions are projected down to the image plane, losing a great deal of information. The recent emergence of 3D data, both in broadcast content, and commercial depth sensors, provides the possibility to overcome this issue. This paper presents a new dataset, for benchmarking action recognition algorithms in natural environments, while making use of 3D information. The dataset contains around 650 video clips, across 14 classes. In addition, two state of the art action recognition algorithms are extended to make use of the 3D data, and five new interest point detection strategies are also proposed, that extend to the 3D data. Our evaluation compares all 4 feature descriptors, using 7 different types of interest point, over a variety of threshold levels, for the Hollywood3D dataset. We make the dataset including stereo video, estimated depth maps and all code required to reproduce the benchmark results, available to the wider community.

****

3D Pictorial Structures for Multiple View Articulated Pose Estimation

Magnus Burenius, Josephine Sullivan, Stefan Carlsson; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3618-3625

We consider the problem of automatically estimating the 3D pose of humans from images, taken from multiple calibrated views. We show that it is possible and tractable to extend the pictorial structures framework, popular for 2D pose estimation, to 3D. We discuss how to use this framework to impose view, skeleton, joint angle and intersection constraints in 3D. The 3D pictorial structures are evaluated on multiple view data from a professional football game. The evaluation is focused on computational tractability, but we also demonstrate how a simple 2D part detector can be plugged into the framework.

****

Improving the Visual Comprehension of Point Sets

Sagi Katz, Ayellet Tal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 121-128

Point sets are the standard output of many 3D scanning systems and depth cameras. Presenting the set of points as is, might "hide" the prominent features of the object from which the points are sampled. Our goal is to reduce the number of points in a point set, for improving the visual comprehension from a given viewpoint. This is done by controlling the density of the reduced point set, so as to create bright regions (low density) and dark regions (high density), producing an effect of shading. This data reduction is achieved by leveraging a limitation of a solution to the classical problem of determining visibility from a viewpoint. In addition, we introduce a new dual problem, for determining visibility of a point from infinity, and show how a limitation of its solution can be leveraged in a similar way.

****

Kernel Methods on the Riemannian Manifold of Symmetric Positive Definite Matrices

Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, Mehrtash Harandi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 73-80

Symmetric Positive Definite (SPD) matrices have become popular to encode image information. Accounting for the geometry of the Riemannian manifold of SPD matrices has proven key to the success of many algorithms. However, most existing methods only approximate the true shape of the manifold locally by its tangent plane. In this paper, inspired by kernel methods, we propose to map SPD matrices to a high dimensional Hilbert space where Euclidean geometry applies. To encode the geometry of the manifold in the mapping, we introduce a family of provably positive definite kernels on the Riemannian manifold of SPD matrices. These kernels are derived from the Gaussian kernel, but exploit different metrics on the manifold. This lets us extend kernel-based algorithms developed for Euclidean spaces, such as SVM and kernel PCA, to the Riemannian manifold of SPD matrices. We demonstrate the benefits of our approach on the problems of pedestrian detection, object categorization, texture analysis, 2D motion segmentation and Diffusion Tensor Imaging (DTI) segmentation.

****

Graph Transduction Learning with Connectivity Constraints with Application to Multiple Foreground Cosegmentation

Tianyang Ma, Longin Jan Latecki; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1955-1962

The proposed approach is based on standard graph transduction, semi-supervised learning (SSL) framework. Its key novelty is the integration of global connectivity constraints into this framework. Although connectivity leads to higher order constraints and their number is an exponential, finding the most violated connectivity constraint can be done efficiently in polynomial time. Moreover, each such constraint can be represented as a linear inequality. Based on this fact, we design a cutting-plane algorithm to solve the integrated problem. It iterates between solving a convex quadratic problem of label propagation with linear inequality constraints, and finding the most violated constraint. We demonstrate the benefits of the proposed approach on a realistic and very challenging problem of cosegmentation of multiple foreground objects in photo collections in which the foreground objects are not present in all photos. The obtained results not only demonstrate performance boost induced by the connectivity constraints, but also show a significant improvement over the state-of-the-art methods.

****

A Max-Margin Riffled Independence Model for Image Tag Ranking

Tian Lan, Greg Mori; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3103-3110

We propose Max-Margin Riffled Independence Model (MMRIM), a new method for image tag ranking modeling the structured preferences among tags. The goal is to predict a ranked tag list for a given image, where tags are ordered by their importance or relevance to the image content. Our model integrates the max-margin formalism with riffled independence factorizations proposed in [10], which naturally allows for structured learning and efficient ranking. Experimental results on the SUN Attribute and LabelMe datasets demonstrate the superior performance of the proposed model compared with baseline tag ranking methods. We also apply the predicted rank list of tags to several higher-level computer vision applications in image understanding and retrieval, and demonstrate that MMRIM significantly improves the accuracy of these applications.

****

Label Propagation from ImageNet to 3D Point Clouds

Yan Wang, Rongrong Ji, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3135-3142

Recent years have witnessed a growing interest in understanding the semantics of point clouds in a wide variety of applications. However, point cloud labeling remains an open problem, due to the difficulty in acquiring sufficient 3D point l abels towards training effective classifiers. In this paper, we overcome this challenge by utilizing the existing massive 2D semantic labeled datasets from decadelong community efforts, such as ImageNet and LabelMe, and a novel "cross-domain" label propagation approach. Our proposed method consists of two major novel components, Exemplar SVM based label propagation, which effectively addresses the cross-domain issue, and a graphical model based contextual refinement incorporating 3D constraints. Most importantly, the entire process does not require any training data from the target scenes, also with good scalability towards large scale applications. We evaluate our approach on the well-known Cornell Point Cloud Dataset, achieving much greater efficiency and comparable accuracy even without any 3D training data. Our approach shows further major gains in accuracy when the training data from the target scenes is used, outperforming state-ofthe-art approaches with far better efficiency.

****

Supervised Semantic Gradient Extraction Using Linear-Time Optimization
Shulin Yang, Jue Wang, Linda Shapiro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2826-2833

This paper proposes a new supervised semantic edge and gradient extraction approach, which allows the user to roughly scribble over the desired region to extract semantically-dominant and coherent edges in it. Our approach first extracts low-level edgelets (small edge clusters) from the input image as primitives and build a graph upon them, by jointly considering both the geometric and appearance compatibility of edgelets. Given the characteristics of the graph, it cannot be effectively optimized by commonly-used energy minimization tools such as graph cuts. We thus propose an efficient linear algorithm for precise graph optimization, by taking advantage of the special structure of the graph. Objective evaluations show that the proposed method significantly outperforms previous semantic edge detection algorithms. Finally, we demonstrate the effectiveness of the system in various image editing tasks.

****

Deep Learning Shape Priors for Object Segmentation
Fei Chen, Huimin Yu, Roland Hu, Xunxun Zeng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1870-1877

In this paper we introduce a new shape-driven approach for object segmentation. Given a training set of shapes, we first use deep Boltzmann machine to learn the hierarchical architecture of shape priors. This learned hierarchical architecture is then used to model shape variations of global and local structures in an energetic form. Finally, it is applied to data-driven variational methods to perform object extraction of corrupted data based on shape probabilistic representation. Experiments demonstrate that our model can be applied to dataset of arbitrary prior shapes, and can cope with image noise and clutter, as well as partial occlusions.

****

Consensus of k-NNs for Robust Neighborhood Selection on Graph-Based Manifolds
Vittal Premachandran, Ramakrishna Kakarala; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1594-1601

Propagating similarity information along the data manifold requires careful selection of local neighborhood. Selecting a "good" neighborhood in an unsupervised setting, given an affinity graph, has been a difficult task. The most common way to select a local neighborhood has been to use the k-nearest neighborhood (k-NN) selection criterion. However, it has the tendency to include noisy edges. In this paper, we propose a way to select a robust neighborhood using the consensus of multiple rounds of k-NNs. We explain how using consensus information can give better control over neighborhood selection. We also explain in detail the problems with another recently proposed neighborhood selection criteria, i.e., Dominant Neighbors, and show that our method is immune to those problems. Finally, we show the results from experiments in which we compare our method to other neighborhood selection approaches. The results corroborate our claims that consensus of k-NNs does indeed help in selecting more robust and stable localities.

****

Semi-supervised Learning with Constraints for Person Identification in Multimedia Data

Martin Bauml, Makarand Tapaswi, Rainer Stiefelhagen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3602-3609

We address the problem of person identification in TV series. We propose a unified learning framework for multiclass classification which incorporates labeled and unlabeled data, and constraints between pairs of features in the training. We apply the framework to train multinomial logistic regression classifiers for multi-class face recognition. The method is completely automatic, as the labeled data is obtained by tagging speaking faces using subtitles and fan transcripts of the videos. We demonstrate our approach on six episodes each of two diverse TV series and achieve state-of-the-art performance.

****

Capturing Layers in Image Collections with Componential Models: From the Layered Epitome to the Componential Counting Grid

Alessandro Perina, Nebojsa Jojic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 500-507

Recently, the Counting Grid (CG) model [5] was developed to represent each input image as a point in a large grid of feature counts. This latent point is a corner of a window of grid points which are all uniformly combined to match the (normalized) feature counts in the image. Being a bag of word model with spatial layout in the latent space, the CG model has superior handling of field of view changes in comparison to other bag of word models, but with the price of being essentially a mixture, mapping each scene to a single window in the grid. In this paper we introduce a family of componential models, dubbed the Componential Counting Grid, whose members represent each input image by multiple latent locations, rather than just one. In this way, we make a substantially more flexible admixture model which captures layers or parts of images and maps them to separate windows in a Counting Grid. We tested the models on scene and place classification where their componential nature helped to extract objects, to capture parallax effects, thus better fitting the data and outperforming Counting Grids and Latent Dirichlet Allocation, especially on sequences taken with wearable cameras.

****

Layer Depth Denoising and Completion for Structured-Light RGB-D Cameras

Ju Shen, Sen-Ching S. Cheung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1187-1194

The recent popularity of structured-light depth sensors has enabled many new applications from gesture-based user interface to 3D reconstructions. The quality of the depth measurements of these systems, however, is far from perfect. Some depth values can have significant errors, while others can be missing altogether. The uncertainty in depth measurements among these sensors can significantly degrade the performance of any subsequent vision processing. In this paper, we propose a novel probabilistic model to capture various types of uncertainties in the depth measurement process among structured-light systems. The key to our model is the use of depth layers to account for the differences between foreground objects and background scene, the missing depth value phenomenon, and the correlation between color and depth channels. The depth layer labeling is solved as a maximum a-posteriori estimation problem, and a Markov Random Field attuned to the uncertainty in measurements is used to spatially smooth the labeling process. Using the depth-layer labels, we propose a depth correction and completion algorithm that outperforms other techniques in the literature.

****

Adaptive Compressed Tomography Sensing

Oren Barkan, Jonathan Weill, Amir Averbuch, Shai Dekel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2195-2202

One of the main challenges in Computed Tomography (CT) is how to balance between the amount of radiation the patient is exposed to during scan time and the quality of the CT image. We propose a mathematical model for adaptive CT acquisition whose goal is to reduce dosage levels while maintaining high image quality at t he same time. The adaptive algorithm iterates between selective limited acquisition and improved reconstruction, with the goal of applying only the dose level required for sufficient image quality. The theoretical foundation of the algorithm is nonlinear Ridgelet approximation and a discrete form of Ridgelet analysis is used to compute the selective acquisition steps that best capture the image edges. We show experimental results where for the same number of line projections, the adaptive model produces higher image quality, when compared with standard limited angle, non-adaptive acquisition algorithms.

****

Detection of Manipulation Action Consequences (MAC)

Yezhou Yang, Cornelia Fermuller, Yiannis Aloimonos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2563-2570

The problem of action recognition and human activity has been an active research area in Computer Vision and Robotics. While full-body motions can be characterized by movement and change of posture, no characterization, that holds invariance, has yet been proposed for the description of manipulation actions. We propose that a fundamental concept in understanding such actions, are the consequences of actions. There is a small set of fundamental primitive action consequences that provides a systematic high-level classification of manipulation actions. In this paper a technique is developed to recognize these action consequences. At the heart of the technique lies a novel active tracking and segmentation method that monitors the changes in appearance and topological structure of the manipulated object. These are then used in a visual semantic graph (VSG) based procedure applied to the time sequence of the monitored object to recognize the action consequence. We provide a new dataset, called Manipulation Action Consequences (MAC 1.0), which can serve as testbed for other studies on this topic. Several experiments on this dataset demonstrates that our method can robustly track objects and detect their deformations and division during the manipulation. Quantitative tests prove the effectiveness and efficiency of the method.

****

Efficient Color Boundary Detection with Color-Opponent Mechanisms

Kaifu Yang, Shaobing Gao, Chaoyi Li, Yongjie Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2810-2817

Color information plays an important role in better understanding of natural scenes by at least facilitating discriminating boundaries of objects or areas. In this study, we propose a new framework for boundary detection in complex natural scenes based on the color-opponent mechanisms of the visual system. The red-green and blue-yellow color opponent channels in the human visual system are regarded as the building blocks for various color perception tasks such as boundary detection. The proposed framework is a feedforward hierarchical model, which has direct counterpart to the color-opponent mechanisms involved in from the retina to the primary visual cortex (V1). Results show that our simple framework has excellent ability to flexibly capture both the structured chromatic and achromatic boundaries in complex scenes.

