ECCV 2012

12TH EUROPEAN CONFERENCE
ON COMPUTER VISION

Florence, Italy, 7-13 October 2012

Abstracts

PREFACE

The booklet that is now in your hands contains, in chronological order, all the abstracts of the ECCV 2012 main conference papers (40 orals, 368 posters). It was conceived as a practical resource by which you can have a quick glance at the contents of each technical session before it starts, and select the papers most appealing to you. The booklet is meant to be a physical complement to the web resources for laptop, tablet and smartphone also implemented for this edition. In particular, we believe that selecting in advance which posters to see will help you to make the best out of each poster session.

The eight poster sessions of ECCV 2012 have no titles, and are all alike, in the sense that they include papers from all the main thematic areas of computer vision. The rationale behind this choice was to uniformly distribute paper topics—and hence audience interest!—across all poster sessions, thus keeping the number of must-see papers for the general attendee approximately constant throughout the conference. Although topic distribution inside each session is uniform, paper order is by no means chaotic. Indeed, posters on the same subject are grouped together, both in the proceedings volumes and inside conference sessions, so as to be displayed close to each other. The thematic areas used for the grouping were: (1) Geometry, Shape and Reconstruction, (2) Recognition and Classification, (3) Features and Matching, (4) Action and Activities, (5) Models, Optimization and Learning, (6) Tracking and Registration, (7) Lighting and Color, (8) Segmentation. According to this scheme, the first posters of each session are on geometry, followed by posters on recognition, etc.

We hope that you'll find this booklet a pleasant and useful tool for exploring the main conference technical program. Enjoy!

Roberto Cipolla, Carlo Colombo and Alberto Del Bimbo
ECCV 2012 General Chairs

TABLE OF CONTENTS

Paper coding. The code beside each paper title should be read as follows: Sk-(O/P)n[A/B], where k is the session number, n is the paper number (O=oral, P=poster) inside the session, and A and B are the rooms where posters are displayed (see the companion Main Conference booklet for the location of poster rooms). Thus S3-P11B is the eleventh poster of session 3 in room B, while S5-O4 is the fourth oral paper of session 5.

ORAL SESSION 1
GEOMETRY: THEORY AND APPLICATION

Monday, October 8
10:05 - 10:35

[S1-O1]

A QCQP Approach to Triangulation

Chris Aholt, Sameer Agarwal, and Rekha Thomas

Triangulation of a three-dimensional point from $n \geq 2$ two-dimensional images can be formulated as a quadratically constrained quadratic program. We propose an algorithm to extract candidate solutions to this problem from its semidefinite programming relaxations. We then describe a sufficient condition and a polynomial time test for certifying when such a solution is optimal. This test has no false positives. Experiments indicate that false negatives are rare, and the algorithm has excellent performance in practice. We explain this phenomenon in terms of the geometry of the triangulation problem.

[S1-O2]

Reconstructing the World's Museums

Jianxiong Xiao and Yasutaka Furukawa

Photorealistic maps are a useful navigational guide for large indoor environments, such as museums and businesses. However, it is impossible to acquire photographs covering a large indoor environment from aerial viewpoints. This paper presents a 3D reconstruction and visualization system to automatically produce clean and well-regularized texture-mapped 3D models for large indoor scenes, from ground-level photographs and 3D laser points. The key component is a new algorithm called "Inverse CSG" for reconstructing a scene in a Constructive Solid Geometry (CSG) representation consisting of volumetric primitives, which imposes powerful regularization constraints to exploit structural regularities. We also propose several techniques to adjust the 3D model to make it suitable for rendering the 3D maps from aerial viewpoints. The visualization system enables users to easily browse a large scale indoor environment from a bird's-eye view, locate specific room interiors, fly into a place of interest, view immersive ground-level panorama views, and zoom out again, all with seamless 3D transitions. We demonstrate our system on various museums, including the Metropolitan Museum of Art in New York City – one of the largest art galleries in the world.

POSTER SESSION 1

Monday, October 8
10:40 - 13:10

[S1-P1A]

Lie Bodies: A Manifold Representation of 3D Human Shape

Oren Freifeld and Michael J. Black

Three-dimensional object shape is commonly represented in terms of deformations of a triangular mesh from an exemplar shape. Existing models, however, are based on a Euclidean representation of shape deformations. In contrast, we argue that shape has a manifold structure: For example, summing the shape deformations for two people does not necessarily yield a deformation corresponding to a valid human shape, nor does the Euclidean difference of these two deformations provide a meaningful measure of shape dissimilarity. Consequently, we define a novel manifold for shape representation, with emphasis on body shapes, using a new Lie group of deformations. This has several advantages. First we define triangle deformations exactly, removing non-physical deformations and redundant degrees of freedom common to previous methods. Second, the Riemannian structure of Lie Bodies enables a more meaningful definition of body shape similarity by measuring distance between bodies on the manifold of body shape deformations. Third, the group structure allows the valid composition of deformations. This is important for models that factor body shape deformations into multiple causes or represent shape as a linear combination of basis shapes. Finally, body shape variation is modeled using statistics on manifolds. Instead of modeling Euclidean shape variation with Principal Component Analysis we capture shape variation on the manifold using Principal Geodesic Analysis. Our experiments show consistent visual and quantitative advantages of Lie Bodies over traditional Euclidean models of shape deformation and our representation can be easily incorporated into existing methods.

[S1-P2A]

Worldwide Pose Estimation Using 3D Point Clouds

Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua

We address the problem of determining where a photo was taken by estimating a full 6-DOF-plus-intrincs camera pose with respect to a large geo-registered 3D point cloud, bringing together research on image localization, landmark recognition, and 3D pose estimation. Our method scales to datasets with hundreds of thousands of images and tens of millions of 3D points through the use of two new techniques: a co-occurrence prior for RANSAC and bidirectional matching of image features with 3D points. We evaluate our method on several large data sets, and show state-of-the-art results on landmark recognition as well as the ability to locate cameras to within meters, requiring only seconds per query.

[S1-P3A]

Improved Reconstruction of Deforming Surfaces by Cancelling Ambient Occlusion

Thabo Beeler, Derek Bradley, Henning Zimmer, and Markus Gross

We present a general technique for improving space-time reconstructions of deforming surfaces, which are captured in an video-based reconstruction scenario under uniform illumination. Our approach simultaneously improves both the acquired shape as well as the tracked motion of the deforming surface. The method is based on factoring out surface shading, computed by a fast approximation to global illumination called ambient occlusion. This allows us to improve the performance of optical flow tracking that mainly relies on constancy of image features, such as intensity. While cancelling the local shading, we also optimize the surface shape to minimize the residual between the ambient occlusion of the 3D geometry and that of the image, yielding more accurate surface details in the reconstruction. Our enhancement is independent of the actual space-time reconstruction algorithm. We experimentally measure the quantitative improvements produced by our algorithm using a synthetic example of deforming skin, where ground truth shape and motion is available. We further demonstrate our enhancement on a real-world sequence of human face reconstruction.

[S1-P4A]

On the Statistical Determination of Optimal Camera Configurations in Large Scale Surveillance Networks

Junbin Liu, Clinton Fookes, Tim Wark, and Sridha Sridharan

The selection of optimal camera configurations (camera locations, orientations etc.) for multi-camera networks remains an unsolved problem. Previous approaches largely focus on proposing various objective functions to achieve different tasks. Most of them, however, do not generalize well to large scale networks. To tackle this, we introduce a statistical formulation of the optimal selection of camera configurations as well as propose a Trans-Dimensional Simulated Annealing (TDSA) algorithm to effectively solve the problem. We compare our approach with a state-of-the-art method based on Binary Integer Programming (BIP) and show that our approach offers similar performance on small scale problems. However, we also demonstrate the capability of our approach in dealing with large scale problems and show that our approach produces better results than 2 alternative heuristics designed to deal with the scalability issue of BIP.

[S1-P5A]

The Scale of Geometric Texture

Geoffrey Oxholm, Prabin Bariya, and Ko Nishino

The most defining characteristic of texture is its underlying geometry. Although the appearance of texture is as dynamic as its illumination and viewing conditions, its geometry remains constant. In this work, we study the fundamental characteristic properties of texture geometry—self similarity and scale variability—and exploit them to perform surface normal estimation, and geometric texture classification. Textures, whether they are regular or stochastic, exhibit some form of repetition in their underlying geometry. We use this property to derive a photometric stereo method uniquely tailored to utilize the redundancy in geometric texture. Using basic observations about the scale variability of texture geometry, we derive a compact, rotation invariant, scale-space representation of geometric texture. To evaluate this representation we introduce an extensive new texture database that contains multiple distances as well as in-plane and out-of plane rotations. The high accuracy of the classification results indicate the descriptive yet compact nature of our texture representation, and demonstrates the importance of geometric texture analysis, pointing the way towards improvements in appearance modeling and synthesis.

[S1-P6A]

Efficient Articulated Trajectory Reconstruction Using Dynamic Programming and Filters

Jack Valmadre, Yingying Zhu, Sridha Sridharan, and Simon Lucey

This paper considers the problem of reconstructing the motion of a 3D articulated tree from 2D point correspondences subject to some temporal prior. Hitherto, smooth motion has been encouraged using a trajectory basis, yielding a hard combinatorial problem with time complexity growing exponentially in the number of frames. Branch and bound strategies have previously attempted to curb this complexity whilst maintaining global optimality. However, they provide no guarantee of being more efficient than exhaustive search. Inspired by recent work which reconstructs general trajectories using compact high-pass filters, we develop a dynamic programming approach which scales linearly in the number of frames, leveraging the intrinsically local nature of filter interactions. Extension to affine projection enables reconstruction without estimating cameras.

[S1-P7A]

Object Co-detection

Sid Yingze Bao, Yu Xiang, and Silvio Savarese

In this paper we introduce a new problem which we call object co-detection. Given a set of images with objects observed from two or multiple images, the goal of co-detection is to detect the objects, establish the identity of individual object instance, as well as estimate the viewpoint transformation of corresponding object instances. In designing a co-detector, we follow the intuition that an object has consistent appearance when observed from the same or different viewpoints. By modeling an object using state-of-the-art part-based representations such as [1,2], we measure appearance consistency between objects by comparing part appearance and geometry across images. This allows to effectively account for object self-occlusions and viewpoint transformations. Extensive experimental evaluation indicates that our co-detector obtains more accurate detection results than if objects were to be detected from each image individually. Moreover, we demonstrate the relevance of our co-detection scheme to other recognition problems such as single instance object recognition, wide-baseline matching, and image query.

[S1-P8A]

Morphable Displacement Field Based Image Matching for Face Recognition across Pose

Shaoxin Li, Xin Liu, Xiujuan Chai, Haihong Zhang, Shihong Lao, and Shiguang Shan

Fully automatic Face Recognition Across Pose (FRAP) is one of the most desirable techniques, however, also one of the most challenging tasks in face recognition field. Matching a pair of face images in different poses can be converted into matching their pixels corresponding to the same semantic facial point. Following this idea, given two images G and P in different poses, we propose a novel method, named Morphable Displacement Field (MDF), to match G with P's virtual view under G's pose. By formulating MDF as a convex combination of a number of template displacement fields generated from a 3D face database, our model satisfies both global conformity and local consistency. We further present an approximate but effective solution of the proposed MDF model, named implicit Morphable Displacement Field (iMDF), which synthesizes virtual view implicitly via an MDF by minimizing matching residual. This formulation not only avoids intractable optimization of the high-dimensional displacement field but also facilitates a constrained quadratic optimization. The proposed method can work well even when only 2 facial landmarks are labeled, which makes it especially suitable for fully automatic FRAP system. Extensive evaluations on FERET, PIE and Multi-PIE databases show considerable improvement over state-of-the-art FRAP algorithms in both semi-automatic and fully automatic evaluation protocols.