****

Better Exploiting Motion for Better Action Recognition

Mihir Jain, Herve Jegou, Patrick Bouthemy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2555-2562

Several recent works on action recognition have attested the importance of explicitly integrating motion characteristics in the video description. This paper establishes that adequately decomposing visual motion into dominant and residual motions, both in the extraction of the space-time trajectories and for the computation of descriptors, significantly improves action recognition algorithms. Then, we design a new motion descriptor, the DCS descriptor, based on differential motion scalar quantities, divergence, curl and shear features. It captures additional information on the local motion patterns enhancing results. Finally, applying the recent VLAD coding technique proposed in image retrieval provides a substantial improvement for action recognition. Our three contributions are complementary and lead to outperform all reported results by a significant margin on three challenging datasets, namely Hollywood 2, HMDB51 and Olympic Sports.

****

Constraints as Features

Shmuel Asafi, Daniel Cohen-Or; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1634-1641

In this paper, we introduce a new approach to constrained clustering which treats the constraints as features. Our method augments the original feature space with additional dimensions, each of which derived from a given Cannot-link constraints. The specified Cannot-link pair gets extreme coordinates values, and the rest of the points get coordinate values that express their spatial influence from the specified constrained pair. After augmenting all the new features, a standard unconstrained clustering algorithm can be performed, like k-means or spectral clustering. We demonstrate the efficacy of our method for active semi-supervised learning applied to image segmentation and compare it to alternative methods. We also evaluate the performance of our method on the four most commonly evaluated datasets from the UCI machine learning repository.

****

Graph-Laplacian PCA: Closed-Form Solution and Robustness

Bo Jiang, Chris Ding, Bio Luo, Jin Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3492-3498

Principal Component Analysis (PCA) is a widely used to learn a low-dimensional representation. In many applications, both vector data X and graph data W are available. Laplacian embedding is widely used for embedding graph data. We propose a graph-Laplacian PCA (gLPCA) to learn a low dimensional representation of X that incorporates graph structures encoded in W . This model has several advantages : (1) It is a data representation model. (2) It has a compact closed-form solution and can be efficiently computed. (3) It is capable to remove corruptions. Extensive experiments on 8 datasets show promising results on image reconstruction and significant improvement on clustering and classification.

****

Determining Motion Directly from Normal Flows Upon the Use of a Spherical Eye Platform

Tak-Wai Hui, Ronald Chung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2267-2274

We address the problem of recovering camera motion from video data, which does not require the establishment of feature correspondences or computation of optical flows but from normal flows directly. We have designed an imaging system that has a wide field of view by fixating a number of cameras together to form an approximate spherical eye. With a substantially widened visual field, we discover that estimating the directions of translation and rotation components of the motion separately are possible and particularly efficient. In addition, the inherent ambiguities between translation and rotation also disappear. Magnitude of rotation is recovered subsequently. Experimental results on synthetic and real image data are provided. The results show that not only the accuracy of motion estimation is comparable to those of the state-of-the-art methods that require explicit feature correspondences or optical flows, but also a faster computation time.

****

Visual Place Recognition with Repetitive Structures

Akihiko Torii, Josef Sivic, Tomas Pajdla, Masatoshi Okutomi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 883-890

Repeated structures such as building facades, fences or road markings often represent a significant challenge for place recognition. Repeated structures are notoriously hard for establishing correspondences using multi-view geometry. Even more importantly, they violate the feature independence assumed in the bag-of-visual-words representation which often leads to over-counting evidence and significant degradation of retrieval performance. In this work we show that repeated structures are not a nuisance but, when appropriately represented, they form an important distinguishing feature for many places. We describe a representation of repeated structures suitable for scalable retrieval. It is based on robust detection of repeated image structures and a simple modification of weights in the ba g-of-visual-word model. Place recognition results are shown on datasets of street-level imagery from Pittsburgh and San Francisco demonstrating significant gains in recognition performance compared to the standard bag-of-visual-words baseline and more recently proposed burstiness weighting.

****

Single-Pedestrian Detection Aided by Multi-pedestrian Detection

Wanli Ouyang, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3198-3205

In this paper, we address the challenging problem of detecting pedestrians who appear in groups and have interaction. A new approach is proposed for single-pedestrian detection aided by multi-pedestrian detection. A mixture model of multi-pedestrian detectors is designed to capture the unique visual cues which are formed by nearby multiple pedestrians but cannot be captured by single-pedestrian detectors. A probabilistic framework is proposed to model the relationship between the configurations estimated by singleand multi-pedestrian detectors, and to refine the single-pedestrian detection result with multi-pedestrian detection. It can integrate with any single-pedestrian detector without significantly increasing the computation load. 15 state-of-the-art single-pedestrian detection approaches are investigated on three widely used public datasets: Caltech, TUD-Brussels and ETH. Experimental results show that our framework significantly improves all these approaches. The average improvement is 9% on the Caltech-Test dataset, aveaon the TUD-Brussels dataset and teh-on the ETH dataset in terms of average miss rate. The lowest average miss rate is reduced from 48% to 43% on the Caltech-Test dataset, from edueto fom4on the TUD-Brussels dataset and from etfoto 55%ton the ETH dataset.

****

Understanding Bayesian Rooms Using Composite 3D Object Models

Luca Del Pero, Joshua Bowdish, Bonnie Kermgard, Emily Hartley, Kobus Barnard; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 153-160

We develop a comprehensive Bayesian generative model for understanding indoor scenes. While it is common in this domain to approximate objects with 3D bounding boxes, we propose using strong representations with finer granularity. For example, we model a chair as a set of four legs, a seat and a backrest. We find that modeling detailed geometry improves recognition and reconstruction, and enables more refined use of appearance for scene understanding. We demonstrate this with a new likelihood function that rewards 3D object hypotheses whose 2D projection is more uniform in color distribution. Such a measure would be confused by background pixels if we used a bounding box to represent a concave object like a chair. Complex objects are modeled using a set or re-usable 3D parts, and we show that this representation captures much of the variation among object instances with relatively few parameters. We also designed specific data-driven inference mechanisms for each part that are shared by all objects containing that part, which helps make inference transparent to the modeler. Further, we show how to exploit contextual relationships to detect more objects, by, for example,proposing chairs around and underneath tables. We present results showing the benefits of each of these innovations. The performance of our approach often exceeds that of state-of-the-art methods on the two tasks of room layout estimation and object recognition, as evaluated on two bench mark data sets used in this domain.

****

Groupwise Registration via Graph Shrinkage on the Image Manifold

Shihui Ying, Guorong Wu, Qian Wang, Dinggang Shen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2323-2330

Recently, groupwise registration has been investigated for simultaneous alignment of all images without selecting any individual image as the template, thus avoiding the potential bias in image registration. However, none of current groupwise registration method fully utilizes the image distribution to guide the registration. Thus, the registration performance usually suffers from large inter-subject variations across individual images. To solve this issue, we propose a novel groupwise registration algorithm for large population dataset, guided by the im

age distribution on the manifold. Specifically, we first use a graph to model the distribution of all image data sitting on the image manifold, with each node representing an image and each edge representing the geodesic pathway between two nodes (or images). Then, the procedure of warping all images to their population center turns to the dynamic shrinking of the graph nodes along their graph edges until all graph nodes become close to each other. Thus, the topology of image distribution on the image manifold is always preserved during the groupwise registration. More importantly, by modeling the distribution of all images via a graph, we can potentially reduce registration error since every time each image is warped only according to its nearby images with similar structures in the graph. We have evaluated our proposed groupwise registration method on both synthetic and real datasets, with comparison to the two state-of-the-art groupwise registration methods. All experimental results show that our proposed method achieves the best performance in terms of registration accuracy and robustness.
****
On a Link Between Kernel Mean Maps and Fraunhofer Diffraction, with an Application to Super-Resolution Beyond the Diffraction Limit
Stefan Harmeling, Michael Hirsch, Bernhard Scholkopf; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1083-1090
We establish a link between Fourier optics and a recent construction from the machine learning community termed the kernel mean map. Using the Fraunhofer approximation, it identifies the kernel with the squared Fourier transform of the aperture. This allows us to use results about the invertibility of the kernel mean map to provide a statement about the invertibility of Fraunhofer diffraction, showing that imaging processes with arbitrarily small apertures can in principle be invertible, i.e., do not lose information, provided the objects to be imaged satisfy a generic condition. A real world experiment shows that we can super-resolve beyond the Rayleigh limit.
****
Background Modeling Based on Bidirectional Analysis
Atsushi Shimada, Hajime Nagahara, Rin-ichiro Taniguchi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1979-1986
Background modeling and subtraction is an essential task in video surveillance applications. Most traditional studies use information observed in past frames to create and update a background model. To adapt to background changes, the background model has been enhanced by introducing various forms of information including spatial consistency and temporal tendency. In this paper, we propose a new framework that leverages information from a future period. Our proposed approach realizes a low-cost and highly accurate background model. The proposed framework is called bidirectional background modeling, and performs background subtraction based on bidirectional analysis; i.e., analysis from past to present and analysis from future to present. Although a result will be output with some delay because information is taken from a future period, our proposed approach improves the accuracy by about 30% if only a 33-millisecond of delay is acceptable. Furthermore, the memory cost can be reduced by about 65% relative to typical background modeling.
****
Minimum Uncertainty Gap for Robust Visual Tracking
Junseok Kwon, Kyoung Mu Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2355-2362
We propose a novel tracking algorithm that robustly tracks the target by finding the state which minimizes uncertainty of the likelihood at current state. The uncertainty of the likelihood is estimated by obtaining the gap between the lower and upper bounds of the likelihood. By minimizing the gap between the two bounds, our method finds the confident and reliable state of the target. In the paper, the state that gives the Minimum Uncertainty Gap (MUG) between likelihood bounds is shown to be more reliable than the state which gives the maximum likelihood only, especially when there are severe illumination changes, occlusions, and pose variations. A rigorous derivation of the lower and upper bounds of the likel

ihood for the visual tracking problem is provided to address this issue. Additionally, an efficient inference algorithm using Interacting Markov Chain Monte Carlo is presented to find the best state that maximizes the average of the lower and upper bounds of the likelihood and minimizes the gap between two bounds simultaneously. Experimental results demonstrate that our method successfully tracks the target in realistic videos and outperforms conventional tracking methods.

****

Real-Time No-Reference Image Quality Assessment Based on Filter Learning

Peng Ye, Jayant Kumar, Le Kang, David Doermann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 987-994

This paper addresses the problem of general-purpose No-Reference Image Quality Assessment (NR-IQA) with the goal of developing a real-time, cross-domain model that can predict the quality of distorted images without prior knowledge of non-distorted reference images and types of distortions present in these images. The contributions of our work are two-fold: first, the proposed method is highly efficient. NR-IQA measures are often used in real-time imaging or communication systems, therefore it is important to have a fast NR-IQA algorithm that can be used in these real-time applications. Second, the proposed method has the potential to be used in multiple image domains. Previous work on NR-IQA focus primarily on predicting quality of natural scene image with respect to human perception, yet, in other image domains, the final receiver of a digital image may not be a human. The proposed method consists of the following components: (1) a local feature extractor; (2) a global feature extractor and (3) a regression model. While previous approaches usually treat local feature extraction and regression model training independently, we propose a supervised method based on back-projection, which links the two steps by learning a compact set of filters which can be applied to local image patches to obtain discriminative local features. Using a small set of filters, the proposed method is extremely fast. We have tested this method on various natural scene and document image datasets and obtained stateof-the-art results.

****

City-Scale Change Detection in Cadastral 3D Models Using Images

Aparna Taneja, Luca Ballan, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 113-120

In this paper, we propose a method to detect changes in the geometry of a city using panoramic images captured by a car driving around the city. We designed our approach to account for all the challenges involved in a large scale application of change detection, such as, inaccuracies in the input geometry, errors in the geo-location data of the images, as well as, the limited amount of information due to sparse imagery. We evaluated our approach on an area of 6 square kilometers inside a city, using 3420 images downloaded from Google StreetView. These images besides being publicly available, are also a good example of panoramic images captured with a driving vehicle, and hence demonstrating all the possible challenges resulting from such an acquisition. We also quantitatively compared the performance of our approach with respect to a ground truth, as well as to prior work. This evaluation shows that our approach outperforms the current state of the art.