[S1-P9A]

Combining Per-frame and Per-track Cues for Multi-person Action Recognition

Sameh Khamis, Vlad I. Morariu, and Larry S. Davis

We propose a model to combine per-frame and per-track cues for action recognition. With multiple targets in a scene, our model simultaneously captures the natural harmony of an individual's action in a scene and the flow of actions of an individual in a video sequence, inferring valid tracks in the process. Our motivation is based on the unlikely discordance of an action in a structured scene, both at the track level and the frame level (e.g., a person dancing in a crowd of joggers). While we can utilize sampling approaches for inference in our model, we instead devise a global inference algorithm by decomposing the problem and solving the subproblems exactly and efficiently, recovering a globally optimal joint solution in several cases. Finally, we improve on the state-of-the-art action recognition results for two publicly available datasets.

[S1-P10A]

Joint Image and Word Sense Discrimination for Image Retrieval

Aurelien Lucchi and Jason Weston

We study the task of learning to rank images given a text query, a problem that is complicated by the issue of multiple senses. That is, the senses of interest are typically the visually distinct concepts that a user wishes to retrieve. In this paper, we propose to learn a ranking function that optimizes the ranking cost of interest and simultaneously discovers the disambiguated senses of the query that are optimal for the supervised task. Note that no supervised information is given about the senses. Experiments performed on web images and the ImageNet dataset show that using our approach leads to a clear gain in performance.

[S1-P11A]

Script Data for Attribute-Based Recognition of Composite Activities

Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele

State-of-the-art human activity recognition methods build on discriminative learning which requires a representative training set for good performance. This leads to scalability issues for the recognition of large sets of highly diverse activities. In this paper we leverage the fact that many human activities are compositional and that the essential components of the activities can be obtained from textual descriptions or scripts. To share and transfer knowledge between composite activities we model them by a common set of attributes corresponding to basic actions and object participants. This attribute representation allows to incorporate script data that delivers new variations of a composite activity or even to unseen composite activities. In our experiments on 41 composite cooking tasks, we found that script data to successfully capture the high variability of composite activities. We show improvements in a supervised case where training data for all composite cooking tasks is available, but we are also able to recognize unseen composites by just using script data and without any manual video annotation.

[S1-P12A]

Undoing the Damage of Dataset Bias

Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and Antonio Torralba

The presence of bias in existing object recognition datasets is now well-known in the computer vision community. While it remains in question whether creating an unbiased dataset is possible given limited resources, in this work we propose a discriminative framework that directly exploits dataset bias during training. In particular, our model learns two sets of weights: (1) bias vectors associated with each individual dataset, and (2) visual world weights that are common to all datasets, which are learned by undoing the associated bias from each dataset. The visual world weights are expected to be our best possible approximation to the object model trained on an unbiased dataset, and thus tend to have good generalization ability. We demonstrate the effectiveness of our model by applying the learned weights to a novel, unseen dataset, and report superior results for both classification and detection tasks compared to a classical SVM that does not account for the presence of bias. Overall, we find that it is beneficial to explicitly account for bias when combining multiple datasets.

[S1-P13A]

Dog Breed Classification Using Part Localization

Jiongxin Liu, Angjoo Kanazawa, David Jacobs, and Peter Belhumeur

We propose a novel approach to fine-grained image classification in which instances from different classes share common parts but have wide variation in shape and appearance. We use dog breed identification as a test case to show that extracting corresponding parts improves classification performance. This domain is especially challenging since the appearance of corresponding parts can vary dramatically, e.g., the faces of bulldogs and beagles are very different. To find accurate correspondences, we build exemplar-based geometric and appearance models of dog breeds and their face parts. Part correspondence allows us to extract and compare descriptors in like image locations. Our approach also features a hierarchy of parts (e.g., face and eyes) and breed-specific part localization. We achieve 67% recognition rate on a large real-world dataset including 133 dog breeds and 8,351 images, and experimental results show that accurate part localization significantly increases classification performance compared to state-of-the-art approaches.

[S1-P14A]

A Dictionary Learning Approach for Classification: Separating the Particularity and the Commonality

Shu Kong and Donghui Wang

Empirically, we find that, despite the class-specific features owned by the objects appearing in the images, the objects from different categories usually share some common patterns, which do not contribute to the discrimination of them. Concentrating on this observation and under the general dictionary learning (DL) framework, we propose a novel method to explicitly learn a common pattern pool (the commonality) and class-specific dictionaries (the particularity) for classification. We call our method DL-COPAR, which can learn the most compact and most discriminative class-specific dictionaries used for classification. The proposed DL-COPAR is extensively evaluated both on synthetic data and on benchmark image databases in comparison with existing DL-based classification methods. The experimental results demonstrate that DL-COPAR achieves very promising performances in various applications, such as face recognition, handwritten digit recognition, scene classification and object recognition.

[S1-P15A]

Learning to Efficiently Detect Repeatable Interest Points in Depth Data

Stefan Holzer, Jamie Shotton, and Pushmeet Kohli

Interest point (IP) detection is an important component of many computer vision methods. While there are a number of methods for detecting IPs in RGB images, modalities such as depth images and range scans have seen relatively little work. In this paper, we approach the IP detection problem from a machine learning viewpoint and formulate it as a regression problem. We learn a regression forest (RF) model that, given an image patch, tells us if there is an IP in the center of the patch. Our RF based method for IP detection allows an easy trade-off between speed and repeatability by adapting the depth and number of trees used for approximating the interest point response maps. The data used for training the RF model is obtained by running state-of-the-art IP detection methods on the depth images. We show further how the IP response map used for training the RF can be specifically designed to increase repeatability by employing 3D models of scenes generated by reconstruction systems such as KinectFusion [1]. Our experiments demonstrate that the use of such data leads to considerably improved IP detection.

[S1-P16A]

Effective Use of Frequent Itemset Mining for Image Classification

Basura Fernando, Elisa Fromont, and Tinne Tuytelaars

In this paper we propose a new and effective scheme for applying frequent itemset mining to image classification tasks. We refer to the new set of obtained patterns as Frequent Local Histograms or FLHs. During the construction of the FLHs, we pay special attention to keep all the local histogram information during the mining process and to select the most relevant reduced set of FLH patterns for classification. The careful choice of the visual primitives and some proposed extensions to exploit other visual cues such as colour or global spatial information allow us to build powerful bag-of-FLH-based image representations. We show that these bag-of-FLHs are more discriminative than traditional bag-of-words and yield state-of-the art results on various image classification benchmarks.

[S1-P17A]

Efficient Discriminative Projections for Compact Binary Descriptors

Tomasz Trzcinski and Vincent Lepetit

Binary descriptors of image patches are increasingly popular given that they require less storage and enable faster processing. This, however, comes at a price of lower recognition performances. To boost these performances, we project the image patches to a more discriminative subspace, and threshold their coordinates to build our binary descriptor. However, applying complex projections to the patches is slow, which negates some of the advantages of binary descriptors. Hence, our key idea is to learn the discriminative projections so that they can be decomposed into a small number of simple filters for which the responses can be computed fast. We show that with as few as 32 bits per descriptor we outperform the state-of-the-art binary descriptors in terms of both accuracy and efficiency.

[S1-P18A]

Descriptor Learning Using Convex Optimisation

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman

The objective of this work is to learn descriptors suitable for the sparse feature detectors used in viewpoint invariant matching. We make a number of novel contributions towards this goal: first, it is shown that learning the pooling regions for the descriptor can be formulated as a convex optimisation problem selecting the regions using sparsity; second, it is shown that dimensionality reduction can also be formulated as a convex optimisation problem, using the nuclear norm to reduce dimensionality. Both of these problems use large margin discriminative learning methods. The third contribution is a new method of obtaining the positive and negative training data in a weakly supervised manner. And, finally, we employ a state-of-the-art stochastic optimizer that is efficient and well matched to the non-smooth cost functions proposed here. It is demonstrated that the new learning methods improve over the state of the art in descriptor learning for large scale matching, Brown et al. [2], and large scale object retrieval, Philbin et al. [10].

[S1-P19A]

Bottom-Up Perceptual Organization of Images into Object Part Hypotheses

Maruthi Narayanan and Benjamin Kimia

The demise of ''segmentation-then-recognition'' strategy led to a paradigm shift toward feature-based discriminative recognition with significant success. However, increased complexity in multi-class datasets reveals that local low-level features may not be sufficiently discriminative, requiring the construction and use of more complex structural features which are necessarily category independent. The paper proposes a bottom-up procedure for generating fragment features which are intended to be object part hypotheses. Suggesting that the demise of segmentation to generate a representation suitable for recognition was due to prematurely committing to a grouping option in the face of ambiguities, the proposed framework considers and tracks multiple alternate grouping options. This approach is made tractable by (i) using a medial fragment representation which allows for the simultaneous use of multiple cues, (ii) a set of transforms to effect grouping operations, (iii) a containment graph representation which avoids duplicate consideration of possibilities, and the estimation of the likelihood of a grouping sequence to retain only plausible groupings. The resulting hypotheses are evaluated intrinsically by measuring their ability to represent objects with a few fragments. They are also evaluated by comparison to algorithms which aim to generate full object segments, with results that match or exceed the state of art, thus demonstrating the suitability of the proposed mid-level representation.

[S1-P20A]

Match Graph Construction for Large Image Databases

Kwang In Kim, James Tompkin, Martin Theobald, Jan Kautz, and Christian Theobalt

How best to efficiently establish correspondence among a large set of images or video frames is an interesting unanswered question. For large databases, the high computational cost of performing pair-wise image matching is a major problem. However, for many applications, images are inherently sparsely connected, and so current techniques try to correctly estimate small potentially matching subsets of databases upon which to perform expensive pair-wise matching. Our contribution is to pose the identification of potential matches as a link prediction problem in an image correspondence graph, and to propose an effective algorithm to solve this problem. Our algorithm facilitates incremental image matching: initially, the match graph is very sparse, but it becomes dense as we alternate between link prediction and verification. We demonstrate the effectiveness of our algorithm by comparing it with several existing alternatives on large-scale databases. Our resulting match graph is useful for many different applications. As an example, we show the benefits of our graph construction method to a label propagation application which propagates user-provided sparse object labels to other instances of that object in large image collections.

[S1-P21A]

Modeling Complex Temporal Composition of Actionlets for Activity Prediction

Kang Li, Jie Hu, and Yun Fu

Early prediction of ongoing activity has been more and more valuable in a large variety of time-critical applications. To build an effective representation for prediction, human activities can be characterized by a complex temporal composition of constituent simple actions. Different from early recognition on short-duration simple activities, we propose a novel framework for long-duration complex activity prediction by discovering the causal relationships between constituent actions and the predictable characteristics of activities. The major contributions of our work include: (1) we propose a novel activity decomposition method by monitoring motion velocity which encodes a temporal decomposition of long activities into a sequence of meaningful action units; (2) Probabilistic Suffix Tree (PST) is introduced to represent both large and small order Markov dependencies between action units; (3) we present a Predictive Accumulative Function (PAF) to depict the predictability of each kind of activity. The effectiveness of the proposed method is evaluated on two experimental scenarios: activities with middle-level complexity and activities with high-level complexity. Our method achieves promising results and can predict global activity classes and local action units.

[S1-P22A]

Learning Human Interaction by Interactive Phrases

Yu Kong, Yunde Jia, and Yun Fu

In this paper, we present a novel approach for human interaction recognition from videos. We introduce high-level descriptions called interactive phrases to express binary semantic motion relationships between interacting people. Interactive phrases naturally exploit human knowledge to describe interactions and allow us to construct a more descriptive model for recognizing human interactions. We propose a novel hierarchical model to encode interactive phrases based on the latent SVM framework where interactive phrases are treated as latent variables. The interdependencies between interactive phrases are explicitly captured in the model to deal with motion ambiguity and partial occlusion in interactions. We evaluate our method on a newly collected BIT-Interaction dataset and UT-Interaction dataset. Promising results demonstrate the effectiveness of the proposed method.