****

Occlusion Patterns for Object Class Detection

Bojan Pepikj, Michael Stark, Peter Gehler, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3286-3293

Despite the success of recent object class recognition systems, the long-standing problem of partial occlusion remains a major challenge, and a principled solution is yet to be found. In this paper we leave the beaten path of methods that treat occlusion as just another source of noise instead, we include the occluder itself into the modelling, by mining distinctive, reoccurring occlusion patterns from annotated training data. These patterns are then used as training data for dedicated detectors of varying sophistication. In particular, we evaluate and compare models that range from standard object class detectors to hierarchical, p

art-based representations of occluder/occludee pairs. In an extensive evaluation we derive insights that can aid further developments in tackling the occlusion challenge.
****
Local Fisher Discriminant Analysis for Pedestrian Re-identification
Sateesh Pedagadi, James Orwell, Sergio Velastin, Boghos Boghossian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3318-3325
Metric learning methods, for person re-identification, estimate a scaling for distances in a vector space that is optimized for picking out observations of the same individual. This paper presents a novel approach to the pedestrian re-identification problem that uses metric learning to improve the state-of-the-art performance on standard public datasets. Very high dimensional features are extracted from the source color image. A first processing stage performs unsupervised PCA dimensionality reduction, constrained to maintain the redundancy in color-space representation. A second stage further reduces the dimensionality, using a Local Fisher Discriminant Analysis defined by a training set. A regularization step is introduced to avoid singular matrices during this stage. The experiments conducted on three publicly available datasets confirm that the proposed method outperforms the state-of-the-art performance, including all other known metric learning methods. Furthermore, the method is an effective way to process observations comprising multiple shots, and is non-iterative: the computation times are relatively modest. Finally, a novel statistic is derived to characterize the Match Characteristic: the normalized entropy reduction can be used to define the 'Proportion of Uncertainty Removed' (P UR). This measure is invariant to test set size and provides an intuitive indication of performance.
****
Semi-supervised Node Splitting for Random Forest Construction
Xiao Liu, Mingli Song, Dacheng Tao, Zicheng Liu, Luming Zhang, Chun Chen, Jiajun Bu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 492-499
Node splitting is an important issue in Random Forest but robust splitting requires a large number of training samples. Existing solutions fail to properly partition the feature space if there are insufficient training data. In this paper, we present semi-supervised splitting to overcome this limitation by splitting nodes with the guidance of both labeled and unlabeled data. In particular, we derive a nonparametric algorithm to obtain an accurate quality measure of splitting by incorporating abundant unlabeled data. To avoid the curse of dimensionality, we project the data points from the original high-dimensional feature space onto a low-dimensional subspace before estimation. A unified optimization framework is proposed to select a coupled pair of subspace and separating hyperplane such that the smoothness of the subspace and the quality of the splitting are guaranteed simultaneously. The proposed algorithm is compared with state-of-the-art supervised and semi-supervised algorithms for typical computer vision applications such as object categorization and image segmentation. Experimental results on publicly available datasets demonstrate the superiority of our method.
****
Vantage Feature Frames for Fine-Grained Categorization
Asma Rejeb Sfar, Nozha Boujemaa, Donald Geman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 835-842
We study fine-grained categorization, the task of distinguishing among (sub)categories of the same generic object class (e.g., birds), focusing on determining botanical species (leaves and orchids) from scanned images. The strategy is to focus attention around several vantage points, which is the approach taken by botanists, but using features dedicated to the individual categories. Our implementation of the strategy is based on vantage feature frames, a novel object representation consisting of two components: a set of coordinate systems centered at the most discriminating local viewpoints for the generic object class and a set of category-dependent features computed in these frames. The features are pooled over frames to build the classifier. Categorization then proceeds from coarse-grai

ned (finding the frames) to fine-grained (finding the category), and hence the vantage feature frames must be both detectable and discriminating. The proposed method outperforms state-of-the art algorithms, in particular those using more distributed representations, on standard databases of leaves.

****

A Video Representation Using Temporal Superpixels

Jason Chang, Donglai Wei, John W. Fisher III; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2051-2058

We develop a generative probabilistic model for temporally consistent superpixels in video sequences. In contrast to supervoxel methods, object parts in different frames are tracked by the same temporal superpixel. We explicitly model flow between frames with a bilateral Gaussian process and use this information to propagate superpixels in an online fashion. We consider four novel metrics to quantify performance of a temporal superpixel representation and demonstrate superior performance when compared to supervoxel methods.

****

Structure Preserving Object Tracking

Lu Zhang, Laurens van der Maaten; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1838-1845

Model-free trackers can track arbitrary objects based on a single (bounding-box) annotation of the object. Whilst the performance of model-free trackers has recently improved significantly, simultaneously tracking multiple objects with similar appearance remains very hard. In this paper, we propose a new multi-object model-free tracker (based on tracking-by-detection) that resolves this problem by incorporating spatial constraints between the objects. The spatial constraints are learned along with the object detectors using an online structured SVM algorithm. The experimental evaluation of our structure-preserving object tracker (SPOT) reveals significant performance improvements in multi-object tracking. We also show that SPOT can improve the performance of single-object trackers by simultaneously tracking different parts of the object.

****

Unsupervised Salience Learning for Person Re-identification

Rui Zhao, Wanli Ouyang, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3586-3593

Human eyes can recognize person identities based on some small salient regions. However, such valuable salient information is often hidden when computing similarities of images with existing approaches. Moreover, many existing approaches learn discriminative features and handle drastic viewpoint change in a supervised way and require labeling new training data for a different pair of camera views. In this paper, we propose a novel perspective for person re-identification based on unsupervised salience learning. Distinctive features are extracted without requiring identity labels in the training procedure. First, we apply adjacency constrained patch matching to build dense correspondence between image pairs, which shows effectiveness in handling misalignment caused by large viewpoint and pose variations. Second, we learn human salience in an unsupervised manner. To improve the performance of person re-identification, human salience is incorporated in patch matching to find reliable and discriminative matched patches. The effectiveness of our approach is validated on the widely used VIPeR dataset and ETHZ dataset.

****

Spatiotemporal Deformable Part Models for Action Detection

Yicong Tian, Rahul Sukthankar, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2642-2649

Deformable part models have achieved impressive performance for object detection, even on difficult image datasets. This paper explores the generalization of deformable part models from 2D images to 3D spatiotemporal volumes to better study their effectiveness for action detection in video. Actions are treated as spatiotemporal patterns and a deformable part model is generated for each action from a collection of examples. For each action model, the most discriminative 3D subvolumes are automatically selected as parts and the spatiotemporal relations bet ween their locations are learned. By focusing on the most distinctive parts of each action, our models adapt to intra-class variation and show robustness to clutter. Extensive experiments on several video datasets demonstrate the strength of spatiotemporal DPMs for classifying and localizing actions.

****

Beyond Point Clouds: Scene Understanding by Reasoning Geometry and Physics

Bo Zheng, Yibiao Zhao, Joey C. Yu, Katsushi Ikeuchi, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3127-3134

In this paper, we present an approach for scene understanding by reasoning physical stability of objects from point cloud. We utilize a simple observation that, by human design, objects in static scenes should be stable with respect to gravity. This assumption is applicable to all scene categories and poses useful constraints for the plausible interpretations (parses) in scene understanding. Our method consists of two major steps: 1) geometric reasoning: recovering solid 3D volumetric primitives from defective point cloud; and 2) physical reasoning: grouping the unstable primitives to physically stable objects by optimizing the stability and the scene prior. We propose to use a novel disconnectivity graph (DG) to represent the energy landscape and use a Swendsen-Wang Cut (MCMC) method for optimization. In experiments, we demonstrate that the algorithm achieves substantially better performance for i) object segmentation, ii) 3D volumetric recovery of the scene, and iii) better parsing result for scene understanding in comparison to state-of-the-art methods in both public dataset and our own new dataset.

****

Recovering Stereo Pairs from Anaglyphs

Armand Joulin, Sing Bing Kang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 289-296

An anaglyph is a single image created by selecting complementary colors from a stereo color pair; the user can perceive depth by viewing it through color-filtered glasses. We propose a technique to reconstruct the original color stereo pair given such an anaglyph. We modified SIFT-Flow and use it to initially match the different color channels across the two views. Our technique then iteratively refines the matches, selects the good matches (which defines the "anchor" colors), and propagates the anchor colors. We use a diffusion-based technique for the color propagation, and added a step to suppress unwanted colors. Results on a variety of inputs demonstrate the robustness of our technique. We also extended our method to anaglyph videos by using optic flow between time frames.

****

Axially Symmetric 3D Pots Configuration System Using Axis of Symmetry and Break Curve

Kilho Son, Eduardo B. Almeida, David B. Cooper; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 257-264

This paper introduces a novel approach for reassembling pot sherds found at archaeological excavation sites, for the purpose of reconstructing clay pots that had been made on a wheel. These pots and the sherds into which they have broken are axially symmetric. The reassembly process can be viewed as 3D puzzle solving or generalized cylinder learning from broken fragments. The estimation exploits both local and semi-global geometric structure, thus making it a fundamental problem of geometry estimation from noisy fragments in computer vision and pattern recognition. The data used are densely digitized 3D laser scans of each fragment's outer surface. The proposed reassembly system is automatic and functions when the pile of available fragments is from one or multiple pots, and even when pieces are missing from any pot. The geometric structure used are curves on the pot along which the surface had broken and the silhouette of a pot with respect to an axis, called axisprofile curve (APC). For reassembling multiple pots with or without missing pieces, our algorithm estimates the APC from each fragment, then reassembles into configurations the ones having distinctive APC. Further growth of configurations is based on adding remaining fragments such that their APC and break curves are consistent with those of a configuration. The method is novel, more robust and handles the largest numbers of fragments to date.

****

Learning a Manifold as an Atlas

Nikolaos Pitelis, Chris Russell, Lourdes Agapito; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1642-1649

In this work, we return to the underlying mathematical definition of a manifold and directly characterise learning a manifold as finding an atlas, or a set of overlapping charts, that accurately describe local structure. We formulate the problem of learning the manifold as an optimisation that simultaneously refines the continuous parameters defining the charts, and the discrete assignment of points to charts. In contrast to existing methods, this direct formulation of a manifold does not require "unwrapping" the manifold into a lower dimensional space and allows us to learn closed manifolds of interest to vision, such as those corresponding to gait cycles or camera pose. We report state-ofthe-art results for manifold based nearest neighbour classification on vision datasets, and show how the same techniques can be applied to the 3D reconstruction of human motion from a single image.

****

Label-Embedding for Attribute-Based Classification

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, Cordelia Schmid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 819-826

Attributes are an intermediate representation, which enables parameter sharing between classes, a must when training data is scarce. We propose to view attribute-based image classification as a label-embedding problem: each class is embedded in the space of attribute vectors. We introduce a function which measures the compatibility between an image and a label embedding. The parameters of this function are learned on a training set of labeled samples to ensure that, given an image, the correct classes rank higher than the incorrect ones. Results on the Animals With Attributes and Caltech-UCSD-Birds datasets show that the proposed framework outperforms the standard Direct Attribute Prediction baseline in a zero-shot learning scenario. The label embedding framework offers other advantages such as the ability to leverage alternative sources of information in addition to attributes (e.g. class hierarchies) or to transition smoothly from zero-shot learning to learning with large quantities of data.

****

Dynamic Scene Classification: Learning Motion Descriptors with Slow Features Analysis

Christian Theriault, Nicolas Thome, Matthieu Cord; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2603-2610

In this paper, we address the challenging problem of categorizing video sequences composed of dynamic natural scenes. Contrarily to previous methods that rely on handcrafted descriptors, we propose here to represent videos using unsupervised learning of motion features. Our method encompasses three main contributions: 1) Based on the Slow Feature Analysis principle, we introduce a learned local motion descriptor which represents the principal and more stable motion components of training videos. 2) We integrate our local motion feature into a global coding/pooling architecture in order to provide an effective signature for each video sequence. 3) We report state of the art classification performances on two challenging natural scenes data sets. In particular, an outstanding improvement of 11% in classification score is reached on a data set introduced in 2012.

****

The Episolar Constraint: Monocular Shape from Shadow Correspondence

Austin Abrams, Kylia Miskell, Robert Pless; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1407-1414

Shadows encode a powerful geometric cue: if one pixel casts a shadow onto another, then the two pixels are colinear with the lighting direction. Given many images over many lighting directions, this constraint can be leveraged to recover the depth of a scene from a single viewpoint. For outdoor scenes with solar illumination, we term this the episolar constraint, which provides a convex optimization to solve for the sparse depth of a scene from shadow correspondences, a metho d to reduce the search space when finding shadow correspondences, and a method to geometrically calibrate a camera using shadow constraints. Our method constructs a dense network of nonlocal constraints which complements recent work on outdoor photometric stereo and cloud based cues for 3D. We demonstrate results across a variety of time-lapse sequences from webcams "in the wild."

****

Learning and Calibrating Per-Location Classifiers for Visual Place Recognition

Petr Gronat, Guillaume Obozinski, Josef Sivic, Tomas Pajdla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 907-914

The aim of this work is to localize a query photograph by finding other images depicting the same place in a large geotagged image database. This is a challenging task due to changes in viewpoint, imaging conditions and the large size of the image database. The contribution of this work is two-fold. First, we cast the place recognition problem as a classification task and use the available geotags to train a classifier for each location in the database in a similar manner to per-exemplar SVMs in object recognition. Second, as only few positive training examples are available for each location, we propose a new approach to calibrate all the per-location SVM classifiers using only the negative examples. The calibration we propose relies on a significance measure essentially equivalent to the p-values classically used in statistical hypothesis testing. Experiments are performed on a database of 25,000 geotagged street view images of Pittsburgh and demonstrate improved place recognition accuracy of the proposed approach over the previous work.

****

Blind Deconvolution of Widefield Fluorescence Microscopic Data by Regularization of the Optical Transfer Function (OTF)

Margret Keuper, Thorsten Schmidt, Maja Temerinac-Ott, Jan Padeken, Patrick Heun, Olaf Ronneberger, Thomas Brox; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2179-2186

With volumetric data from widefield fluorescence microscopy, many emerging questions in biological and biomedical research are being investigated. Data can be recorded with high temporal resolution while the specimen is only exposed to a low amount of phototoxicity. These advantages come at the cost of strong recording blur caused by the infinitely extended point spread function (PSF). For widefield microscopy, its magnitude only decays with the square of the distance to the focal point and consists of an airy bessel pattern which is intricate to describe in the spatial domain. However, the Fourier transform of the incoherent PSF (denoted as Optical Transfer Function (OTF)) is well localized and smooth. In this paper, we present a blind deconvolution method that improves results of state-of-theart deconvolution methods on widefield data by exploiting the properties of the widefield OTF.

****

Tensor-Based Human Body Modeling

Yinpeng Chen, Zicheng Liu, Zhengyou Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 105-112

In this paper, we present a novel approach to model 3D human body with variations on both human shape and pose, by exploring a tensor decomposition technique. 3D human body modeling is important for 3D reconstruction and animation of realistic human body, which can be widely used in Tele-presence and video game applications. It is challenging due to a wide range of shape variations over different people and poses. The existing SCAPE model [4] is popular in computer vision for modeling 3D human body. However, it considers shape and pose deformations separately, which is not accurate since pose deformation is persondependent. Our tensor-based model addresses this issue by jointly modeling shape and pose deformations. Experimental results demonstrate that our tensor-based model outperforms the SCAPE model quite significantly. We also apply our model to capture human body using Microsoft Kinect sensors with excellent results.

****

Segment-Tree Based Cost Aggregation for Stereo Matching

Xing Mei, Xun Sun, Weiming Dong, Haitao Wang, Xiaopeng Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 313-320

This paper presents a novel tree-based cost aggregation method for dense stereo matching. Instead of employing the minimum spanning tree (MST) and its variants, a new tree structure, "Segment-Tree", is proposed for non-local matching cost aggregation. Conceptually, the segment-tree is constructed in a three-step process: first, the pixels are grouped into a set of segments with the reference color or intensity image; second, a tree graph is created for each segment; and in the final step, these independent segment graphs are linked to form the segment-tree structure. In practice, this tree can be efficiently built in time nearly linear to the number of the image pixels. Compared to MST where the graph connectivity is determined with local edge weights, our method introduces some 'non-local' decision rules: the pixels in one perceptually consistent segment are more likely to share similar disparities, and therefore their connectivity within the segment should be first enforced in the tree construction process. The matching costs are then aggregated over the tree within two passes. Performance evaluation on 19 Middlebury data sets shows that the proposed method is comparable to previous state-of-the-art aggregation methods in disparity accuracy and processing speed. Furthermore, the tree structure can be refined with the estimated disparities, which leads to consistent scene segmentation and significantly better aggregation results.

****

Category Modeling from Just a Single Labeling: Use Depth Information to Guide the Learning of 2D Models

Quanshi Zhang, Xuan Song, Xiaowei Shao, Ryosuke Shibasaki, Huijing Zhao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 193-200

An object model base that covers a large number of object categories is of great value for many computer vision tasks. As artifacts are usually designed to have various textures, their structure is the primary distinguishing feature between different categories. Thus, how to encode this structural information and how to start the model learning with a minimum of human labeling become two key challenges for the construction of the model base. We design a graphical model that uses object edges to represent object structures, and this paper aims to incrementally learn this category model from one labeled object and a number of casually captured scenes. However, the incremental model learning may be biased due to the limited human labeling. Therefore, we propose a new strategy that uses the depth information in RGBD images to guide the model learning for object detection in ordinary RGB images. In experiments, the proposed method achieves superior performance as good as the supervised methods that require the labeling of all target objects.

****

Human Pose Estimation Using a Joint Pixel-wise and Part-wise Formulation

Lubor Ladicky, Philip H.S. Torr, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3578-3585

Our goal is to detect humans and estimate their 2D pose in single images. In particular, handling cases of partial visibility where some limbs may be occluded or one person is partially occluding another. Two standard, but disparate, approaches have developed in the field: the first is the part based approach for layout type problems, involving optimising an articulated pictorial structure; the second is the pixel based approach for image labelling involving optimising a random field graph defined on the image. Our novel contribution is a formulation for pose estimation which combines these two models in a principled way in one optimisation problem and thereby inherits the advantages of both of them. Inference on this joint model finds the set of instances of persons in an image, the location of their joints, and a pixel-wise body part labelling. We achieve near or state of the art results on standard human pose data sets, and demonstrate the correct estimation for cases of self-occlusion, person overlap and image truncation.

****

Learning Separable Filters

Roberto Rigamonti, Amos Sironi, Vincent Lepetit, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2754-2761

Learning filters to produce sparse image representations in terms of overcomplete dictionaries has emerged as a powerful way to create image features for many different purposes. Unfortunately, these filters are usually both numerous and non-separable, making their use computationally expensive. In this paper, we show that such filters can be computed as linear combinations of a smaller number of separable ones, thus greatly reducing the computational complexity at no cost in terms of performance. This makes filter learning approaches practical even for large images or 3D volumes, and we show that we significantly outperform state-of-theart methods on the linear structure extraction task, in terms of both accuracy and speed. Moreover, our approach is general and can be used on generic filter banks to reduce the complexity of the convolutions.

****

Tracking Human Pose by Tracking Symmetric Parts

Varun Ramakrishna, Takeo Kanade, Yaser Sheikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3728-3735

The human body is structurally symmetric. Tracking by detection approaches for human pose suffer from double counting, where the same image evidence is used to explain two separate but symmetric parts, such as the left and right feet. Double counting, if left unaddressed can critically affect subsequent processes, such as action recognition, affordance estimation, and pose reconstruction. In this work, we present an occlusion aware algorithm for tracking human pose in an image sequence, that addresses the problem of double counting. Our key insight is that tracking human pose can be cast as a multi-target tracking problem where the "targets" are related by an underlying articulated structure. The human body is modeled as a combination of singleton parts (such as the head and neck) and symmetric pairs of parts (such as the shoulders, knees, and feet). Symmetric body parts are jointly tracked with mutual exclusion constraints to prevent double counting by reasoning about occlusion. We evaluate our algorithm on an outdoor dataset with natural background clutter, a standard indoor dataset (HumanEva-I), and compare against a state of the art pose estimation algorithm.

****

Facial Feature Tracking Under Varying Facial Expressions and Face Poses Based on Restricted Boltzmann Machines

Yue Wu, Zuoguan Wang, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3452-3459

Facial feature tracking is an active area in computer vision due to its relevance to many applications. It is a nontrivial task, since faces may have varying facial expressions, poses or occlusions. In this paper, we address this problem by proposing a face shape prior model that is constructed based on the Restricted Boltzmann Machines (RBM) and their variants. Specifically, we first construct a model based on Deep Belief Networks to capture the face shape variations due to varying facial expressions for near-frontal view. To handle pose variations, the frontal face shape prior model is incorporated into a 3-way RBM model that could capture the relationship between frontal face shapes and non-frontal face shapes. Finally, we introduce methods to systematically combine the face shape prior models with image measurements of facial feature points. Experiments on benchmark databases show that with the proposed method, facial feature points can be tracked robustly and accurately even if faces have significant facial expressions and poses.

****

Weakly Supervised Learning of Mid-Level Features with Beta-Bernoulli Process Restricted Boltzmann Machines

Roni Mittelman, Honglak Lee, Benjamin Kuipers, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 476-483

The use of semantic attributes in computer vision problems has been gaining increased popularity in recent years. Attributes provide an intermediate feature representation in between low-level features and the class categories, leading to improved learning on novel categories from few examples. However, a major caveat is that learning semantic attributes is a laborious task, requiring a significant amount of time and human intervention to provide labels. In order to address this issue, we propose a weakly supervised approach to learn mid-level features, where only class-level supervision is provided during training. We develop a novel extension of the restricted Boltzmann machine (RBM) by incorporating a Beta-Bernoulli process factor potential for hidden units. Unlike the standard RBM, our model uses the class labels to promote category-dependent sharing of learned features, which tends to improve the generalization performance. By using semantic attributes for which annotations are available, we show that we can find correspondences between the learned mid-level features and the labeled attributes. Therefore, the mid-level features have distinct semantic characterization which is similar to that given by the semantic attributes, even though their labeling was not provided during training. Our experimental results on object recognition tasks show significant performance gains, outperforming existing methods which rely on manually labeled semantic attributes.

****

K-Means Hashing: An Affinity-Preserving Quantization Method for Learning Binary Compact Codes

Kaiming He, Fang Wen, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2938-2945

In computer vision there has been increasing interest in learning hashing codes whose Hamming distance approximates the data similarity. The hashing functions play roles in both quantizing the vector space and generating similarity-preserving codes. Most existing hashing methods use hyper-planes (or kernelized hyper-planes) to quantize and encode. In this paper, we present a hashing method adopting the k-means quantization. We propose a novel Affinity-Preserving K-means algorithm which simultaneously performs k-means clustering and learns the binary indices of the quantized cells. The distance between the cells is approximated by the Hamming distance of the cell indices. We further generalize our algorithm to a product space for learning longer codes. Experiments show our method, named as K-means Hashing (KMH), outperforms various state-of-the-art hashing encoding methods.

****

Rolling Riemannian Manifolds to Solve the Multi-class Classification Problem

Rui Caseiro, Pedro Martins, Joao F. Henriques, Fatima Silva Leite, Jorge Batista; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 41-48

In the past few years there has been a growing interest on geometric frameworks to learn supervised classification models on Riemannian manifolds [31, 27]. A popular framework, valid over any Riemannian manifold, was proposed in [31] for binary classification. Once moving from binary to multi-class classification this paradigm is not valid anymore, due to the spread of multiple positive classes on the manifold [27]. It is then natural to ask whether the multi-class paradigm could be extended to operate on a large class of Riemannian manifolds. We propose a mathematically well-founded classification paradigm that allows to extend the work in [31] to multi-class models, taking into account the structure of the space. The idea is to project all the data from the manifold onto an affine tangent space at a particular point. To mitigate the distortion induced by local diffeomorphisms, we introduce for the first time in the computer vision community a well-founded mathematical concept, so-called Rolling map [21, 16]. The novelty in this alternate school of thought is that the manifold will be firstly rolled (without slipping or twisting) as a rigid body, then the given data is unwrapped onto the affine tangent space, where the classification is performed.

****

Mesh Based Semantic Modelling for Indoor and Outdoor Scenes

Julien P.C. Valentin, Sunando Sengupta, Jonathan Warrell, Ali Shahrokni, Philip

H.S. Torr; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2067-2074

Semantic reconstruction of a scene is important for a variety of applications such as 3D modelling, object recognition and autonomous robotic navigation. However, most object labelling methods work in the image domain and fail to capture the information present in 3D space. In this work we propose a principled way to generate object labelling in 3D. Our method builds a triangulated meshed representation of the scene from multiple depth estimates. We then define a CRF over this mesh, which is able to capture the consistency of geometric properties of the objects present in the scene. In this framework, we are able to generate object hypotheses by combining information from multiple sources: geometric properties (from the 3D mesh), and appearance properties (from images). We demonstrate the robustness of our framework in both indoor and outdoor scenes. For indoor scenes we created an augmented version of the NYU indoor scene dataset (RGB D images) with object labelled meshes for training and evaluation. For outdoor scenes, we created ground truth object labellings for the KITTI odometry dataset (stereo image sequence). We observe a significant speed-up in the inference stage by performing labelling on the mesh, and additionally achieve higher accuracies.

****

A Bayesian Approach to Multimodal Visual Dictionary Learning

Go Irie, Dong Liu, Zhenguo Li, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 329-336

nary learning methods rely on image descriptors alone or together with class labels. However, Web images are often associated with text data which may carry substantial information regarding image semantics, and may be exploited for visual dictionary learning. This paper explores this idea by leveraging relational information between image descriptors and textual words via co-clustering, in addition to information of image descriptors. Existing co-clustering methods are not optimal for this problem because they ignore the structure of image descriptors in the continuous space, which is crucial for capturing visual characteristics of images. We propose a novel Bayesian co-clustering model to jointly estimate the underlying distributions of the continuous image descriptors as well as the relationship between such distributions and the textual words through a unified Bayesian inference. Extensive experiments on image categorization and retrieval have validated the substantial value of the proposed joint modeling in improving visual dictionary learning, where our model shows superior performance over several recent methods.

****

Photometric Ambient Occlusion

Daniel Hauagge, Scott Wehrwein, Kavita Bala, Noah Snavely; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2515-2522

We present a method for computing ambient occlusion (AO) for a stack of images of a scene from a fixed viewpoint. Ambient occlusion, a concept common in computer graphics, characterizes the local visibility at a point: it approximates how much light can reach that point from different directions without getting blocked by other geometry. While AO has received surprisingly little attention in vision, we show that it can be approximated using simple, per-pixel statistics over image stacks, based on a simplified image formation model. We use our derived AO measure to compute reflectance and illumination for objects without relying on additional smoothness priors, and demonstrate state-of-the art performance on the MIT Intrinsic Images benchmark. We also demonstrate our method on several synthetic and real scenes, including 3D printed objects with known ground truth geometry.

****

Beyond Physical Connections: Tree Models in Human Pose Estimation

Fang Wang, Yi Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 596-603

Simple tree models for articulated objects prevails in the last decade. However, it is also believed that these simple tree models are not capable of capturing large variations in many scenarios, such as human pose estimation. This paper attempts to address three questions: 1) are simple tree models sufficient? more specifically, 2) how to use tree models effectively in human pose estimation? and 3) how shall we use combined parts together with single parts efficiently? Assuming we have a set of single parts and combined parts, and the goal is to estimate a joint distribution of their locations. We surprisingly find that no latent variables are introduced in the Leeds Sport Dataset (LSP) during learning latent trees for deformable model, which aims at approximating the joint distributions of body part locations using minimal tree structure. This suggests one can straightforwardly use a mixed representation of single and combined parts to approximate their joint distribution in a simple tree model. As such, one only needs to build Visual Categories of the combined parts, and then perform inference on the learned latent tree. Our method outperformed the state of the art on the LSP, both in the scenarios when the training images are from the same dataset and from the PARSE dataset. Experiments on animal images from the VOC challenge further support our findings.