[S1-P1B]

Learning to Recognize Daily Actions Using Gaze

Alireza Fathi, Yin Li, and James M. Rehg

We present a probabilistic generative model for simultaneously recognizing daily actions and predicting gaze locations in videos recorded from an egocentric camera. We focus on activities requiring eye-hand coordination and model the spatio-temporal relationship between the gaze point, the scene objects, and the action label. Our model captures the fact that the distribution of both visual features and object occurrences in the vicinity of the gaze point is correlated with the verb-object pair describing the action. It explicitly incorporates known properties of gaze behavior from the psychology literature, such as the temporal delay between fixation and manipulation events. We present an inference method that can predict the best sequence of gaze locations and the associated action label from an input sequence of images. We demonstrate improvements in action recognition rates and gaze prediction accuracy relative to state-of-the-art methods, on two new datasets that contain egocentric videos of daily activities and gaze.

[S1-P2B]

Gait Recognition by Ranking

Raúl Martín-Félez and Tao Xiang

The advantage of gait over other biometrics such as face or fingerprint is that it can operate from a distance and without subject cooperation. However, this also makes gait subject to changes in various covariate conditions including carrying, clothing, surface and view angle. Existing approaches attempt to address these condition changes by feature selection, feature transformation or discriminant subspace learning. However, they suffer from lack of training samples from each subject, can only cope with changes in a subset of conditions with limited success, and are based on the invalid assumption that the covariate conditions are known a priori. They are thus unable to perform gait recognition under a genuine uncooperative setting. We propose a novel approach which casts gait recognition as a bipartite ranking problem and leverages training samples from different classes/people and even from different datasets. This makes our approach suitable for recognition under a genuine uncooperative setting and robust against any covariate types, as demonstrated by our extensive experiments.

[S1-P3B]

Semi-intrinsic Mean Shift on Riemannian Manifolds

Rui Caseiro, João F. Henriques, Pedro Martins, and Jorge Batista

The original mean shift algorithm [1] on Euclidean spaces (MS) was extended in [2] to operate on general Riemannian manifolds. This extension is extrinsic (Ext-MS) since the mode seeking is performed on the tangent spaces [3], where the underlying curvature is not fully considered (tangent spaces are only valid in a small neighborhood). In [3] was proposed an intrinsic mean shift designed to operate on two particular Riemannian manifolds (IntGS-MS), i.e. Grassmann and Stiefel manifolds (using manifold-dedicated density kernels). It is then natural to ask whether mean shift could be intrinsically extended to work on a large class of manifolds. We propose a novel paradigm to intrinsically reformulate the mean shift on general Riemannian manifolds. This is accomplished by embedding the Riemannian manifold into a Reproducing Kernel Hilbert Space (RKHS) by using a general and mathematically well-founded Riemannian kernel function, i.e. heat kernel [5]. The key issue is that when the data is implicitly mapped to the Hilbert space, the curvature of the manifold is taken into account (i.e. exploits the underlying information of the data). The inherent optimization is then performed on the embedded space. Theoretic analysis and experimental results demonstrate the promise and effectiveness of this novel paradigm.

[S1-P4B]

Efficient Nonlocal Regularization for Optical Flow

Philipp Krähenbühl and Vladlen Koltun

Dense optical flow estimation in images is a challenging problem because the algorithm must coordinate the estimated motion across large regions in the image, while avoiding inappropriate smoothing over motion boundaries. Recent works have advocated for the use of nonlocal regularization to model long-range correlations in the flow. However, incorporating nonlocal regularization into an energy optimization framework is challenging due to the large number of pairwise penalty terms. Existing techniques either substitute intermediate filtering of the flow field for direct optimization of the nonlocal objective, or suffer substantial performance penalties when the range of the regularizer increases. In this paper, we describe an optimization algorithm that efficiently handles a general type of nonlocal regularization objectives for optical flow estimation. The computational complexity of the algorithm is independent of the range of the regularizer. We show that nonlocal regularization improves estimation accuracy at longer ranges than previously reported, and is complementary to intermediate filtering of the flow field. Our algorithm is simple and is compatible with many optical flow models.

[S1-P5B]

Fast Fusion Moves for Multi-model Estimation

Andrew Delong, Olga Veksler, and Yuri Boykov

We develop a fast, effective algorithm for minimizing a well-known objective function for robust multi-model estimation. Our work introduces a combinatorial step belonging to a family of powerful move-making methods like α-expansion and fusion. We also show that our subproblem can be quickly transformed into a comparatively small instance of minimum-weighted vertex-cover. In practice, these vertex-cover subproblems are almost always bipartite and can be solved exactly by specialized network flow algorithms. Experiments indicate that our approach achieves the robustness of methods like affinity propagation, whilst providing the speed of fast greedy heuristics.

[S1-P6B]

Approximate MRF Inference Using Bounded Treewidth Subgraphs

Alexander Fix, Joyce Chen, Endre Boros, and Ramin Zabih

Graph cut algorithms [9], commonly used in computer vision, solve a first-order MRF over binary variables. The state of the art for this NP-hard problem is QPBO [1,2], which finds the values for a subset of the variables in the global minimum. While QPBO is very effective overall there are still many difficult problems where it can only label a small subset of the variables. We propose a new approach that, instead of optimizing the original graphical model, instead optimizes a tractable sub-model, defined as an energy function that uses a subset of the pairwise interactions of the original, but for which exact inference can be done efficiently. Our Bounded Treewidth Subgraph (k-BTS) algorithm greedily computes a large weight treewidth-k subgraph of the signed graph, then solves the energy minimization problem for this subgraph by dynamic programming. The edges omitted by our greedy method provide a per-instance lower bound. We demonstrate promising experimental results for binary deconvolution, a challenging problem used to benchmark QPBO [2]: our algorithm performs an order of magnitude better than QPBO or its common variants [4], both in terms of energy and accuracy, and the visual quality of our output is strikingly better as well. We also obtain a significant improvement in energy and accuracy on a stereo benchmark with 2nd order priors [5], although the improvement in visual quality is more modest. Our method's running time is comparable to QPBO.

[S1-P7B]

Recursive Bilateral Filtering

Qingxiong Yang

This paper proposes a recursive implementation of the bilateral filter. Unlike previous methods, this implementation yields an bilateral filter whose computational complexity is linear in both input size and dimensionality. The proposed implementation demonstrates that the bilateral filter can be as efficient as the recent edge-preserving filtering methods, especially for high-dimensional images. Let the number of pixels contained in the image be N, and the number of channels be D, the computational complexity of the proposed implementation will be O(ND). It is more efficient than the state-of-the-art bilateral filtering methods that have a computational complexity of $O(ND^2)$ [1] (linear in the image size but polynomial in dimensionality) or O(Nlog(N)D) [2] (linear in the dimensionality thus faster than [1] for high-dimensional filtering). Specifically, the proposed implementation takes about 43 ms to process a one megapixel color image (and about 14 ms to process a 1 megapixel grayscale image) which is about 18× faster than [1] and 86× faster than [2]. The experiments were conducted on a MacBook Air laptop computer with a 1.8 GHz Intel Core i7 CPU and 4 GB memory. The memory complexity of the proposed implementation is also low: as few as the image memory will be required (memory for the images before and after filtering is excluded). This paper also derives a new filter named gradient domain bilateral filter from the proposed recursive implementation. Unlike the bilateral filter, it performs bilateral filtering on the gradient domain. It can be used for edge-preserving filtering but avoids sharp edges that are observed to cause visible artifacts in some computer graphics tasks. The proposed implementations were proved to be effective for a number of computer vision and computer graphics applications, including stylization, tone mapping, detail enhancement and stereo matching.

[S1-P8B]

Accelerated Large Scale Optimization by Concomitant Hashing

Yadong Mu, John Wright, and Shih-Fu Chang

Traditional locality-sensitive hashing (LSH) techniques aim to tackle the curse of explosive data scale by guaranteeing that similar samples are projected onto proximal hash buckets. Despite the success of LSH on numerous vision tasks like image retrieval and object matching, however, its potential in large-scale optimization is only realized recently. In this paper we further advance this nascent area. We first identify two common operations known as the computational bottleneck of numerous optimization algorithms in a large-scale setting, i.e., min/max inner product. We propose a hashing scheme for accelerating min/max inner product, which exploits properties of order statistics of statistically correlated random vectors. Compared with other schemes, our algorithm exhibits improved recall at a lower computational cost. The effectiveness and efficiency of the proposed method are corroborated by theoretic analysis and several important applications. Especially, we use the proposed hashing scheme to perform approximate l_1 regularized least squares with dictionaries with millions of elements, a scale which is beyond the capability of currently known exact solvers. Nonetheless, it is highlighted that the focus of this paper is not on a new hashing scheme for approximate nearest neighbor problem. It exploits a new application for the hashing techniques and proposes a general framework for accelerating a large variety of optimization procedures in computer vision.

[S1-P9B]

Graph Degree Linkage: Agglomerative Clustering on a Directed Graph

Wei Zhang, Xiaogang Wang, Deli Zhao, and Xiaoou Tang

This paper proposes a simple but effective graph-based agglomerative algorithm, for clustering high-dimensional data. We explore the different roles of two fundamental concepts in graph theory, indegree and outdegree, in the context of clustering. The average indegree reflects the density near a sample, and the average outdegree characterizes the local geometry around a sample. Based on such insights, we define the affinity measure of clusters via the product of average indegree and average outdegree. The product-based affinity makes our algorithm robust to noise. The algorithm has three main advantages: good performance, easy implementation, and high computational efficiency. We test the algorithm on two fundamental computer vision problems: image clustering and object matching. Extensive experiments demonstrate that it outperforms the state-of-the-arts in both applications.

[S1-P10B]

Supervised Earth Mover's Distance Learning and Its Computer Vision Applications

Fan Wang and Leonidas J. Guibas

The Earth Mover's Distance (EMD) is an intuitive and natural distance metric for comparing two histograms or probability distributions. It provides a distance value as well as a flow-network indicating how the probability mass is optimally transported between the bins. In traditional EMD, the ground distance between the bins is pre-defined. Instead, we propose to jointly optimize the ground distance matrix and the EMD flow-network based on a partial ordering of histogram distances in an optimization framework. Our method is further extended to accept information from general labeled pairs. The trained ground distance better reflects the cross-bin relationships, hence produces more accurate EMD values and flow-networks. Two computer vision applications are used to demonstrate the effectiveness of the algorithm: first, we apply the optimized EMD value to face verification, and achieve state-of-the-art performance on the PubFig and the LFW data sets; second, the learned EMD flow-network is used to analyze face attribute changes, obtaining consistent paths that demonstrate intuitive transitions on certain facial attributes.

[S1-P11B]

Global Optimization of Object Pose and Motion from a Single Rolling Shutter Image with Automatic 2D-3D Matching

Ludovic Magerand, Adrien Bartoli, Omar Ait-Aider, and Daniel Pizarro

Low cost CMOS cameras can have an acquisition mode called rolling shutter which sequentially exposes the scan-lines. When a single object moves with respect to the camera, this creates image distortions. Assuming 2D-3D correspondences known, previous work showed that the object pose and kinematics can be estimated from a single rolling shutter image. This was achieved using a suboptimal initialization followed by local iterative optimization. We propose a polynomial projection model for rolling shutter cameras and a constrained global optimization of its parameters. This is done by means of a semidefinite programming problem obtained from the generalized problem of moments method. Contrarily to previous work, our optimization does not require an initialization and ensures that the global minimum is achieved. This allows us to build automatically robust 2D-3D correspondences using a template to provide an initial set of correspondences. Experiments show that our method slightly improves previous work on both simulated and real data. This is due to local minima into which previous methods get trapped. We also successfully experimented building 2D-3D correspondences automatically with both simulated and real data.

[S1-P12B]

Online Learning of Linear Predictors for Real-Time Tracking

Stefan Holzer, Marc Pollefeys, Slobodan Ilic, David Joseph Tan, and Nassir Navab

Although fast and reliable, real-time template tracking using linear predictors requires a long training time. The lack of the ability to learn new templates online prevents their use in applications that require fast learning. This especially holds for applications where the scene is not known a priori and multiple templates have to be added online. So far, linear predictors had to be either learned offline [1] or in an iterative manner by starting with a small sized template and growing it over time [2]. In this paper, we propose a fast and simple reformulation of the learning procedure that allows learning new linear predictors online.

[S1-P13B]

Online Learned Discriminative Part-Based Appearance Models for Multi-human Tracking

Bo Yang and Ram Nevatia

We introduce an online learning approach to produce discriminative part-based appearance models (DPAMs) for tracking multiple humans in real scenes by incorporating association based and category free tracking methods. Detection responses are gradually associated into tracklets in multiple levels to produce final tracks. Unlike most previous multi-target tracking approaches which do not explicitly consider occlusions in appearance modeling, we introduce a part based model that explicitly finds unoccluded parts by occlusion reasoning in each frame, so that occluded parts are removed in appearance modeling. Then DPAMs for each tracklet is online learned to distinguish a tracklet with others as well as the background, and is further used in a conservative category free tracking approach to partially overcome the missed detection problem as well as to reduce difficulties in tracklet associations under long gaps. We evaluate our approach on three public data sets, and show significant improvements compared with state-of-art methods.

[S1-P14B]

Exposure Stacks of Live Scenes with Hand-Held Cameras

Jun Hu, Orazio Gallo, and Kari Pulli

Many computational photography applications require the user to take multiple pictures of the same scene with different camera settings. While this allows to capture more information about the scene than what is possible with a single image, the approach is limited by the requirement that the images be perfectly registered. In a typical scenario the camera is hand-held and is therefore prone to moving during the capture of an image burst, while the scene is likely to contain moving objects. Combining such images without careful registration introduces annoying artifacts in the final image. This paper presents a method to register exposure stacks in the presence of both camera motion and scene changes. Our approach warps and modifies the content of the images in the stack to match that of a reference image. Even in the presence of large, highly non-rigid displacements we show that the images are correctly registered to the reference.

[S1-P15B]

Dual-Force Metric Learning for Robust Distracter-Resistant Tracker

Zhibin Hong, Xue Mei, and Dacheng Tao

In this paper, we propose a robust distracter-resistant tracking approach by learning a discriminative metric that adaptively learns the importance of features on-the-fly. The proposed metric is elaborately designed for the tracking problem by forming a margin objective function which systematically includes distance margin maximization and reconstruction error constraint that acts as a force to push distracters away from the positive space and into the negative space. Due to the variety of negative samples in the tracking problem, we specifically introduce the similarity propagation technique that gives distracters a second force from the negative space. Consequently, the discriminative metric obtained helps to preserve the most discriminative information to separate the target from distracters while ensuring the stability of the optimal metric. We seamlessly combine it with the popular L1 minimization tracker. Our tracker is therefore not only resistant to distracters, but also inherits the merit of occlusion robustness from the L1 tracker. Quantitative comparisons with several state-of-the-art algorithms have been conducted in many challenging video sequences. The results show that our method resists distracters excellently and achieves superior performance.

[S1-P16B]

Shape and Reflectance from Natural Illumination

Geoffrey Oxholm and Ko Nishino

We introduce a method to jointly estimate the BRDF and geometry of an object from a single image under known, but uncontrolled, natural illumination. We show that this previously unexplored problem becomes tractable when one exploits the orientation clues embedded in the lighting environment. Intuitively, unique regions in the lighting environment act analogously to the point light sources of traditional photometric stereo; they strongly constrain the orientation of the surface patches that reflect them. The reflectance, which acts as a bandpass filter on the lighting environment, determines the necessary scale of such regions. Accurate reflectance estimation, however, relies on accurate surface orientation information. Thus, these two factors must be estimated jointly. To do so, we derive a probabilistic formulation and introduce priors to address situations where the reflectance and lighting environment do not sufficiently constrain the geometry of the object. Through extensive experimentation we show what this space looks like, and offer insights into what problems become solvable in various categories of real-world natural illumination environments.

[S1-P17B]

Frequency Analysis of Transient Light Transport with Applications in Bare Sensor Imaging

Di Wu, Gordon Wetzstein, Christopher Barsi, Thomas Willwacher, Matthew O'Toole, Nikhil Naik, Qionghai Dai, Kyros Kutulakos, and Ramesh Raskar

Light transport has been analyzed extensively, in both the primal domain and the frequency domain; the latter provides intuition of effects introduced by free space propagation and by optical elements, and allows for optimal designs of computational cameras for tailored, efficient information capture. Here, we relax the common assumption that the speed of light is infinite and analyze free space propagation in the frequency domain considering spatial, temporal, and angular light variation. Using this analysis, we derive analytic expressions for cross-dimensional information transfer and show how this can be exploited for designing a new, time-resolved bare sensor imaging system.

[S1-P18B]

Nonuniform Lattice Regression for Modeling the Camera Imaging Pipeline

Hai Ting Lin, Zheng Lu, Seon Joo Kim, and Michael S. Brown

We describe a method to construct a sparse lookup table (LUT) that is effective in modeling the camera imaging pipeline that maps a RAW camera values to their sRGB output. This work builds on the recent in-camera color processing model proposed by Kim et al. [1] that included a 3D gamut-mapping function. The major drawback in [1] is the high computational cost of the 3D mapping function that uses radial basis functions (RBF) involving several thousand control points. We show how to construct a LUT using a novel nonuniform lattice regression method that adapts the LUT lattice to better fit the 3D gamut-mapping function. Our method offers not only a performance speedup of an order of magnitude faster than RBF, but also a compact mechanism to describe the imaging pipeline.

[S1-P19B]

Context-Based Automatic Local Image Enhancement

Sung Ju Hwang, Ashish Kapoor, and Sing Bing Kang

In this paper, we describe a technique to automatically enhance the perceptual quality of an image. Unlike previous techniques, where global statistics of the image are used to determine enhancement operation, our method is local and relies on local scene descriptors and context in addition to high-level image statistics. We cast the problem of image enhancement as searching for the best transformation for each pixel in the given image and then discovering the enhanced image using a formulation based on Gaussian Random Fields. The search is done in a coarse-to-fine manner, namely by finding the best candidate images, followed by pixels. Our experiments indicate that such context-based local enhancement is better than global enhancement schemes. A user study using Mechanical Turk shows that the subjects prefer contextual and local enhancements over the ones provided by existing schemes.

[S1-P20B]

Segmentation with Non-linear Regional Constraints via Line-Search Cuts

Lena Gorelick, Frank R. Schmidt, Yuri Boykov, Andrew Delong, and Aaron Ward

This paper is concerned with energy-based image segmentation problems. We introduce a general class of regional functionals defined as an arbitrary non-linear combination of regional unary terms. Such (high-order) functionals are very useful in vision and medical applications and some special cases appear in prior art. For example, our general class of functionals includes but is not restricted to soft constraints on segment volume, its appearance histogram, or shape. Our overall segmentation energy combines regional functionals with standard length-based regularizers and/or other submodular terms. In general, regional functionals make the corresponding energy minimization NP-hard. We propose a new greedy algorithm based on iterative line search. A parametric max-flow technique efficiently explores all solutions along the direction (line) of the steepest descent of the energy. We compute the best "step size", i.e. the globally optimal solution along the line. This algorithm can make large moves escaping weak local minima, as demonstrated on many real images.

[S1-P21B]

Hausdorff Distance Constraint for Multi-surface Segmentation

Frank R. Schmidt and Yuri Boykov

It is well known that multi-surface segmentation can be cast as a multi-labeling problem. Different segments may belong to the same semantic object which may impose various inter-segment constraints [1]. In medical applications, there are a lot of scenarios where upper bounds on the Hausdorff distances between subsequent surfaces are known. We show that incorporating these priors into multi-surface segmentation is potentially NP-hard. To cope with this problem we develop a submodular-supermodular procedure that converges to a locally optimal solution well-approximating the problem. While we cannot guarantee global optimality, only feasible solutions are considered during the optimization process. Empirically, we get useful solutions for many challenging medical applications including MRI and ultrasound images.

[S1-P22B]

Background Subtraction Using Low Rank and Group Sparsity Constraints

Xinyi Cui, Junzhou Huang, Shaoting Zhang, and Dimitris N. Metaxas

Background subtraction has been widely investigated in recent years. Most previous work has focused on stationary cameras. Recently, moving cameras have also been studied since videos from mobile devices have increased significantly. In this paper, we propose a unified and robust framework to effectively handle diverse types of videos, e.g., videos from stationary or moving cameras. Our model is inspired by two observations: 1) background motion caused by orthographic cameras lies in a low rank subspace, and 2) pixels belonging to one trajectory tend to group together. Based on these two observations, we introduce a new model using both low rank and group sparsity constraints. It is able to robustly decompose a motion trajectory matrix into foreground and background ones. After obtaining foreground and background trajectories, the information gathered on them is used to build a statistical model to further label frames at the pixel level. Extensive experiments demonstrate very competitive performance on both synthetic data and real videos.

[S1-P23B]

Free Hand-Drawn Sketch Segmentation

Zhenbang Sun, Changhu Wang, Liqing Zhang, and Lei Zhang

In this paper, we study the problem of how to segment a freehand sketch at the object level. By carefully considering the basic principles of human perceptual organization, a real-time solution is presented to automatically segment a user's sketch during his/her drawing. First, a graph-based sketch segmentation algorithm is proposed to segment a cluttered sketch into multiple parts based on the factor of proximity. Then, to improve the ability of detecting semantically meaningful objects, a semantic-based approach is introduced to simulate the past experience in the perceptual system by leveraging a web-scale clipart database. Finally, other important factors learnt from past experience, such as similarity, symmetry, direction, and closure, are also taken into account to make the approach more robust and practical. The proposed sketch segmentation framework has ability to handle complex sketches with overlapped objects. Extensive experimental results show the effectiveness of the proposed framework and algorithms.

[S1-P24B]

Auto-Grouped Sparse Representation for Visual Analysis

Jiashi Feng, Xiaotong Yuan, Zilei Wang, Huan Xu, and Shuicheng Yan

In this work, we investigate how to automatically uncover the underlying group structure of a feature vector such that each group characterizes certain object-specific patterns, e.g., visual pattern or motion trajectories from one object. By mining the group structure, we can effectively alleviate the mutual inference of multiple objects and improve the performance in various visual analysis tasks. To this end, we propose a novel auto-grouped sparse representation (ASR) method. ASR groups semantically correlated feature elements together through optimally fusing their multiple sparse representations. Due to the intractability of primal objective function, we also propose well-behaved convex relaxation and smooth approximation to guarantee obtaining a global optimal solution effectively. Finally, we apply ASR in two important visual analysis tasks: multi-label image classification and motion segmentation. Comprehensive experimental evaluations show that ASR is able to achieve superior performance compared with the state-of-the-arts on these two tasks.

POSTER SESSION 2

Monday, October 8
14:30 - 17:00

[S2-P1A]

Background Inpainting for Videos with Dynamic Objects and a Free-Moving Camera

Miguel Granados, Kwang In Kim, James Tompkin, Jan Kautz, and Christian Theobalt

We propose a method for removing marked dynamic objects from videos captured with a free-moving camera, so long as the objects occlude parts of the scene with a static background. Our approach takes as input a video, a mask marking the object to be removed, and a mask marking the dynamic objects to remain in the scene. To inpaint a frame, we align other candidate frames in which parts of the missing region are visible. Among these candidates, a single source is chosen to fill each pixel so that the final arrangement is color-consistent. Intensity differences between sources are smoothed using gradient domain fusion. Our frame alignment process assumes that the scene can be approximated using piecewise planar geometry: A set of homographies is estimated for each frame pair, and one each is selected for aligning pixels such that the color-discrepancy is minimized and the epipolar constraints are maintained. We provide experimental validation with several real-world video sequences to demonstrate that, unlike in previous work, inpainting videos shot with free-moving cameras does not necessarily require estimation of absolute camera positions and per-frame per-pixel depth maps.