****

Patch Match Filter: Efficient Edge-Aware Filtering Meets Randomized Search for Fast Correspondence Field Estimation

Jiangbo Lu, Hongsheng Yang, Dongbo Min, Minh N. Do; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1854-1861

Though many tasks in computer vision can be formulated elegantly as pixel-labeling problems, a typical challenge discouraging such a discrete formulation is often due to computational efficiency. Recent studies on fast cost volume filtering based on efficient edge-aware filters have provided a fast alternative to solve discrete labeling problems, with the complexity independent of the support window size. However, these methods still have to step through the entire cost volume exhaustively, which makes the solution speed scale linearly with the label space size. When the label space is huge, which is often the case for (subpixelaccurate) stereo and optical flow estimation, their computational complexity becomes quickly unacceptable. Developed to search approximate nearest neighbors rapidly, the PatchMatch method can significantly reduce the complexity dependency on the search space size. But, its pixel-wise randomized search and fragmented data access within the 3D cost volume seriously hinder the application of efficient cost slice filtering. This paper presents a generic and fast computational framework for general multi-labeling problems called PatchMatch Filter (PMF). For the very first time, we explore effective and efficient strategies to weave together these two fundamental techniques developed in isolation, i.e., PatchMatch-based randomized search and efficient edge-aware image filtering. By decompositing an image into compact superpixels, we also propose superpixelbased novel search strategies that generalize and improve the original PatchMatch method. Focusing on dense correspondence field estimation in this paper, we demonstrate PMF's applications in stereo and optical flow. Our PMF methods achieve state-of-the-art correspondence accuracy but run much faster than other competing methods, often giving over 10-times speedup for large label space cases.

****

Generalized Domain-Adaptive Dictionaries

Sumit Shekhar, Vishal M. Patel, Hien V. Nguyen, Rama Chellappa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 361-368

Data-driven dictionaries have produced state-of-the-art results in various classification tasks. However, when the target data has a different distribution than the source data, the learned sparse representation may not be optimal. In this paper, we investigate if it is possible to optimally represent both source and target by a common dictionary. Specifically, we describe a technique which jointly learns projections of data in the two domains, and a latent dictionary which can succinctly represent both the domains in the projected low-dimensional space. An efficient optimization technique is presented, which can be easily kernelized and extended to multiple domains. The algorithm is modified to learn a common discriminative dictionary, which can be further used for classification. The pro

posed approach does not require any explicit correspondence between the source and target domains, and shows good results even when there are only a few labels available in the target domain. Various recognition experiments show that the method performs on par or better than competitive stateof-the-art methods.
****

Supervised Descent Method and Its Applications to Face Alignment
Xuehan Xiong, Fernando De la Torre; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 532-539
Many computer vision problems (e.g., camera calibration, image alignment, structure from motion) are solved through a nonlinear optimization method. It is generally accepted that 2 nd order descent methods are the most robust, fast and reliable approaches for nonlinear optimization of a general smooth function. However, in the context of computer vision, 2 nd order descent methods have two main drawbacks: (1) The function might not be analytically differentiable and numerical approximations are impractical. (2) The Hessian might be large and not positive definite. To address these issues, this paper proposes a Supervised Descent Method (SDM) for minimizing a Non-linear Least Squares (NLS) function. During training, the SDM learns a sequence of descent directions that minimizes the mean of NLS functions sampled at different points. In testing, SDM minimizes the NLS objective using the learned descent directions without computing the Jacobian nor the Hessian. We illustrate the benefits of our approach in synthetic and real examples, and show how SDM achieves state-ofthe-art performance in the problem of facial feature detection. The code is available at www.humansensing.cs. cmu.edu/intraface.
****

Self-Paced Learning for Long-Term Tracking
James S. Supancic III, Deva Ramanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2379-2386
We address the problem of long-term object tracking, where the object may become occluded or leave-the-view. In this setting, we show that an accurate appearance model is considerably more effective than a strong motion model. We develop simple but effective algorithms that alternate between tracking and learning a good appearance model given a track. We show that it is crucial to learn from the "right" frames, and use the formalism of self-paced curriculum learning to automatically select such frames. We leverage techniques from object detection for learning accurate appearance-based templates, demonstrating the importance of using a large negative training set (typically not used for tracking). We describe both an offline algorithm (that processes frames in batch) and a linear-time online (i.e. causal) algorithm that approaches real-time performance. Our models significantly outperform prior art, reducing the average error on benchmark videos by a factor of 4.
****

A Machine Learning Approach for Non-blind Image Deconvolution
Christian J. Schuler, Harold Christopher Burger, Stefan Harmeling, Bernhard Scholkopf; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1067-1074
Image deconvolution is the ill-posed problem of recovering a sharp image, given a blurry one generated by a convolution. In this work, we deal with space-invariant nonblind deconvolution. Currently, the most successful methods involve a regularized inversion of the blur in Fourier domain as a first step. This step amplifies and colors the noise, and corrupts the image information. In a second (and arguably more difficult) step, one then needs to remove the colored noise, typically using a cleverly engineered algorithm. However, the methods based on this two-step approach do not properly address the fact that the image information has been corrupted. In this work, we also rely on a two-step procedure, but learn the second step on a large dataset of natural images, using a neural network. We will show that this approach outperforms the current state-ofthe-art on a large dataset of artificially blurred images. We demonstrate the practical applicability of our method in a real-world example with photographic out-of-focus blur.
****

Correlation Filters for Object Alignment

Vishnu Naresh Boddeti, Takeo Kanade, B.V.K. Vijaya Kumar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2291-2298

Alignment of 3D objects from 2D images is one of the most important and well studied problems in computer vision. A typical object alignment system consists of a landmark appearance model which is used to obtain an initial shape and a shape model which refines this initial shape by correcting the initialization errors. Since errors in landmark initialization from the appearance model propagate through the shape model, it is critical to have a robust landmark appearance model. While there has been much progress in designing sophisticated and robust shape models, there has been relatively less progress in designing robust landmark detection models. In this paper we present an efficient and robust landmark detection model which is designed specifically to minimize localization errors thereby leading to state-of-the-art object alignment performance. We demonstrate the efficacy and speed of the proposed approach on the challenging task of multi-view car alignment.

****

Voxel Cloud Connectivity Segmentation - Supervoxels for Point Clouds

Jeremie Papon, Alexey Abramov, Markus Schoeler, Florentin Worgotter; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2027-2034

Unsupervised over-segmentation of an image into regions of perceptually similar pixels, known as superpixels, is a widely used preprocessing step in segmentation algorithms. Superpixel methods reduce the number of regions that must be considered later by more computationally expensive algorithms, with a minimal loss of information. Nevertheless, as some information is inevitably lost, it is vital that superpixels not cross object boundaries, as such errors will propagate through later steps. Existing methods make use of projected color or depth information, but do not consider three dimensional geometric relationships between observed data points which can be used to prevent superpixels from crossing regions of empty space. We propose a novel over-segmentation algorithm which uses voxel relationships to produce over-segmentations which are fully consistent with the spatial geometry of the scene in three dimensional, rather than projective, space. Enforcing the constraint that segmented regions must have spatial connectivity prevents label flow across semantic object boundaries which might otherwise be violated. Additionally, as the algorithm works directly in 3D space, observations from several calibrated RGB+D cameras can be segmented jointly. Experiments on a large data set of human annotated RGB+D images demonstrate a significant reduction in occurrence of clusters crossing object boundaries, while maintaining speeds comparable to state-of-the-art 2D methods.

****

Adherent Raindrop Detection and Removal in Video

Shaodi You, Robby T. Tan, Rei Kawakami, Katsushi Ikeuchi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1035-1042

Raindrops adhered to a windscreen or window glass can significantly degrade the visibility of a scene. Detecting and removing raindrops will, therefore, benefit many computer vision applications, particularly outdoor surveillance systems and intelligent vehicle systems. In this paper, a method that automatically detects and removes adherent raindrops is introduced. The core idea is to exploit the local spatiotemporal derivatives of raindrops. First, it detects raindrops based on the motion and the intensity temporal derivatives of the input video. Second, relying on an analysis that some areas of a raindrop completely occludes the scene, yet the remaining areas occludes only partially, the method removes the two types of areas separately. For partially occluding areas, it restores them by retrieving as much as possible information of the scene, namely, by solving a blending function on the detected partially occluding areas using the temporal intensity change. For completely occluding areas, it recovers them by using a video completion technique. Experimental results using various real videos show the e ffectiveness of the proposed method.

****

Recovering Line-Networks in Images by Junction-Point Processes

Dengfeng Chai, Wolfgang Forstner, Florent Lafarge; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1894-1901

The automatic extraction of line-networks from images is a well-known computer vision issue. Appearance and shape considerations have been deeply explored in the literature to improve accuracy in presence of occlusions, shadows, and a wide variety of irrelevant objects. However most existing works have ignored the structural aspect of the problem. We present an original method which provides structurally-coherent solutions. Contrary to the pixelbased and object-based methods, our result is a graph in which each node represents either a connection or an ending in the line-network. Based on stochastic geometry, we develop a new family of point processes consisting in sampling junction-points in the input image by using a Monte Carlo mechanism. The quality of a configuration is measured by a probability density which takes into account both image consistency and shape priors. Our experiments on a variety of problems illustrate the potential of our approach in terms of accuracy, flexibility and efficiency.

****

Continuous Inference in Graphical Models with Polynomial Energies

Mathieu Salzmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1744-1751

In this paper, we tackle the problem of performing inference in graphical models whose energy is a polynomial function of continuous variables. Our energy minimization method follows a dual decomposition approach, where the global problem is split into subproblems defined over the graph cliques. The optimal solution to these subproblems is obtained by making use of a polynomial system solver. Our algorithm inherits the convergence guarantees of dual decomposition. To speed up optimization, we also introduce a variant of this algorithm based on the augmented Lagrangian method. Our experiments illustrate the diversity of computer vision problems that can be expressed with polynomial energies, and demonstrate the benefits of our approach over existing continuous inference methods.

****

Attribute-Based Detection of Unfamiliar Classes with Humans in the Loop

Catherine Wah, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 779-786

Recent work in computer vision has addressed zero-shot learning or unseen class detection, which involves categorizing objects without observing any training examples. However, these problems assume that attributes or defining characteristics of these unobserved classes are known, leveraging this information at test time to detect an unseen class. We address the more realistic problem of detecting categories that do not appear in the dataset in any form. We denote such a category as an unfamiliar class; it is neither observed at train time, nor do we possess any knowledge regarding its relationships to attributes. This problem is one that has received limited attention within the computer vision community. In this work, we propose a novel approach to the unfamiliar class detection task that builds on attribute-based classification methods, and we empirically demonstrate how classification accuracy is impacted by attribute noise and dataset "difficulty," as quantified by the separation of classes in the attribute space. We also present a method for incorporating human users to overcome deficiencies in attribute detection. We demonstrate results superior to existing methods on the challenging CUB-200-2011 dataset.

****

Locally Aligned Feature Transforms across Views

Wei Li, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3594-3601

In this paper, we propose a new approach for matching images observed in different camera views with complex cross-view transforms and apply it to person reidentification. It jointly partitions the image spaces of two camera views into different configurations according to the similarity of cross-view transforms. The v isual features of an image pair from different views are first locally aligned by being projected to a common feature space and then matched with softly assigned metrics which are locally optimized. The features optimal for recognizing identities are different from those for clustering cross-view transforms. They are jointly learned by utilizing sparsityinducing norm and information theoretical regularization. This approach can be generalized to the settings where test images are from new camera views, not the same as those in the training set. Extensive experiments are conducted on public datasets and our own dataset. Comparisons with the state-of-the-art metric learning and person re-identification methods show the superior performance of our approach.

****

Sensing and Recognizing Surface Textures Using a GelSight Sensor

Rui Li, Edward H. Adelson; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1241-1247

Sensing surface textures by tou ch is a valuable was difficult to build capability for robots. Until recently it a a compliant sensor with high sennnsitivity and high resolution. The GelSight sensor is cooompliant and offers sensitivity and resolution exceeding that of the human fingertips. This opens the possibility of measuring and recognizing highly detailed surface texxxtures. The GelSight sensor, when pressed against a surfaccce, delivers a height map. This can be treated as an image, aaand processed using the tools of visual texture analysis. W WW have devised a simple yet effective texture recognitiooon system based on local binary patterns, and enhanced it by the use of a multi-scale pyramid and a Hellinger dddistance metric. We built a database with 40 classes of taaactile textures using materials such as fabric, wood, and sannndpaper. Our system can correctly categorize materials from m this database with high accuracy. This suggests that the G GelSight sensor can be useful for material recognition by rooobots.

****

Universality of the Local Marginal Polytope

Daniel Prusa, Tomas Werner; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1738-1743

We show that solving the LP relaxation of the MAP inference problem in graphical models (also known as the minsum problem, energy minimization, or weighted constraint satisfaction) is not easier than solving any LP. More precisely, any polytope is linear-time representable by a local marginal polytope and any LP can be reduced in linear time to a linear optimization (allowing infinite weights) over a local marginal polytope.

****

Graph Matching with Anchor Nodes: A Learning Approach

Nan Hu, Raif M. Rustamov, Leonidas Guibas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2906-2913

In this paper, we consider the weighted graph matching problem with partially disclosed correspondences between a number of anchor nodes. Our construction exploits recently introduced node signatures based on graph Laplacians, namely the Laplacian family signature (LFS) on the nodes, and the pairwise heat kernel map on the edges. In this paper, without assuming an explicit form of parametric dependence nor a distance metric between node signatures, we formulate an optimization problem which incorporates the knowledge of anchor nodes. Solving this problem gives us an optimized proximity measure specific to the graphs under consideration. Using this as a first order compatibility term, we then set up an integer quadratic program (IQP) to solve for a near optimal graph matching. Our experiments demonstrate the superior performance of our approach on randomly generated graphs and on two widelyused image sequences, when compared with other existing signature and adjacency matrix based graph matching methods.

****

Blocks That Shout: Distinctive Parts for Scene Classification

Mayank Juneja, Andrea Vedaldi, C.V. Jawahar, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 923-930

The automatic discovery of distinctive parts for an object or scene class is cha llenging since it requires simultaneously to learn the part appearance and also to identify the part occurrences in images. In this paper, we propose a simple, efficient, and effective method to do so. We address this problem by learning parts incrementally, starting from a single part occurrence with an Exemplar SVM. In this manner, additional part instances are discovered and aligned reliably before being considered as training examples. We also propose entropy-rank curves as a means of evaluating the distinctiveness of parts shareable between categories and use them to select useful parts out of a set of candidates. We apply the new representation to the task of scene categorisation on the MIT Scene 67 benchmark. We show that our method can learn parts which are significantly more informative and for a fraction of the cost, compared to previous part-learning methods such as Singh et al. [28]. We also show that a well constructed bag of words or Fisher vector model can substantially outperform the previous state-ofthe-art classification performance on this data.

****

Megastereo: Constructing High-Resolution Stereo Panoramas

Christian Richardt, Yael Pritch, Henning Zimmer, Alexander Sorkine-Hornung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1256-1263

We present a solution for generating high-quality stereo panoramas at megapixel resolutions. While previous approaches introduced the basic principles, we show that those techniques do not generalise well to today's high image resolutions and lead to disturbing visual artefacts. As our first contribution, we describe the necessary correction steps and a compact representation for the input images in order to achieve a highly accurate approximation to the required ray space. Our second contribution is a flow-based upsampling of the available input rays which effectively resolves known aliasing issues like stitching artefacts. The required rays are generated on the fly to perfectly match the desired output resolution, even for small numbers of input images. In addition, the upsampling is real-time and enables direct interactive control over the desired stereoscopic depth effect. In combination, our contributions allow the generation of stereoscopic panoramas at high output resolutions that are virtually free of artefacts such as seams, stereo discontinuities, vertical parallax and other mono-/stereoscopic shape distortions. Our process is robust, and other types of multiperspective panoramas, such as linear panoramas, can also benefit from our contributions. We show various comparisons and high-resolution results.

****

Augmenting Bag-of-Words: Data-Driven Discovery of Temporal and Structural Information for Activity Recognition

Vinay Bettadapura, Grant Schindler, Thomas Ploetz, Irfan Essa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2619-2626

We present data-driven techniques to augment Bag of Words (BoW) models, which allow for more robust modeling and recognition of complex long-term activities, especially when the structure and topology of the activities are not known a priori. Our approach specifically addresses the limitations of standard BoW approaches, which fail to represent the underlying temporal and causal information that is inherent in activity streams. In addition, we also propose the use of randomly sampled regular expressions to discover and encode patterns in activities. We demonstrate the effectiveness of our approach in experimental evaluations where we successfully recognize activities and detect anomalies in four complex datasets.

****

Dense 3D Reconstruction from Severely Blurred Images Using a Single Moving Camera

Hee Seok Lee, Kuoung Mu Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 273-280

Motion blur frequently occurs in dense 3D reconstruction using a single moving camera, and it degrades the quality of the 3D reconstruction. To handle motion blur caused by rapid camera shakes, we propose a blur-aware depth reconstruction m ethod, which utilizes a pixel correspondence that is obtained by considering the effect of motion blur. Motion blur is dependent on 3D geometry, thus parameterizing blurred appearance of images with scene depth given camera motion is possible and a depth map can be accurately estimated from the blur-considered pixel correspondence. The estimated depth is then converted into pixel-wise blur kernels, and non-uniform motion blur is easily removed with low computational cost. The obtained blur kernel is depth-dependent, thus it effectively addresses scene-depth variation, which is a challenging problem in conventional non-uniform deblurring methods.

****

A Practical Rank-Constrained Eight-Point Algorithm for Fundamental Matrix Estimation

Yinqiang Zheng, Shigeki Sugimoto, Masatoshi Okutomi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1546-1553

Due to its simplicity, the eight-point algorithm has been widely used in fundamental matrix estimation. Unfortunately, the rank-2 constraint of a fundamental matrix is enforced via a posterior rank correction step, thus leading to non-optimal solutions to the original problem. To address this drawback, existing algorithms need to solve either a very high order polynomial or a sequence of convex relaxation problems, both of which are computationally ineffective and numerically unstable. In this work, we present a new rank-2 constrained eight-point algorithm, which directly incorporates the rank-2 constraint in the minimization process. To avoid singularities, we propose to solve seven subproblems and retrieve their globally optimal solutions by using tailored polynomial system solvers. Our proposed method is noniterative, computationally efficient and numerically stable. Experiment results have verified its superiority over existing algebraic error based algorithms in terms of accuracy, as well as its advantages when used to initialize geometric error based algorithms.

****

Accurate Localization of 3D Objects from RGB-D Data Using Segmentation Hypotheses

Byung-soo Kim, Shili Xu, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3182-3189

In this paper we focus on the problem of detecting objects in 3D from RGB-D images. We propose a novel framework that explores the compatibility between segmentation hypotheses of the object in the image and the corresponding 3D map. Our framework allows to discover the optimal location of the object using a generalization of the structural latent SVM formulation in 3D as well as the definition of a new loss function defined over the 3D space in training. We evaluate our method using two existing RGB-D datasets. Extensive quantitative and qualitative experimental results show that our proposed approach outperforms state-of-theart as methods well as a number of baseline approaches for both 3D and 2D object recognition tasks.

****

Pixel-Level Hand Detection in Ego-centric Videos

Cheng Li, Kris M. Kitani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3570-3577

We address the task of pixel-level hand detection in the context of ego-centric cameras. Extracting hand regions in ego-centric videos is a critical step for understanding handobject manipulation and analyzing hand-eye coordination. However, in contrast to traditional applications of hand detection, such as gesture interfaces or sign-language recognition, ego-centric videos present new challenges such as rapid changes in illuminations, significant camera motion and complex hand-object manipulations. To quantify the challenges and performance in this new domain, we present a fully labeled indoor/outdoor ego-centric hand detection benchmark dataset containing over 200 million labeled pixels, which contains hand images taken under various illumination conditions. Using both our dataset and a publicly available ego-centric indoors dataset, we give extensive analysis of detection performance using a wide range of local appearance features. Our analysis highlights the effectiveness of sparse features and the importance of modeling

global illumination. We propose a modeling strategy based on our findings and show that our model outperforms several baseline approaches.
****

Geometric Context from Videos
S. Hussain Raza, Matthias Grundmann, Irfan Essa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3081-3088
We present a novel algorithm for estimating the broad 3D geometric structure of outdoor video scenes. Leveraging spatio-temporal video segmentation, we decompose a dynamic scene captured by a video into geometric classes, based on predictions made by region-classifiers that are trained on appearance and motion features. By examining the homogeneity of the prediction, we combine predictions across multiple segmentation hierarchy levels alleviating the need to determine the granularity a priori. We built a novel, extensive dataset on geometric context of video to evaluate our method, consisting of over 100 groundtruth annotated outdoor videos with over 20,000 frames. To further scale beyond this dataset, we propose a semisupervised learning framework to expand the pool of labeled data with high confidence predictions obtained from unlabeled data. Our system produces an accurate prediction of geometric context of video achieving 96% accuracy across main geometric classes.
****

Exploiting the Power of Stereo Confidences
David Pfeiffer, Stefan Gehrig, Nicolai Schneider; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 297-304
Applications based on stereo vision are becoming increasingly common, ranging from gaming over robotics to driver assistance. While stereo algorithms have been investigated heavily both on the pixel and the application level, far less attention has been dedicated to the use of stereo confidence cues. Mostly, a threshold is applied to the confidence values for further processing, which is essentially a sparsified disparity map. This is straightforward but it does not take full advantage of the available information. In this paper, we make full use of the stereo confidence cues by propagating all confidence values along with the measured disparities in a Bayesian manner. Before using this information, a mapping from confidence values to disparity outlier probability rate is performed based on gathered disparity statistics from labeled video data. We present an extension of the so called Stixel World, a generic 3D intermediate representation that can serve as input for many of the applications mentioned above. This scheme is modified to directly exploit stereo confidence cues in the underlying sensor model during a maximum a posteriori estimation process. The effectiveness of this step is verified in an in-depth evaluation on a large real-world traffic data base of which parts are made publicly available. We show that using stereo confidence cues allows both reducing the number of false object detections by a factor of six while keeping the detection rate at a near constant level.
****

Optimizing 1-Nearest Prototype Classifiers
Paul Wohlhart, Martin Kostinger, Michael Donoser, Peter M. Roth, Horst Bischof; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 460-467
The development of complex, powerful classifiers and their constant improvement have contributed much to the progress in many fields of computer vision. However, the trend towards large scale datasets revived the interest in simpler classifiers to reduce runtime. Simple nearest neighbor classifiers have several beneficial properties, such as low complexity and inherent multi-class handling, however, they have a runtime linear in the size of the database. Recent related work represents data samples by assigning them to a set of prototypes that partition the input feature space and afterwards applies linear classifiers on top of this representation to approximate decision boundaries locally linear. In this paper, we go a step beyond these approaches and purely focus on 1-nearest prototype classification, where we propose a novel algorithm for deriving optimal prototypes in a discriminative manner from the training samples. Our method is implicitly multi-class capable, parameter free, avoids noise overfitting and, since during

testing only comparisons to the derived prototypes are required, highly efficient. Experiments demonstrate that we are able to outperform related locally linear methods, while even getting close to the results of more complex classifiers.

****

Efficient 3D Endfiring TRUS Prostate Segmentation with Globally Optimized Rotational Symmetry

Jing Yuan, Wu Qiu, Eranga Ukwatta, Martin Rajchl, Xue-Cheng Tai, Aaron Fenster; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2211-2218

Segmenting 3D endfiring transrectal ultrasound (TRUS) prostate images efficiently and accurately is of utmost importance for the planning and guiding 3D TRUS guided prostate biopsy. Poor image quality and imaging artifacts of 3D TRUS images often introduce a challenging task in computation to directly extract the 3D prostate surface. In this work, we propose a novel global optimization approach to delineate 3D prostate boundaries using its rotational resliced images around a specified axis, which properly enforces the inherent rotational symmetry of prostate shapes to jointly adjust a series of 2D slicewise segmentations in the global 3D sense. We show that the introduced challenging combinatorial optimization problem can be solved globally and exactly by means of convex relaxation. In this regard, we propose a novel coupled continuous max-flow model, which not only provides a powerful mathematical tool to analyze the proposed optimization problem but also amounts to a new and efficient duality-based algorithm. Extensive experiments demonstrate that the proposed method significantly outperforms the state-of-art methods in terms of efficiency, accuracy, reliability and less user-interactions, and reduces the execution time by a factor of 100.

****

Robust Canonical Time Warping for the Alignment of Grossly Corrupted Sequences

Yannis Panagakis, Mihalis A. Nicolaou, Stefanos Zafeiriou, Maja Pantic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 540-547

Temporal alignment of human behaviour from visual data is a very challenging problem due to a numerous reasons, including possible large temporal scale differences, inter/intra subject variability and, more importantly, due to the presence of gross errors and outliers. Gross errors are often in abundance due to incorrect localization and tracking, presence of partial occlusion etc. Furthermore, such errors rarely follow a Gaussian distribution, which is the de-facto assumption in machine learning methods. In this paper, building on recent advances on rank minimization and compressive sensing, a novel, robust to gross errors temporal alignment method is proposed. While previous approaches combine the dynamic time warping (DTW) with low-dimensional projections that maximally correlate two sequences, we aim to learn two underlying projection matrices (one for each sequence), which not only maximally correlate the sequences but, at the same time, efficiently remove the possible corruptions in any datum in the sequences. The projections are obtained by minimizing the weighted sum of nuclear and 1 norms, by solving a sequence of convex optimization problems, while the temporal alignment is found by applying the DTW in an alternating fashion. The superiority of the proposed method against the state-of-the-art time alignment methods, namely the canonical time warping and the generalized time warping, is indicated by the experimental results on both synthetic and real datasets.