[S2-P2A]

Optimal Templates for Nonrigid Surface Reconstruction

Markus Moll and Luc Van Gool

This paper addresses the problem of reconstructing a deforming surface from point observations in a monocular video sequence. Recent state-of-the-art approaches divide the surface into smaller patches to simplify the problem. Among these, one very promising approach reconstructs the patches individually using a quadratic deformation model. In this paper, we demonstrate limitations that affect its applicability to real-world data and propose an approach that overcomes these problems. In particular, we show how to eliminate the need for manually picking a template that is used to model the deformations. We evaluate our algorithm on both synthetic and real-world data sets and show that it systematically reduces the reconstruction error by a factor of up to ten.

[S2-P3A]

Learning Domain Knowledge for Facade Labelling

Dengxin Dai, Mukta Prasad, Gerhard Schmitt, and Luc Van Gool

This paper presents an approach to address the problem of image façade labelling. In the architectural literature, domain knowledge is usually expressed geometrically in the final design, so façade labelling should on the one hand conform to visual evidence, and on the other hand to the architectural principles – how individual assets (e.g. doors, windows) interact with each other to form a façade as a whole. To this end, we first propose a recursive splitting method to segment façades into a bunch of tiles for semantic recognition. The segmentation improves the processing speed, guides visual recognition on suitable scales and renders the extraction of architectural principles easy. Given a set of segmented training façades with their label maps, we then identify a set of meta-features to capture both the visual evidence and the architectural principles. The features are used to train our façade labelling model. In the test stage, the features are extracted from segmented façades and the inferred label maps. The following three steps are iterated until the optimal labelling is reached: 1) proposing modifications to the current labelling; 2) extracting new features for the proposed labelling; 3) feeding the new features to the labelling model to decide whether to accept the modifications. In experiments, we evaluated our method on the ECP façade dataset and achieved higher precision than the state-of-the-art at both the pixel level and the structural level.

[S2-P4A]

Simultaneous Shape and Pose Adaption of Articulated Models Using Linear Optimization

Matthias Straka, Stefan Hauswiesner, Matthias Rüther, and Horst Bischof

We propose a novel formulation to express the attachment of a polygonal surface to a skeleton using purely linear terms. This enables to simultaneously adapt the pose and shape of an articulated model in an efficient way. Our work is motivated by the difficulty to constrain a mesh when adapting it to multi-view silhouette images. However, such an adaption is essential when capturing the detailed temporal evolution of skin and clothing of a human actor without markers. While related work is only able to ensure surface consistency during mesh adaption, our coupled optimization of the skeleton creates structural stability and minimizes the sensibility to occlusions and outliers in input images. We demonstrate the benefits of our approach in an extensive evaluation. The skeleton attachment considerably reduces implausible deformations, especially when the number of input views is limited.

[S2-P5A]

Robust Fitting for Multiple View Geometry

Olof Enqvist, Erik Ask, Fredrik Kahl, and Kalle Åström

How hard are geometric vision problems with outliers? We show that for most fitting problems, a solution that minimizes the number of outliers can be found with an algorithm that has polynomial time-complexity in the number of points (independent of the rate of outliers). Further, and perhaps more interestingly, other cost functions such as the truncated L2-norm can also be handled within the same framework with the same time complexity. We apply our framework to triangulation, relative pose problems and stitching, and give several other examples that fulfill the required conditions. Based on efficient polynomial equation solvers, it is experimentally demonstrated that these problems can be solved reliably, in particular for low-dimensional models. Comparisons to standard random sampling solvers are also given.

[S2-P6A]

Improving Image-Based Localization by Active Correspondence Search

Torsten Sattler, Bastian Leibe, and Leif Kobbelt

We propose a powerful pipeline for determining the pose of a query image relative to a point cloud reconstruction of a large scene consisting of more than one million 3D points. The key component of our approach is an efficient and effective search method to establish matches between image features and scene points needed for pose estimation. Our main contribution is a framework for actively searching for additional matches, based on both 2D-to-3D and 3D-to-2D search. A unified formulation of search in both directions allows us to exploit the distinct advantages of both strategies, while avoiding their weaknesses. Due to active search, the resulting pipeline is able to close the gap in registration performance observed between efficient search methods and approaches that are allowed to run for multiple seconds, without sacrificing run-time efficiency. Our method achieves the best registration performance published so far on three standard benchmark datasets, with run-times comparable or superior to the fastest state-of-the-art methods.

[S2-P7A]

From Meaningful Contours to Discriminative Object Shape

Pradeep Yarlagadda and Björn Ommer

Shape is a natural, highly prominent characteristic of objects that human vision utilizes everyday. But despite its expressiveness, shape poses significant challenges for category-level object detection in cluttered scenes: Object form is an emergent property that cannot be perceived locally but becomes only available once the whole object has been detected and segregated from the background. Thus we address the detection of objects and the assembling of their shape simultaneously. A dictionary of meaningful contours is obtained by clustering based on contour co-activation in all training images. We seek a joint, consistent placement of all contours in an image, since placing them independently from another is not reliable due to the emergence of shape. Therefore, the characteristic object shape is learned by discovering spatially consistent configurations of all dictionary contours using maximum margin multiple instance learning. During recognition, objects are detected and their shape is explained simultaneously by optimizing a single cost function. We demonstrate the benefit of our approach on standard shape benchmarks.

[S2-P8A]

A Particle Filter Framework for Contour Detection

Nicolas Widynski and Max Mignotte

We investigate the contour detection task in complex natural images. We propose a novel contour detection algorithm which locally tracks small pieces of edges called edgelets. The combination of the Bayesian modeling and the edgelets enables the use of semi-local prior information and image-dependent likelihoods. We use a mixed offline and online learning strategy to detect the most relevant edgelets. The detection problem is then modeled as a sequential Bayesian tracking task, estimated using a particle filtering technique. Experiments on the Berkeley Segmentation Datasets show that the proposed Particle Filter Contour Detector method performs well compared to competing state-of-the-art methods.

[S2-P9A]

TriCoS: A Tri-level Class-Discriminative Co-segmentation Method for Image Classification

Yuning Chai, Esa Rahtu, Victor Lempitsky, Luc Van Gool, and Andrew Zisserman

The aim of this paper is to leverage foreground segmentation to improve classification performance on weakly annotated datasets – those with no additional annotation other than class labels. We introduce TriCoS, a new co-segmentation algorithm that looks at all training images jointly and automatically segments out the most class-discriminative foregrounds for each image. Ultimately, those foreground segmentations are used to train a classification system. TriCoS solves the co-segmentation problem by minimizing losses at three different levels: the category level for foreground/background consistency across images belonging to the same category, the image level for spatial continuity within each image, and the dataset level for discrimination between classes. In an extensive set of experiments, we evaluate the algorithm on three benchmark datasets: the UCSD-Caltech Birds-200-2010, the Stanford Dogs, and the Oxford Flowers 102. With the help of a modern image classifier, we show superior performance compared to previously published classification methods and other co-segmentation methods.

[S2-P10A]

Multi-view Discriminant Analysis

Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen

The same object can be observed at different viewpoints or even by different sensors, thus generating multiple distinct even heterogeneous samples. Nowadays, more and more applications need to recognize object from distinct views. Some seminal works have been proposed for object recognition across two views and applied to multiple views in some inefficient pairwise manner. In this paper, we propose a Multi-view Discriminant Analysis (MvDA) method, which seeks for a discriminant common space by jointly learning multiple view-specific linear transforms for robust object recognition from multiple views, in a non-pairwise manner. Specifically, our MvDA is formulated to jointly solve the multiple linear transforms by optimizing a generalized Rayleigh quotient, i.e., maximizing the between-class variations and minimizing the within-class variations of the low-dimensional embeddings from both intra-view and inter-view in the common space. By reformulating this problem as a ratio trace problem, an analytical solution can be achieved by using the generalized eigenvalue decomposition. The proposed method is applied to three multi-view face recognition problems: face recognition across poses, photo-sketch face recognition, and Visual (VIS) image vs. Near Infrared (NIR) image face recognition. Evaluations are conducted respectively on Multi-PIE, CUFSF and HFB databases. Intensive experiments show that MvDA can achieve a more discriminant common space, with up to 13% improvement compared with the best known results.

[S2-P11A]

Multi-scale Patch Based Collaborative Representation for Face Recognition with Margin Distribution Optimization

Pengfei Zhu, Lei Zhang, Qinghua Hu, and Simon C.K. Shiu

Small sample size is one of the most challenging problems in face recognition due to the difficulty of sample collection in many real-world applications. By representing the query sample as a linear combination of training samples from all classes, the so-called collaborative representation based classification (CRC) shows very effective face recognition performance with low computational cost. However, the recognition rate of CRC will drop dramatically when the available training samples per subject are very limited. One intuitive solution to this problem is operating CRC on patches and combining the recognition outputs of all patches. Nonetheless, the setting of patch size is a non-trivial task. Considering the fact that patches on different scales can have complementary information for classification, we propose a multi-scale patch based CRC method, while the ensemble of multi-scale outputs is achieved by regularized margin distribution optimization. Our extensive experiments validated that the proposed method outperforms many state-of-the-art patch based face recognition algorithms.

[S2-P12A]

Object Detection Using Strongly-Supervised Deformable Part Models

Hossein Azizpour and Ivan Laptev

Deformable part-based models [1, 2] achieve state-of-the-art performance for object detection, but rely on heuristic initialization during training due to the optimization of non-convex cost function. This paper investigates limitations of such an initialization and extends earlier methods using additional supervision. We explore strong supervision in terms of annotated object parts and use it to (i) improve model initialization, (ii) optimize model structure, and (iii) handle partial occlusions. Our method is able to deal with sub-optimal and incomplete annotations of object parts and is shown to benefit from semi-supervised learning setups where part-level annotation is provided for a fraction of positive examples only. Experimental results are reported for the detection of six animal classes in PASCAL VOC 2007 and 2010 datasets. We demonstrate significant improvements in detection performance compared to the LSVM [1] and the Poselet [3] object detectors.

[S2-P13A]

Efficient Misalignment-Robust Representation for Real-Time Face Recognition

Meng Yang, Lei Zhang, and David Zhang

Sparse representation techniques for robust face recognition have been widely studied in the past several years. Recently face recognition with simultaneous misalignment, occlusion and other variations has achieved interesting results via robust alignment by sparse representation (RASR). In RASR, the best alignment of a testing sample is sought subject by subject in the database. However, such an exhaustive search strategy can make the time complexity of RASR prohibitive in large-scale face databases. In this paper, we propose a novel scheme, namely misalignment robust representation (MRR), by representing the misaligned testing sample in the transformed face space spanned by all subjects. The MRR seeks the best alignment via a two-step optimization with a coarse-to-fine search strategy, which needs only two deformation-recovery operations. Extensive experiments on representative face databases show that MRR has almost the same accuracy as RASR in various face recognition and verification tasks but it runs tens to hundreds of times faster than RASR. The running time of MRR is less than 1 second in the large-scale Multi-PIE face database, demonstrating its great potential for real-time face recognition.

[S2-P14A]

Monocular Object Detection Using 3D Geometric Primitives

Peter Carr, Yaser Sheikh, and Iain Matthews

Multiview object detection methods achieve robustness in adverse imaging conditions by exploiting projective consistency across views. In this paper, we present an algorithm that achieves performance comparable to multiview methods from a single camera by employing geometric primitives as proxies for the true 3D shape of objects, such as pedestrians or vehicles. Our key insight is that for a calibrated camera, geometric primitives produce predetermined location-specific patterns in occupancy maps. We use these to define spatially-varying kernel functions of projected shape. This leads to an analytical formation model of occupancy maps as the convolution of locations and projected shape kernels. We estimate object locations by deconvolving the occupancy map using an efficient template similarity scheme. The number of objects and their positions are determined using the mean shift algorithm. The approach is highly parallel because the occupancy probability of a particular geometric primitive at each ground location is an independent computation. The algorithm extends to multiple cameras without requiring significant bandwidth. We demonstrate comparable performance to multiview methods and show robust, realtime object detection on full resolution HD video in a variety of challenging imaging conditions.

[S2-P15A]

Object-Centric Spatial Pooling for Image Classification

Olga Russakovsky, Yuanqing Lin, Kai Yu, and Li Fei-Fei

Spatial pyramid matching (SPM) based pooling has been the dominant choice for state-of-art image classification systems. In contrast, we propose a novel object-centric spatial pooling (OCP) approach, following the intuition that knowing the location of the object of interest can be useful for image classification. OCP consists of two steps: (1) inferring the location of the objects, and (2) using the location information to pool foreground and background features separately to form the image-level representation. Step (1) is particularly challenging in a typical classification setting where precise object location annotations are not available during training. To address this challenge, we propose a framework that learns object detectors using only image-level class labels, or so-called weak labels. We validate our approach on the challenging PASCAL07 dataset. Our learned detectors are comparable in accuracy with state-of-the-art weakly supervised detection methods. More importantly, the resulting OCP approach significantly outperforms SPM-based pooling in image classification.

[S2-P16A]

Statistics of Patch Offsets for Image Completion

Kaiming He and Jian Sun

Image completion involves filling missing parts in images. In this paper we address this problem through the statistics of patch offsets. We observe that if we match similar patches in the image and obtain their offsets (relative positions), the statistics of these offsets are sparsely distributed. We further observe that a few dominant offsets provide reliable information for completing the image. With these offsets we fill the missing region by combining a stack of shifted images via optimization. A variety of experiments show that our method yields generally better results and is faster than existing state-of-the-art methods.

[S2-P17A]

Spectral Demons – Image Registration via Global Spectral Correspondence

Herve Lombaert, Leo Grady, Xavier Pennec, Nicholas Ayache, and Farida Cheriet

Image registration is a building block for many applications in computer vision and medical imaging. However the current methods are limited when large and highly non-local deformations are present. In this paper, we introduce a new direct feature matching technique for non-parametric image registration where efficient nearest-neighbor searches find global correspondences between intensity, spatial and geometric information. We exploit graph spectral representations that are invariant to isometry under complex deformations. Our direct feature matching technique is used within the established Demons framework for diffeomorphic image registration. Our method, called Spectral Demons, can capture very large, complex and highly non-local deformations between images. We evaluate the improvements of our method on 2D and 3D images and demonstrate substantial improvement over the conventional Demons algorithm for large deformations.

[S2-P18A]

MatchMiner: Efficient Spanning Structure Mining in Large Image Collections

Yin Lou, Noah Snavely, and Johannes Gehrke

Many new computer vision applications are utilizing large-scale data-sets of places derived from the many billions of photos on the Web. Such applications often require knowledge of the visual connectivity structure of these image collections—describing which images overlap or are otherwise related—and an important step in understanding this structure is to identify connected components of this underlying image graph. As the structure of this graph is often initially unknown, this problem can be posed as one of exploring the connectivity between images as quickly as possible, by intelligently selecting a subset of image pairs for feature matching and geometric verification, without having to test all $O(n^2)$ possible pairs. We propose a novel, scalable algorithm called MatchMiner that efficiently explores visual relations between images, incorporating ideas from relevance feedback to improve decision making over time, as well as a simple yet effective rank distance measure for detecting outlier images. Using these ideas, our algorithm automatically prioritizes image pairs that can potentially connect or contribute to large connected components, using an information-theoretic algorithm to decide which image pairs to test next. Our experimental results show that MatchMiner can efficiently find connected components in large image collections, significantly outperforming state-of-the-art image matching methods.

[S2-P19A]

V1-Inspired Features Induce a Weighted Margin in SVMs

Hilton Bristow and Simon Lucey

Image representations derived from simplified models of the primary visual cortex (V1), such as HOG and SIFT, elicit good performance in a myriad of visual classification tasks including object recognition/detection, pedestrian detection and facial expression classification. A central question in the vision, learning and neuroscience communities regards why these architectures perform so well. In this paper, we offer a unique perspective to this question by subsuming the role of V1-inspired features directly within a linear support vector machine (SVM). We demonstrate that a specific class of such features in conjunction with a linear SVM can be reinterpreted as inducing a weighted margin on the Kronecker basis expansion of an image. This new viewpoint on the role of V1-inspired features allows us to answer fundamental questions on the uniqueness and redundancies of these features, and offer substantial improvements in terms of computational and storage efficiency.

[S2-P20A]

Unsupervised Discovery of Mid-Level Discriminative Patches

Saurabh Singh, Abhinav Gupta, and Alexei A. Efros

The goal of this paper is to discover a set of discriminative patches which can serve as a fully unsupervised mid-level visual representation. The desired patches need to satisfy two requirements: 1) to be representative, they need to occur frequently enough in the visual world; 2) to be discriminative, they need to be different enough from the rest of the visual world. The patches could correspond to parts, objects, "visual phrases", etc. but are not restricted to be any one of them. We pose this as an unsupervised discriminative clustering problem on a huge dataset of image patches. We use an iterative procedure which alternates between clustering and training discriminative classifiers, while applying careful cross-validation at each step to prevent overfitting. The paper experimentally demonstrates the effectiveness of discriminative patches as an unsupervised mid-level visual representation, suggesting that it could be used in place of visual words for many tasks. Furthermore, discriminative patches can also be used in a supervised regime, such as scene classification, where they demonstrate state-of-the-art performance on the MIT Indoor-67 dataset.

[S2-P21A]

Self-similar Sketch

Andrea Vedaldi and Andrew Zisserman

We introduce the self-similar sketch, a new method for the extraction of intermediate image features that combines three principles: detection of self-similarity structures, nonaccidental alignment, and instance-specific modelling. The method searches for self-similar image structures that form nonaccidental patterns, for example collinear arrangements. We demonstrate a simple implementation of this idea where self-similar structures are found by looking for SIFT descriptors that map to the same visual words in image-specific vocabularies. This results in a visual word map which is searched for elongated connected components. Finally, segments are fitted to these connected components, extracting linear image structures beyond the ones that can be captured by conventional edge detectors, as the latter implicitly assume a specific appearance for the edges (steps). The resulting collection of segments constitutes a "sketch" of the image. This is applied to the task of estimating vanishing points, horizon, and zenith in standard benchmark data, obtaining state-of-the-art results. We also propose a new vanishing point estimation algorithm based on recently introduced techniques for the continuous-discrete optimisation of energies arising from model selection priors.

[S2-P22A]

Depth Matters: Influence of Depth Cues on Visual Saliency

Congyan Lang, Tam V. Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan

Most previous studies on visual saliency have only focused on static or dynamic 2D scenes. Since the human visual system has evolved predominantly in natural three dimensional environments, it is important to study whether and how depth information influences visual saliency. In this work, we first collect a large human eye fixation database compiled from a pool of 600 2D-vs-3D image pairs viewed by 80 subjects, where the depth information is directly provided by the Kinect camera and the eye tracking data are captured in both 2D and 3D free-viewing experiments. We then analyze the major discrepancies between 2D and 3D human fixation data of the same scenes, which are further abstracted and modeled as novel depth priors. Finally, we evaluate the performances of state-of-the-art saliency detection models over 3D images, and propose solutions to enhance their performances by integrating the depth priors.

[S2-P1B]

Quaternion-Based Spectral Saliency Detection for Eye Fixation Prediction

Boris Schauerte and Rainer Stiefelhagen

In recent years, several authors have reported that spectral saliency detection methods provide state-of-the-art performance in predicting human gaze in images (see, e.g., [1–3]). We systematically integrate and evaluate quaternion DCT- and FFT-based spectral saliency detection [3,4], weighted quaternion color space components [5], and the use of multiple resolutions [1]. Furthermore, we propose the use of the eigenaxes and eigenangles for spectral saliency models that are based on the quaternion Fourier transform. We demonstrate the outstanding performance on the Bruce-Tsotsos (Toronto), Judd (MIT), and Kootstra- Schomacker eye-tracking data sets.

[S2-P2B]

Human Activities as Stochastic Kronecker Graphs

Sinisa Todorovic

A human activity can be viewed as a space-time repetition of activity primitives. Both instances of the primitives, and their repetition are stochastic. They can be modeled by a generative model-graph, where nodes correspond to the primitives, and the graph's adjacency matrix encodes their affinities for probabilistic grouping into observable video features. When a video of the activity is represented by a graph capturing the space-time layout of video features, such a video graph can be viewed as probabilistically sampled from the activity's model-graph. This sampling is formulated as a successive Kronecker multiplication of the model's affinity matrix. The resulting Kronecker-power matrix is taken as a noisy permutation of the adjacency matrix of the video graph. The paper presents our: 1) model-graph; 2) memory- and time-efficient, weakly supervised learning of activity primitives and their affinities; and 3) inference aimed at finding the best expected correspondences between the primitives and observed video features. Our results demonstrate good scalability on UCF50, and superior performance to that of the state of the art on individual, structured, and collective activities of UCF YouTube, Olympic, and Collective datasets.

[S2-P3B]

Facial Action Transfer with Personalized Bilinear Regression

Dong Huang and Fernando De La Torre

Facial Action Transfer (FAT) has recently attracted much attention in computer vision due to its diverse applications in the movie industry, computer games, and privacy protection. The goal of FAT is to "clone" the facial actions from the videos of one person (source) to another person (target). In this paper, we will assume that we have a video of the source person but only one frontal image of the target person. Most successful methods for FAT require a training set with annotated correspondence between expressions of different subjects, sometimes including many images of the target subject. However, labeling expressions is time consuming and error prone (i.e., it is difficult to capture the same intensity of the expression across people). Moreover, in many applications it is not realistic to have many labeled images of the target. This paper proposes a method to learn a personalized facial model, that can produce photo-realistic person-specific facial actions (e.g., synthesize wrinkles for smiling), from only a neutral image of the target person. More importantly, our learning method does not need an explicit correspondence of expressions across subjects. Experiments on the Cohn-Kanade and the RU-FACS databases show the effectiveness of our approach to generate video-realistic images of the target person driven by spontaneous facial actions of the source. Moreover, we illustrate applications of FAT to face de-identification.

[S2-P4B]

Point of Gaze Estimation through Corneal Surface Reflection in an Active Illumination Environment

Atsushi Nakazawa and Christian Nitschke

Eye gaze tracking (EGT) is a common problem with many applications in various fields. While recent methods have achieved improvements in accuracy and usability, current techniques still share several limitations. A major issue is the need for external calibration between the gaze camera system and the scene, which commonly restricts to static planar surfaces and leads to parallax errors. To overcome these issues, the paper proposes a novel scheme that uses the corneal imaging technique to directly analyze reflections from a scene illuminated with structured light. This comprises two major contributions: First, an analytic solution is developed for the forward projection problem to obtain the gaze reflection point (GRP), where light from the point of gaze (PoG) in the scene reflects at the corneal surface into an eye image. We also develop a method to compensate for the individual offset between the optical axis and true visual axis. Second, introducing active coded illumination enables robust and accurate matching at the GRP to obtain the PoG in a scene image, which is the first use of this technique in EGT and corneal reflection analysis. For this purpose, we designed a special high-power IR LED-array projector. Experimental evaluation with a prototype system shows that the proposed scheme achieves considerable accuracy and successfully supports depth-varying environments.

[S2-P5B]

Order-Preserving Sparse Coding for Sequence Classification

Bingbing Ni, Pierre Moulin, and Shuicheng Yan

In this paper, we investigate order-preserving sparse coding for classifying multi-dimensional sequence data. Such a problem is often tackled by first decomposing the input sequence into individual frames and extracting features, then performing sparse coding or other processing for each frame based feature vector independently, and finally aggregating individual responses to classify the input sequence. However, this heuristic approach ignores the underlying temporal order of the input sequence frames, which in turn results in suboptimal discriminative capability. In this work, we introduce a temporal-order-preserving regularizer which aims to preserve the temporal order of the reconstruction coefficients. An efficient Nesterov-type smooth approximation method is developed for optimization of the new regularization criterion, with guaranteed error bounds. Extensive experiments for time series classification on a synthetic dataset, several machine learning benchmarks, and a challenging real-world RGB-D human activity dataset, show that the proposed coding scheme is discriminative and robust, and it outperforms previous art for sequence classification.