****

The Generalized Laplacian Distance and Its Applications for Visual Matching

Elhanan Elboer, Michael Werman, Yacov Hel-Or; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2315-2322

The graph Laplacian operator, which originated in spectral graph theory, is commonly used for learning applications such as spectral clustering and embedding. In this paper we explore the Laplacian distance, a distance function related to the graph Laplacian, and use it for visual search. We show that previous techniques such as Matching by Tone Mapping (MTM) are particular cases of the Laplacian distance. Generalizing the Laplacian distance results in distance measures which are tolerant to various visual distortions. A novel algorithm based on linear d ecomposition makes it possible to compute these generalized distances efficiently. The proposed approach is demonstrated for tone mapping invariant, outlier robust and multimodal template matching.

****

A Sentence Is Worth a Thousand Pixels

Sanja Fidler, Abhishek Sharma, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1995-2002

We are interested in holistic scene understanding where images are accompanied with text in the form of complex sentential descriptions. We propose a holistic conditional random field model for semantic parsing which reasons jointly about which objects are present in the scene, their spatial extent as well as semantic segmentation, and employs text as well as image information as input. We automatically parse the sentences and extract objects and their relationships, and incorporate them into the model, both via potentials as well as by re-ranking candidate detections. We demonstrate the effectiveness of our approach in the challenging UIUC sentences dataset and show segmentation improvements of 12.5% over the visual only model and detection improvements of 5% AP over deformable part-based models [8].

****

Deep Convolutional Network Cascade for Facial Point Detection

Yi Sun, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3476-3483

We propose a new approach for estimation of the positions of facial keypoints with three-level carefully designed convolutional networks. At each level, the outputs of multiple networks are fused for robust and accurate estimation. Thanks to the deep structures of convolutional networks, global high-level features are extracted over the whole face region at the initialization stage, which help to locate high accuracy keypoints. There are two folds of advantage for this. First, the texture context information over the entire face is utilized to locate each keypoint. Second, since the networks are trained to predict all the keypoints simultaneously, the geometric constraints among keypoints are implicitly encoded. The method therefore can avoid local minimum caused by ambiguity and data corruption in difficult image samples due to occlusions, large pose variations, and extreme lightings. The networks at the following two levels are trained to locally refine initial predictions and their inputs are limited to small regions around the initial predictions. Several network structures critical for accurate and robust facial point detection are investigated. Extensive experiments show that our approach outperforms state-ofthe-art methods in both detection accuracy and reliability 1 .

****

Scalable Sparse Subspace Clustering

Xi Peng, Lei Zhang, Zhang Yi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 430-437

In this paper, we address two problems in Sparse Subspace Clustering algorithm (SSC), i.e., scalability issue and out-of-sample problem. SSC constructs a sparse similarity graph for spectral clustering by using sp-minimization based coefficients, has achieved state-of-the-art results for image clustering and motion segmentation. However, the time complexity of SSC is proportion to the cubic of problem size such that it is inefficient to apply SSC into large scale setting. Moreover, SSC does not handle with out-ofsample data that are not used to construct the similarity graph. For each new datum, SSC needs recalculating the cluster membership of the whole data set, which makes SSC is not competitive in fast online clustering. To address the problems, this paper proposes out-of-sample extension of SSC, named as Scalable Sparse Subspace Clustering (SSSC), which makes SSC feasible to cluster large scale data sets. The solution of SSSC adopts a "sampling, clustering, coding, and classifying" strategy. Extensive experimental results on several popular data sets demonstrate the effectiveness and efficiency of our method comparing with the state-of-the-art algorithms.

****

Nonlinearly Constrained MRFs: Exploring the Intrinsic Dimensions of Higher-Order

Cliques

Yun Zeng, Chaohui Wang, Stefano Soatto, Shing-Tung Yau; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1706-1713

This paper introduces an efficient approach to integrating non-local statistics into the higher-order Markov Random Fields (MRFs) framework. Motivated by the observation that many non-local statistics (e.g., shape priors, color distributions) can usually be represented by a small number of parameters, we reformulate the higher-order MRF model by introducing additional latent variables to represent the intrinsic dimensions of the higher-order cliques. The resulting new model, called NC-MRF, not only provides the flexibility in representing the configurations of higher-order cliques, but also automatically decomposes the energy function into less coupled terms, allowing us to design an efficient algorithmic framework for maximum a posteriori (MAP) inference. Based on this novel modeling/inference framework, we achieve state-of-the-art solutions to the challenging problems of class-specific image segmentation and template-based 3D facial expression tracking, which demonstrate the potential of our approach.

****

Seeking the Strongest Rigid Detector

Rodrigo Benenson, Markus Mathias, Tinne Tuytelaars, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3666-3673

The current state of the art solutions for object detection describe each class by a set of models trained on discovered sub-classes (so called "components"), with each model itself composed of collections of interrelated parts (deformable models). These detectors build upon the now classic Histogram of Oriented Gradients+linear SVM combo. In this paper we revisit some of the core assumptions in HOG+SVM and show that by properly designing the feature pooling, feature selection, preprocessing, and training methods, it is possible to reach top quality, at least for pedestrian detections, using a single rigid component. We provide experiments for a large design space, that give insights into the design of classifiers, as well as relevant information for practitioners. Our best detector is fully feed-forward, has a single unified architecture, uses only histograms of oriented gradients and colour information in monocular static images, and improves over 23 other methods on the INRIA, ETH and Caltech-USA datasets, reducing the average miss-rate over HOG+SVM by more than 30%.

****

An Approach to Pose-Based Action Recognition

Chunyu Wang, Yizhou Wang, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 915-922

We address action recognition in videos by modeling the spatial-temporal structures of human poses. We start by improving a state of the art method for estimating human joint locations from videos. More precisely, we obtain the K-best estimations output by the existing method and incorporate additional segmentation cues and temporal constraints to select the "best" one. Then we group the estimated joints into five body parts (e.g. the left arm) and apply data mining techniques to obtain a representation for the spatial-temporal structures of human actions. This representation captures the spatial configurations of body parts in one frame (by spatial-part-sets) as well as the body part movements(by temporal-part-sets) which are characteristic of human actions. It is interpretable, compact, and also robust to errors on joint estimations. Experimental results first show that our approach is able to localize body joints more accurately than existing methods. Next we show that it outperforms state of the art action recognizers on the UCF sport, the Keck Gesture and the MSR-Action3D datasets.

****

Pattern-Driven Colorization of 3D Surfaces

George Leifman, Ayellet Tal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 241-248

Colorization refers to the process of adding color to black & white images or videos. This paper extends the term to handle surfaces in three dimensions. This i s important for applications in which the colors of an object need to be restored and no relevant image exists for texturing it. We focus on surfaces with patterns and propose a novel algorithm for adding colors to these surfaces. The user needs only to scribble a few color strokes on one instance of each pattern, and the system proceeds to automatically colorize the whole surface. For this scheme to work, we address not only the problem of colorization, but also the problem of pattern detection on surfaces.

****

Dense Reconstruction Using 3D Object Shape Priors

Amaury Dame, Victor A. Prisacariu, Carl Y. Ren, Ian Reid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1288-1295

We propose a formulation of monocular SLAM which combines live dense reconstruction with shape priors-based 3D tracking and reconstruction. Current live dense SLAM approaches are limited to the reconstruction of visible surfaces. Moreover, most of them are based on the minimisation of a photo-consistency error, which usually makes them sensitive to specularities. In the 3D pose recovery literature, problems caused by imperfect and ambiguous image information have been dealt with by using prior shape knowledge. At the same time, the success of depth sensors has shown that combining joint image and depth information drastically increases the robustness of the classical monocular 3D tracking and 3D reconstruction approaches. In this work we link dense SLAM to 3D object pose and shape recovery. More specifically, we automatically augment our SLAM system with object specific identity, together with 6D pose and additional shape degrees of freedom for the object(s) of known class in the scene, combining image data and depth information for the pose and shape recovery. This leads to a system that allows for full scaled 3D reconstruction with the known object(s) segmented from the scene. The segmentation enhances the clarity, accuracy and completeness of the maps built by the dense SLAM system, while the dense 3D data aids the segmentation process, yielding faster and more reliable convergence than when using 2D image data alone.

****

Modeling Actions through State Changes

Alireza Fathi, James M. Rehg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2579-2586

In this paper we present a model of action based on the change in the state of the environment. Many actions involve similar dynamics and hand-object relationships, but differ in their purpose and meaning. The key to differentiating these actions is the ability to identify how they change the state of objects and materials in the environment. We propose a weakly supervised method for learning the object and material states that are necessary for recognizing daily actions. Once these state detectors are learned, we can apply them to input videos and pool their outputs to detect actions. We further demonstrate that our method can be used to segment discrete actions from a continuous video of an activity. Our results outperform state-of-the-art action recognition and activity segmentation results.

****

GRASP Recurring Patterns from a Single View

Jingchen Liu, Yanxi Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2003-2010

We propose a novel unsupervised method for discovering recurring patterns from a single view. A key contribution of our approach is the formulation and validation of a joint assignment optimization problem where multiple visual words and object instances of a potential recurring pattern are considered simultaneously. The optimization is achieved by a greedy randomized adaptive search procedure (GRASP) with moves specifically designed for fast convergence. We have quantified systematically the performance of our approach under stressed conditions of the input (missing features, geometric distortions). We demonstrate that our proposed algorithm outperforms state of the art methods for recurring pattern discovery on a diverse set of 400+ real world and synthesized test images.

****

Texture Enhanced Image Denoising via Gradient Histogram Preservation

Wangmeng Zuo, Lei Zhang, Chunwei Song, David Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1203-1210

Image denoising is a classical yet fundamental problem in low level vision, as well as an ideal test bed to evaluate various statistical image modeling methods. One of the most challenging problems in image denoising is how to preserve the fine scale texture structures while removing noise. Various natural image priors, such as gradient based prior, nonlocal self-similarity prior, and sparsity prior, have been extensively exploited for noise removal. The denoising algorithms based on these priors, however, tend to smooth the detailed image textures, degrading the image visual quality. To address this problem, in this paper we propose a texture enhanced image denoising (TEID) method by enforcing the gradient distribution of the denoised image to be close to the estimated gradient distribution of the original image. A novel gradient histogram preservation (GHP) algorithm is developed to enhance the texture structures while removing noise. Our experimental results demonstrate that the proposed GHP based TEID can well preserve the texture features of the denoised images, making them look more natural.

****

Analyzing Semantic Segmentation Using Hybrid Human-Machine CRFs

Roozbeh Mottaghi, Sanja Fidler, Jian Yao, Raquel Urtasun, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3143-3150

Recent trends in semantic image segmentation have pushed for holistic scene understanding models that jointly reason about various tasks such as object detection, scene recognition, shape analysis, contextual reasoning. In this work, we are interested in understanding the roles of these different tasks in aiding semantic segmentation. Towards this goal, we "plug-in" human subjects for each of the various components in a state-of-the-art conditional random field model (CRF) on the MSRC dataset. Comparisons among various hybrid human-machine CRFs give us indications of how much "head room" there is to improve segmentation by focusing research efforts on each of the tasks. One of the interesting findings from our slew of studies was that human classification of isolated super-pixels, while being worse than current machine classifiers, provides a significant boost in performance when plugged into the CRF! Fascinated by this finding, we conducted in depth analysis of the human generated potentials. This inspired a new machine potential which significantly improves state-of-the-art performance on the MRSC dataset.

****

Multi-source Multi-scale Counting in Extremely Dense Crowd Images

Haroon Idrees, Imran Saleemi, Cody Seibert, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2547-2554

We propose to leverage multiple sources of information to compute an estimate of the number of individuals present in an extremely dense crowd visible in a single image. Due to problems including perspective, occlusion, clutter, and few pixels per person, counting by human detection in such images is almost impossible. Instead, our approach relies on multiple sources such as low confidence head detections, repetition of texture elements (using SIFT), and frequency-domain analysis to estimate counts, along with confidence associated with observing individuals, in an image region. Secondly, we employ a global consistency constraint on counts using Markov Random Field. This caters for disparity in counts in local neighborhoods and across scales. We tested our approach on a new dataset of fifty crowd images containing 64K annotated humans, with the head counts ranging from 94 to 4543. This is in stark contrast to datasets used for existing methods which contain not more than tens of individuals. We experimentally demonstrate the efficacy and reliability of the proposed approach by quantifying the counting performance.

****

Non-uniform Motion Deblurring for Bilayer Scenes

Chandramouli Paramanand, Ambasamudram N. Rajagopalan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1115-1122

We address the problem of estimating the latent image of a static bilayer scene (consisting of a foreground and a background at different depths) from motion blurred observations captured with a handheld camera. The camera motion is considered to be composed of in-plane rotations and translations. Since the blur at an image location depends both on camera motion and depth, deblurring becomes a difficult task. We initially propose a method to estimate the transformation spread function (TSF) corresponding to one of the depth layers. The estimated TSF (which reveals the camera motion during exposure) is used to segment the scene into the foreground and background layers and determine the relative depth value. The deblurred image of the scene is finally estimated within a regularization framework by accounting for blur variations due to camera motion as well as depth.

****

Specular Reflection Separation Using Dark Channel Prior

Hyeongwoo Kim, Hailin Jin, Sunil Hadap, Inso Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1460-1467

We present a novel method to separate specular reflection from a single image. Separating an image into diffuse and specular components is an ill-posed problem due to lack of observations. Existing methods rely on a specularfree image to detect and estimate specularity, which however may confuse diffuse pixels with the same hue but a different saturation value as specular pixels. Our method is based on a novel observation that for most natural images the dark channel can provide an approximate specular-free image. We also propose a maximum a posteriori formulation which robustly recovers the specular reflection and chromaticity despite of the hue-saturation ambiguity. We demonstrate the effectiveness of the proposed algorithm on real and synthetic examples. Experimental results show that our method significantly outperforms the state-of-theart methods in separating specular reflection.