[S2-P6B]

Min-Space Integral Histogram

Séverine Dubuisson and Christophe Gonzales

In this paper, we present a new approach for quickly computing the histograms of a set of unrotating rectangular regions. Although it is related to the well-known Integral Histogram (IH), our approach significantly outperforms it, both in terms of memory requirements and of response times. By preprocessing the region of interest (ROI) computing and storing a temporary histogram for each of its pixels, IH is effective only when a large amount of histograms located in a small ROI need be computed by the user. Unlike IH, our approach, called Min-Space Integral Histogram, only computes and stores those temporary histograms that are strictly necessary (less than 4 times the number of regions). Comparative tests highlight its efficiency, which can be up to 75 times faster than IH. In particular, we show that our approach is much less sensitive than IH to histogram quantization and to the size of the ROI.

[S2-P7B]

On Learning Higher-Order Consistency Potentials for Multi-class Pixel Labeling

Kyoungup Park and Stephen Gould

Pairwise Markov random fields are an effective framework for solving many pixel labeling problems in computer vision. However, their performance is limited by their inability to capture higher-order correlations. Recently proposed higher-order models are showing superior performance to their pairwise counterparts. In this paper, we derive two variants of the higher-order lower linear envelop model and show how to perform tractable move-making inference in these models. We propose a novel use of this model for encoding consistency constraints over large sets of pixels. Importantly these pixel sets do not need to be contiguous. However, the consistency model has a large number of parameters to be tuned for good performance. We exploit the structured SVM paradigm to learn optimal parameters and show some practical techniques to overcome huge computation requirements. We evaluate our model on the problems of image denoising and semantic segmentation.

[S2-P8B]

Sparse Coding and Dictionary Learning for Symmetric Positive Definite Matrices: A Kernel Approach

Mehrtash T. Harandi, Conrad Sanderson, Richard Hartley, and Brian C. Lovell

Recent advances suggest that a wide range of computer vision problems can be addressed more appropriately by considering non-Euclidean geometry. This paper tackles the problem of sparse coding and dictionary learning in the space of symmetric positive definite matrices, which form a Riemannian manifold. With the aid of the recently introduced Stein kernel (related to a symmetric version of Bregman matrix divergence), we propose to perform sparse coding by embedding Riemannian manifolds into reproducing kernel Hilbert spaces. This leads to a convex and kernel version of the Lasso problem, which can be solved efficiently. We furthermore propose an algorithm for learning a Riemannian dictionary (used for sparse coding), closely tied to the Stein kernel. Experiments on several classification tasks (face recognition, texture classification, person re-identification) show that the proposed sparse coding approach achieves notable improvements in discrimination accuracy, in comparison to state-of-the-art methods such as tensor sparse coding, Riemannian locality preserving projection, and symmetry-driven accumulation of local features.

[S2-P9B]

Learning Class-to-Image Distance via Large Margin and L1-Norm Regularization

Zhengxiang Wang, Shenghua Gao, and Liang-Tien Chia

Image-to-Class (I2C) distance has demonstrated its effectiveness for object recognition in several single-label datasets. However, for the multi-label problem, where an image may contain several regions belonging to different classes, this distance may not work well since it cannot discriminate local features from different regions in the test image and all local features have to be counted in the I2C distance calculation. In this paper, we propose to use Class-to-Image (C2I) distance and show that this distance performs better than I2C distance for multi-label image classification. However, since the number of local features in a class is huge compared to that in an image, the calculation of C2I distance is much more expensive than I2C distance. Moreover, the label information of training images can be used to help select relevant local features for each class and further improve the recognition performance. Therefore, to make C2I distance faster and perform better, we propose an optimization algorithm using L1-norm regularization and large margin constraint to learn the C2I distance, which will not only reduce the number of local features in the class feature set, but also improve the performance of C2I distance due to the use of label information. Experiments on MSRC, Pascal VOC and MirFlickr datasets show that our method can significantly speed up the C2I distance calculation, while achieves better recognition performance than the original C2I distance and other related methods for multi-labeled datasets.

[S2-P10B]

Taxonomic Multi-class Prediction and Person Layout Using Efficient Structured Ranking

Arpit Mittal, Matthew B. Blaschko, Andrew Zisserman, and Philip H.S. Torr

In computer vision efficient multi-class classification is becoming a key problem as the field develops and the number of object classes to be identified increases. Often objects might have some sort of structure such as a taxonomy in which the mis-classification score for object classes close by, using tree distance within the taxonomy, should be less than for those far apart. This is an example of multi-class classification in which the loss function has a special structure. Another example in vision is for the ubiquitous pictorial structure or parts based model. In this case we would like the mis-classification score to be proportional to the number of parts misclassified. It transpires both of these are examples of structured output ranking problems. However, so far no efficient large scale algorithm for this problem has been demonstrated. In this work we propose an algorithm for structured output ranking that can be trained in a time linear in the number of samples under a mild assumption common to many computer vision problems: that the loss function can be discretized into a small number of values. We show the feasibility of structured ranking on these two core computer vision problems and demonstrate a consistent and substantial improvement over competing techniques. Aside from this, we also achieve state-of-the art results for the PASCAL VOC human layout problem.

[S2-P11B]

Robust Point Matching Revisited: A Concave Optimization Approach

Wei Lian and Lei Zhang

The well-known robust point matching (RPM) method uses deterministic annealing for optimization, and it has two problems. First, it cannot guarantee the global optimality of the solution and tends to align the centers of two point sets. Second, deformation needs to be regularized to avoid the generation of undesirable results. To address these problems, in this paper we first show that the energy function of RPM can be reduced to a concave function with very few non-rigid terms after eliminating the transformation variables and applying linear transformation; we then propose to use concave optimization technique to minimize the resulting energy function. The proposed method scales well with problem size, achieves the globally optimal solution, and does not need regularization for simple transformations such as similarity transform. Experiments on synthetic and real data validate the advantages of our method in comparison with state-of-the-art methods.

[S2-P12B]

Learning Discriminative Spatial Relations for Detector Dictionaries: An Application to Pedestrian Detection

Enver Sangineto, Marco Cristani, Alessio Del Bue, and Vittorio Murino

The recent availability of large scale training sets in conjunction with accurate classifiers (e.g., SVMs) makes it possible to build large sets of ''simple'' object detectors and to develop new classification approaches in which dictionaries of visual features are substituted by dictionaries of object detectors. The responses of this collection of detectors can then be used as a high-level image representation. In this work, we propose to go a step further in this direction by modeling spatial relations among different detector responses. We use Random Forests in order to discriminatively select spatial relations which represent frequent co-occurrences of detector responses. We demonstrate our idea in the specific people detection framework, which is a challenging classification task due to the variability of the human body articulations and appearance, and we use the recently proposed poselets as our basic object dictionary. The use of poselets is not the only possible, actually the proposed method can be applied more in general since few assumptions are made on the basic object detector. The results obtained show sharp improvements with respect to both the original poselet-based people detection method and to other state-of-the-art approaches on two difficult benchmark datasets.

[S2-P13B]

Learning Deformations with Parallel Transport

Donglai Wei, Dahua Lin, and John Fisher III

Many vision problems, such as object recognition and image synthesis, are greatly impacted by deformation of objects. In this paper, we develop a deformation model based on Lie algebraic analysis. This work aims to provide a generative model that explicitly decouples deformation from appearance, which is fundamentally different from the prior work that focuses on deformation-resilient features or metrics. Specifically, the deformation group for each object can be characterized by a set of Lie algebraic basis. Such basis for different objects are related via parallel transport. Exploiting the parallel transport relations, we formulate an optimization problem, and derive an algorithm that jointly estimates the deformation basis for a class of objects, given a set of images resulted from the action of the deformations. We test the proposed model empirically on both character recognition and face synthesis.

[S2-P14B]

Multi-channel Shape-Flow Kernel Descriptors for Robust Video Event Detection and Retrieval

Pradeep Natarajan, Shuang Wu, Shiv Vitaladevuni, Xiaodan Zhuang, Unsang Park, Rohit Prasad, and Premkumar Natarajan

Despite the success of spatio-temporal visual features, they are hand-designed and aggregate image or flow gradients using a pre-specified, uniform set of orientation bins. Kernel descriptors [1] generalize such orientation histograms by defining match kernels over image patches, and have shown superior performance for visual object and scene recognition. In our work, we make two contributions: first, we extend kernel descriptors to the spatio-temporal domain to model salient flow, gradient and texture patterns in video. Further, we apply our kernel descriptors to extract features from different color channels. Second, we present a fast algorithm for kernel descriptor computation of O(1) complexity for each pixel in each video patch, producing two orders of magnitude speedup over conventional kernel descriptors and other popular motion features. Our evaluation results on TRECVID MED 2011 dataset indicate that the proposed multi-channel shape-flow kernel descriptors outperform several other features including SIFT, SURF, STIP and Color SIFT.

[S2-P15B]

Tracking Using Motion Patterns for Very Crowded Scenes

Xuemei Zhao, Dian Gong, and Gérard Medioni

This paper proposes Motion Structure Tracker (MST) to solve the problem of tracking in very crowded structured scenes. It combines visual tracking, motion pattern learning and multi-target tracking. Tracking in crowded scenes is very challenging due to hundreds of similar objects, cluttered background, small object size, and occlusions. However, structured crowded scenes exhibit clear motion pattern(s), which provides rich prior information. In MST, tracking and detection are performed jointly, and motion pattern information is integrated in both steps to enforce scene structure constraint. MST is initially used to track a single target, and further extended to solve a simplified version of the multi-target tracking problem. Experiments are performed on real-world challenging sequences, and MST gives promising results. Our method significantly outperforms several state-of-the-art methods both in terms of track ratio and accuracy.

[S2-P16B]

Long-Range Spatio-Temporal Modeling of Video with Application to Fire Detection

Avinash Ravichandran and Stefano Soatto

We describe a methodology for modeling backgrounds subject to significant variability over time-scales ranging from days to years, where the events of interest exhibit subtle variability relative to the normal mode. The motivating application is fire monitoring from remote stations, where illumination changes spanning the day and the season, meteorological phenomena resembling smoke, and the absence of sufficient training data for the two classes make out-of-the-box classification algorithms ineffective. We exploit low-level descriptors, incorporate explicit modeling of nuisance variability, and learn the residual normal-model variability. Our algorithm achieves state-of-the-art performance not only compared to other anomaly detection schemes, but also compared to human performance, both for untrained and trained operators.

[S2-P17B]

GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs

Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah

Data association is an essential component of any human tracking system. The majority of current methods, such as bipartite matching, incorporate a limited-temporal-locality of the sequence into the data association problem, which makes them inherently prone to IDswitches and difficulties caused by long-term occlusion, cluttered background, and crowded scenes.We propose an approach to data association which incorporates both motion and appearance in a global manner. Unlike limited-temporal-locality methods which incorporate a few frames into the data association problem, we incorporate the whole temporal span and solve the data association problem for one object at a time, while implicitly incorporating the rest of the objects. In order to achieve this, we utilize Generalized Minimum Clique Graphs to solve the optimization problem of our data association method. Our proposed method yields a better formulated approach to data association which is supported by our superior results. Experiments show the proposed method makes significant improvements in tracking in the diverse sequences of Town Center [1], TUD-crossing [2], TUD-Stadtmitte [2], PETS2009 [3], and a new sequence called Parking Lot compared to the state of the art methods.

[S2-P18B]

Heliometric Stereo: Shape from Sun Position

Austin Abrams, Christopher Hawley, and Robert Pless

In this work, we present a method to uncover shape from webcams ''in the wild.'' We present a variant of photometric stereo which uses the sun as a distant light source, so that lighting direction can be computed from known GPS and timestamps. We propose an iterative, non-linear optimization process that optimizes the error in reproducing all images from an extended time-lapse with an image formation model that accounts for ambient lighting, shadows, changing light color, dense surface normal maps, radiometric calibration, and exposure. Unlike many approaches to uncalibrated outdoor image analysis, this procedure is automatic, and we report quantitative results by comparing extracted surface normals to Google Earth 3D models. We evaluate this procedure on data from a varied set of scenes and emphasize the advantages of including imagery from many months.