****

Blessing of Dimensionality: High-Dimensional Feature and Its Efficient Compression for Face Verification

Dong Chen, Xudong Cao, Fang Wen, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3025-3032

Making a high-dimensional (e.g., 100K-dim) feature for face recognition seems not a good idea because it will bring difficulties on consequent training, computation, and storage. This prevents further exploration of the use of a highdimensional feature. In this paper, we study the performance of a highdimensional feature. We first empirically show that high dimensionality is critical to high performance. A 100K-dim feature, based on a single-type Local Binary Pattern (LBP) descriptor, can achieve significant improvements over both its low-dimensional version and the state-of-the-art. We also make the high-dimensional feature practical. With our proposed sparse projection method, named rotated sparse regression, both computation and model storage can be reduced by over 100 times without sacrificing accuracy quality.

****

Robust Estimation of Nonrigid Transformation for Point Set Registration

Jiayi Ma, Ji Zhao, Jinwen Tian, Zhuowen Tu, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2147-2154

We present a new point matching algorithm for robust nonrigid registration. The method iteratively recovers the point correspondence and estimates the transformation between two point sets. In the first step of the iteration, feature descriptors such as shape context are used to establish rough correspondence. In the second step, we estimate the transformation using a robust estimator called L 2 E. This is the main novelty of our approach and it enables us to deal with the noise and outliers which arise in the correspondence step. The transformation is specified in a functional space, more specifically a reproducing kernel Hilbert space. We apply our method to nonrigid sparse image feature correspondence on 2D images and 3D surfaces. Our results quantitatively show that our approach outper forms state-ofthe-art methods, particularly when there are a large number of outliers. Moreover, our method of robustly estimating transformations from correspondences is general and has many other applications.

****

Representing Videos Using Mid-level Discriminative Patches
Arpit Jain, Abhinav Gupta, Mikel Rodriguez, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2571-2578

representation for videos based on mid-level discriminative spatio-temporal patches. These spatio-temporal patches might correspond to a primitive human action, a semantic object, or perhaps a random but informative spatiotemporal patch in the video. What defines these spatiotemporal patches is their discriminative and representative properties. We automatically mine these patches from hundreds of training videos and experimentally demonstrate that these patches establish correspondence across videos and align the videos for label transfer techniques. Furthermore, these patches can be used as a discriminative vocabulary for action classification where they demonstrate stateof-the-art performance on UCF50 and Olympics datasets.

****

Fusing Robust Face Region Descriptors via Multiple Metric Learning for Face Recognition in the Wild
Zhen Cui, Wen Li, Dong Xu, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3554-3561

In many real-world face recognition scenarios, face images can hardly be aligned accurately due to complex appearance variations or low-quality images. To address this issue, we propose a new approach to extract robust face region descriptors. Specifically, we divide each image (resp. video) into several spatial blocks (resp. spatial-temporal volumes) and then represent each block (resp. volume) by sum-pooling the nonnegative sparse codes of position-free patches sampled within the block (resp. volume). Whitened Principal Component Analysis (WPCA) is further utilized to reduce the feature dimension, which leads to our Spatial Face Region Descriptor (SFRD) (resp. Spatial-Temporal Face Region Descriptor, STFRD) for images (resp. videos). Moreover, we develop a new distance metric learning method for face verification called Pairwise-constrained Multiple Metric Learning (PMML) to effectively integrate the face region descriptors of all blocks (resp. volumes) from an image (resp. a video). Our work achieves the stateof-the-art performances on two real-world datasets LFW and YouTube Faces (YTF) according to the restricted protocol.

****

Discriminative Sub-categorization
Minh Hoai, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1666-1673

The objective of this work is to learn sub-categories. Rather than casting this as a problem of unsupervised clustering, we investigate a weakly supervised approach using both positive and negative samples of the category. We make the following contributions: (i) we introduce a new model for discriminative sub-categorization which determines cluster membership for positive samples whilst simultaneously learning a max-margin classifier to separate each cluster from the negative samples; (ii) we show that this model does not suffer from the degenerate cluster problem that afflicts several competing methods (e.g., Latent SVM and Max-Margin Clustering); (iii) we show that the method is able to discover interpretable sub-categories in various datasets. The model is evaluated experimentally over various datasets, and its performance advantages over k-means and Latent SVM are demonstrated. We also stress test the model and show its resilience in discovering sub-categories as the parameters are varied.

****

Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images
Saurabh Gupta, Pablo Arbelaez, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 564-571

We address the problems of contour detection, bottomup grouping and semantic seg mentation using RGB-D data. We focus on the challenging setting of cluttered indoor scenes, and evaluate our approach on the recently introduced NYU-Depth V2 (NYUD2) dataset [27]. We propose algorithms for object boundary detection and hierarchical segmentation that generalize the gP b ucm approach of [2] by making effective use of depth information. We show that our system can label each contour with its type (depth, normal or albedo). We also propose a generic method for long-range amodal completion of surfaces and show its effectiveness in grouping. We then turn to the problem of semantic segmentation and propose a simple approach that classifies superpixels into the 40 dominant object categories in NYUD2. We use both generic and class-specific features to encode the appearance and geometry of objects. We also show how our approach can be used for scene classification, and how this contextual information in turn improves object recognition. In all of these tasks, we report significant improvements over the state-of-the-art.

****

Harvesting Mid-level Visual Concepts from Large-Scale Internet Images

Quannan Li, Jiajun Wu, Zhuowen Tu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 851-858

Obtaining effective mid-level representations has become an increasingly important task in computer vision. In this paper, we propose a fully automatic algorithm which harvests visual concepts from a large number of Internet images (more than a quarter of a million) using text-based queries. Existing approaches to visual concept learning from Internet images either rely on strong supervision with detailed manual annotations or learn image-level classifiers only. Here, we take the advantage of having massive wellorganized Google and Bing image data; visual concepts (around 14, 000) are automatically exploited from images using word-based queries. Using the learned visual concepts, we show state-of-the-art performances on a variety of benchmark datasets, which demonstrate the effectiveness of the learned mid-level representations: being able to generalize well to general natural images. Our method shows significant improvement over the competing systems in image classification, including those with strong supervision.

****

Sample-Specific Late Fusion for Visual Category Recognition

Dong Liu, Kuan-Ting Lai, Guangnan Ye, Ming-Syan Chen, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 803-810

Late fusion addresses the problem of combining the prediction scores of multiple classifiers, in which each score is predicted by a classifier trained with a specific feature. However, the existing methods generally use a fixed fusion weight for all the scores of a classifier, and thus fail to optimally determine the fusion weight for the individual samples. In this paper, we propose a sample-specific late fusion method to address this issue. Specifically, we cast the problem into an information propagation process which propagates the fusion weights learned on the labeled samples to individual unlabeled samples, while enforcing that positive samples have higher fusion scores than negative samples. In this process, we identify the optimal fusion weights for each sample and push positive samples to top positions in the fusion score rank list. We formulate our problem as a L ? norm constrained optimization problem and apply the Alternating Direction Method of Multipliers for the optimization. Extensive experiment results on various visual categorization tasks show that the proposed method consistently and significantly beats the state-of-the-art late fusion methods. To the best knowledge, this is the first method supporting sample-specific fusion weight learning.

****

PISA: Pixelwise Image Saliency by Aggregating Complementary Appearance Contrast Measures with Spatial Priors

Keyang Shi, Keze Wang, Jiangbo Lu, Liang Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2115-2122

Driven by recent vision and graphics applications such as image segmentation and object recognition, assigning pixel-accurate saliency values to uniformly highl ight foreground objects becomes increasingly critical. More often, such fine-grained saliency detection is also desired to have a fast runtime. Motivated by these, we propose a generic and fast computational framework called PISA Pixelwise Image Saliency Aggregating complementary saliency cues based on color and structure contrasts with spatial priors holistically. Overcoming the limitations of previous methods often using homogeneous superpixel-based and color contrast-only treatment, our PISA approach directly performs saliency modeling for each individual pixel and makes use of densely overlapping, feature-adaptive observations for saliency measure computation. We further impose a spatial prior term on each of the two contrast measures, which constrains pixels rendered salient to be compact and also centered in image domain. By fusing complementary contrast measures in such a pixelwise adaptive manner, the detection effectiveness is significantly boosted. Without requiring reliable region segmentation or postrelaxation, PISA exploits an efficient edge-aware image representation and filtering technique and produces spatially coherent yet detail-preserving saliency maps. Extensive experiments on three public datasets demonstrate PISA's superior detection accuracy and competitive runtime speed over the state-of-the-arts approaches.

****

Simultaneous Super-Resolution of Depth and Images Using a Single Camera

Hee Seok Lee, Kuoung Mu Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 281-288

In this paper, we propose a convex optimization framework for simultaneous estimation of super-resolved depth map and images from a single moving camera. The pixel measurement error in 3D reconstruction is directly related to the resolution of the images at hand. In turn, even a small measurement error can cause significant errors in reconstructing 3D scene structure or camera pose. Therefore, enhancing image resolution can be an effective solution for securing the accuracy as well as the resolution of 3D reconstruction. In the proposed method, depth map estimation and image super-resolution are formulated in a single energy minimization framework with a convex function and solved efficiently by a first-order primal-dual algorithm. Explicit inter-frame pixel correspondences are not required for our super-resolution procedure, thus we can avoid a huge computation time and obtain improved depth map in the accuracy and resolution as well as highresolution images with reasonable time. The superiority of our algorithm is demonstrated by presenting the improved depth map accuracy, image super-resolution results, and camera pose estimation.

****

Learning Structured Hough Voting for Joint Object Detection and Occlusion Reasoning

Tao Wang, Xuming He, Nick Barnes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1790-1797

We propose a structured Hough voting method for detecting objects with heavy occlusion in indoor environments. First, we extend the Hough hypothesis space to include both object location and its visibility pattern, and design a new score function that accumulates votes for object detection and occlusion prediction. In addition, we explore the correlation between objects and their environment, building a depth-encoded object-context model based on RGB-D data. Particularly, we design a layered context representation and allow image patches from both objects and backgrounds voting for the object hypotheses. We demonstrate that using a data-driven 2.1D representation we can learn visual codebooks with better quality, and more interpretable detection results in terms of spatial relationship between objects and viewer. We test our algorithm on two challenging RGB-D datasets with significant occlusion and intraclass variation, and demonstrate the superior performance of our method.

****

3D-Based Reasoning with Blocks, Support, and Stability

Zhaoyin Jia, Andrew Gallagher, Ashutosh Saxena, Tsuhan Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1-8

3D volumetric reasoning is important for truly understanding a scene. Humans are able to both segment each object in an image, and perceive a rich 3D interpreta tion of the scene, e.g., the space an object occupies, which objects support other objects, and which objects would, if moved, cause other objects to fall. We propose a new approach for parsing RGB-D images using 3D block units for volumetric reasoning. The algorithm fits image segments with 3D blocks, and iteratively evaluates the scene based on block interaction properties. We produce a 3D representation of the scene based on jointly optimizing over segmentations, block fitting, supporting relations, and object stability. Our algorithm incorporates the intuition that a good 3D representation of the scene is the one that fits the data well, and is a stable, self-supporting (i.e., one that does not topple) arrangement of objects. We experiment on several datasets including controlled and real indoor scenarios. Results show that our stability-reasoning framework improves RGB-D segmentation and scene volumetric representation.

****

Sampling Strategies for Real-Time Action Recognition

Feng Shi, Emil Petriu, Robert Laganiere; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2595-2602

Local spatio-temporal features and bag-of-features representations have become popular for action recognition. A recent trend is to use dense sampling for better performance. While many methods claimed to use dense feature sets, most of them are just denser than approaches based on sparse interest point detectors. In this paper, we explore sampling with high density on action recognition. We also investigate the impact of random sampling over dense grid for computational efficiency. We present a real-time action recognition system which integrates fast random sampling method with local spatio-temporal features extracted from a Local Part Model. A new method based on histogram intersection kernel is proposed to combine multiple channels of different descriptors. Our technique shows high accuracy on the simple KTH dataset, and achieves state-of-the-art on two very challenging real-world datasets, namely, 93% on KTH, 83.3% on UCF50 and 47.6% on HMDB51.

****

SCaLE: Supervised and Cascaded Laplacian Eigenmaps for Visual Object Recognition Based on Nearest Neighbors

Ruobing Wu, Yizhou Yu, Wenping Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 867-874

Recognizing the category of a visual object remains a challenging computer vision problem. In this paper we develop a novel deep learning method that facilitates examplebased visual object category recognition. Our deep learning architecture consists of multiple stacked layers and computes an intermediate representation that can be fed to a nearest-neighbor classifier. This intermediate representation is discriminative and structure-preserving. It is also capable of extracting essential characteristics shared by objects in the same category while filtering out nonessential differences among them. Each layer in our model is a nonlinear mapping, whose parameters are learned through two sequential steps that are designed to achieve the aforementioned properties. The first step computes a discrete mapping called supervised Laplacian Eigenmap. The second step computes a continuous mapping from the discrete version through nonlinear regression. We have extensively tested our method and it achieves state-of-the-art recognition rates on a number of benchmark datasets.

****