[S2-P19B]

Shape from Single Scattering for Translucent Objects

Chika Inoshita, Yasuhiro Mukaigawa, Yasuyuki Matsushita, and Yasushi Yagi

Translucent objects strongly scatter incident light. Scattering makes the problem of estimating shape of translucent objects difficult, because reflective or transmitted light cannot be reliably extracted from the scattering. In this paper, we propose a new shape estimation method by directly utilizing scattering measurements. Although volumetric scattering is a complex phenomenon, single scattering can be relatively easily modeled because it is a simple one-bounce collision of light to a particle in a medium. Based on this observation, our method determines the shape of objects from the observed intensities of the single scattering and its attenuation. We develop a solution method that simultaneously determines scattering parameters and the shape based on energy minimization. We demonstrate the effectiveness of the proposed approach by extensive experiments using synthetic and real data.

[S2-P20B]

Scale Invariant Optical Flow

Li Xu, Zhenlong Dai, and Jiaya Jia

Scale variation commonly arises in images/videos, which cannot be naturally dealt with by optical flow. Invariant feature matching, on the contrary, provides sparse matching and could fail for regions without conspicuous structures. We aim to establish dense correspondence between frames containing objects in different scales and contribute a new framework taking pixel-wise scales into consideration in optical flow estimation. We propose an effective numerical scheme, which iteratively optimizes discrete scale variables and continuous flow ones. This scheme notably expands the practicality of optical flow in natural scenes containing various types of object motion.

[S2-P21B]

Structured Image Segmentation Using Kernelized Features

Aurélien Lucchi, Yunpeng Li, Kevin Smith, and Pascal Fua

Most state-of-the-art approaches to image segmentation formulate the problem using Conditional Random Fields. These models typically include a unary term and a pairwise term, whose parameters must be carefully chosen for optimal performance. Recently, structured learning approaches such as Structured SVMs (SSVM) have made it possible to jointly learn these model parameters. However, they have been limited to linear kernels, since more powerful non-linear kernels cause the learning to become prohibitively expensive. In this paper, we introduce an approach to "kernelize" the features so that a linear SSVM framework can leverage the power of non-linear kernels without incurring the high computational cost. We demonstrate the advantages of this approach in a series of image segmentation experiments on the MSRC data set as well as 2D and 3D datasets containing imagery of neural tissue acquired with electron microscopes.

[S2-P22B]

Salient Object Detection: A Benchmark

Ali Borji, Dicky N. Sihite, and Laurent Itti

Several salient object detection approaches have been published which have been assessed using different evaluation scores and datasets resulting in discrepancy in model comparison. This calls for a methodological framework to compare existing models and evaluate their pros and cons. We analyze benchmark datasets and scoring techniques and, for the first time, provide a quantitative comparison of 35 state-of-the-art saliency detection models. We find that some models perform consistently better than the others. Saliency models that intend to predict eye fixations perform lower on segmentation datasets compared to salient object detection algorithms. Further, we propose combined models which show that integration of the few best models outperforms all models over other datasets. By analyzing the consistency among the best models and among humans for each scene, we identify the scenes where models or humans fail to detect the most salient object. We highlight the current issues and propose future research directions.

[S2-P23B]

Automatic Segmentation of Unknown Objects, with Application to Baggage Security

Leo Grady, Vivek Singh, Timo Kohlberger, Christopher Alvino, and Claus Bahlmann

Computed tomography (CT) is used widely to image patients for medical diagnosis and to scan baggage for threatening materials. Automated reading of these images can be used to reduce the costs of a human operator, extract quantitative information from the images or support the judgements of a human operator. Object quantification requires an image segmentation to make measurements about object size, material composition and morphology. Medical applications mostly require the segmentation of prespecified objects, such as specific organs or lesions, which allows the use of customized algorithms that take advantage of training data to provide orientation and anatomical context of the segmentation targets. In contrast, baggage screening requires the segmentation algorithm to provide segmentation of an unspecified number of objects with enormous variability in size, shape, appearance and spatial context. Furthermore, security systems demand 3D segmentation algorithms that can quickly and reliably detect threats. To address this problem, we present a segmentation algorithm for 3D CT images that makes no assumptions on the number of objects in the image or on the composition of these objects. The algorithm features a new Automatic QUality Measure (AQUA) model that measures the segmentation confidence for any single object (from any segmentation method) and uses this confidence measure to both control splitting and to optimize the segmentation parameters at runtime for each dataset. The algorithm is tested on 27 bags that were packed with a large variety of different objects.

[S2-P24B]

Multi-scale Clustering of Frame-to-Frame Correspondences for Motion Segmentation

Ralf Dragon, Bodo Rosenhahn, and Jörn Ostermann

We present an approach for motion segmentation using independently detected keypoints instead of commonly used tracklets or trajectories. This allows us to establish correspondences over non-consecutive frames, thus we are able to handle multiple object occlusions consistently. On a frame-to-frame level, we extend the classical split-and-merge algorithm for fast and precise motion segmentation. Globally, we cluster multiple of these segmentations of different time scales with an accurate estimation of the number of motions. On the standard benchmarks, our approach performs best in comparison to all algorithms which are able to handle unconstrained missing data. We further show that it works on benchmark data with more than 98% of the input data missing. Finally, the performance is evaluated on a mobile-phone-recorded sequence with multiple objects occluded at the same time.

ORAL SESSION 2 LEARNING AND LARGE-SCALE RECOGNITION

Monday, October 8
17:05 - 18:30

[S2-O1]

Fourier Kernel Learning

Eduard Gabriel Bazavan, Fuxin Li, and Cristian Sminchisescu

Approximations based on random Fourier embeddings have recently emerged as an efficient and formally consistent methodology to design large-scale kernel machines [23]. By expressing the kernel as a Fourier expansion, features are generated based on a finite set of random basis projections, sampled from the Fourier transform of the kernel, with inner products that are Monte Carlo approximations of the original non-linear model. Based on the observation that different kernel-induced Fourier sampling distributions correspond to different kernel parameters, we show that a scalable optimization process in the Fourier domain can be used to identify the different frequency bands that are useful for prediction on training data. This approach allows us to design a family of linear prediction models where we can learn the hyper-parameters of the kernel together with the weights of the feature vectors jointly. Under this methodology, we recover efficient and scalable linear reformulations for both single and multiple kernel learning. Experiments show that our linear models produce fast and accurate predictors for complex datasets such as the Visual Object Challenge 2011 and ImageNet ILSVRC 2011.

[S2-O2]

Efficient Optimization for Low-Rank Integrated Bilinear Classifiers

Takumi Kobayashi and Nobuyuki Otsu

In pattern classification, it is needed to efficiently treat two-way data (feature matrices) while preserving the two-way structure such as spatio-temporal relationships, etc. The classifier for the feature matrix is generally formulated by multiple bilinear forms which result in a matrix. The rank of the matrix, i.e., the number of bilinear forms, should be low from the viewpoint of generalization performance and computational cost. For that purpose, we propose a low-rank bilinear classifier based on the efficient optimization. In the proposed method, the classifier is optimized by minimizing the trace norm of the classifier (matrix), which contributes to the rank reduction for an efficient classifier without any hard constraint on the rank. We formulate the optimization problem in a tractable convex form and propose the procedure to solve it efficiently with the global optimum. In addition, by considering a kernel-based extension of the bilinear method, we induce a novel multiple kernel learning (MKL), called heterogeneous MKL. The method combines both inter kernels between heterogeneous types of features and the ordinary kernels within homogeneous features into a new discriminative kernel in a unified manner using the bilinear model. In the experiments on various classification problems using feature arrays, co-occurrence feature matrices, and multiple kernels, the proposed method exhibits favorable performances compared to the other methods.

[S2-O3]

Metric Learning for Large Scale Image Classification: Generalizing to New Classes at Near-Zero Cost

Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka

We are interested in large-scale image classification and especially in the setting where images corresponding to new or existing classes are continuously added to the training set. Our goal is to devise classifiers which can incorporate such images and classes on-the-fly at (near) zero cost. We cast this problem into one of learning a metric which is shared across all classes and explore k-nearest neighbor (k-NN) and nearest class mean (NCM) classifiers. We learn metrics on the ImageNet 2010 challenge data set, which contains more than 1.2M training images of 1K classes. Surprisingly, the NCM classifier compares favorably to the more flexible k-NN classifier, and has comparable performance to linear SVMs. We also study the generalization performance, among others by using the learned metric on the ImageNet-10K dataset, and we obtain competitive performance. Finally, we explore zero-shot classification, and show how the zero-shot model can be combined very effectively with small training datasets.

[S2-O4]

Leafsnap: A Computer Vision System for Automatic Plant Species Identification

Neeraj Kumar, Peter N. Belhumeur, Arijit Biswas, David W. Jacobs, W. John Kress, Ida C. Lopez, and João V.B. Soares

We describe the first mobile app for identifying plant species using automatic visual recognition. The system – called Leafsnap – identifies tree species from photographs of their leaves. Key to this system are computer vision components for discarding non-leaf images, segmenting the leaf from an untextured background, extracting features representing the curvature of the leaf's contour over multiple scales, and identifying the species from a dataset of the 184 trees in the Northeastern United States. Our system obtains state-of-the-art performance on the real-world images from the new Leafsnap Dataset – the largest of its kind. Throughout the paper, we document many of the practical steps needed to produce a computer vision system such as ours, which currently has nearly a million users.

[S2-O5]

Large Scale Visual Geo-Localization of Images in Mountainous Terrain

Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys

Given a picture taken somewhere in the world, automatic geo-localization of that image is a task that would be extremely useful e.g. for historical and forensic sciences, documentation purposes, organization of the world's photo material and also intelligence applications. While tremendous progress has been made over the last years in visual location recognition within a single city, localization in natural environments is much more difficult, since vegetation, illumination, seasonal changes make appearance-only approaches impractical. In this work, we target mountainous terrain and use digital elevation models to extract representations for fast visual database lookup. We propose an automated approach for very large scale visual localization that can efficiently exploit visual information (contours) and geometric constraints (consistent orientation) at the same time. We validate the system on the scale of a whole country (Switzerland, 40000km^2) using a new dataset of more than 200 landscape query pictures with ground truth.

POSTER SESSION 3

Tuesday, October 9
08:45 - 11:15

[S3-P1A]

Covariance Propagation and Next Best View Planning for 3D Reconstruction

Sebastian Haner and Anders Heyden

This paper examines the potential benefits of applying next best view planning to sequential 3D reconstruction from unordered image sequences. A standard sequential structure-and-motion pipeline is extended with active selection of the order in which cameras are resectioned. To this end, approximate covariance propagation is implemented throughout the system, providing running estimates of the uncertainties of the reconstruction, while also enhancing robustness and accuracy. Experiments show that the use of expensive global bundle adjustment can be reduced throughout the process, while the additional cost of propagation is essentially linear in the problem size.

[S3-P2A]

Dilated Divergence Based Scale-Space Representation for Curve Analysis

Max W.K. Law, KengYeow Tay, Andrew Leung, Gregory J. Garvin, and Shuo Li

This study proposes the novel dilated divergence scale-space representation for multidimensional curve-like image structure analysis. In the proposed framework, image structures are modeled as curves with arbitrary thickness. The dilated divergence analyzes the structure boundaries along the curve normal space in a multi-scale fashion. The dilated divergence based detection is formulated so as to 1) sustain the disturbance introduced by neighboring objects, 2) recognize the curve normal and tangent spaces. The latter enables the innovative formulation of structure eccentricity analysis and curve tangent space-based structure motion analysis, which have been scarcely investigated in literature. The proposed method is validated using 2D, 3D and 4D images. The structure principal direction estimation accuracies, structure scale detection accuracies and detection stabilities are quantified and compared against two scale-space approaches, showing a competitive performance of the proposed approach, under the disturbance introduced by image noise and neighboring objects. Moreover, as an application example employing the dilated divergence detection responses, an automated approach is tailored for spinal cord centerline extraction. The proposed method is shown to be versatile to well suit a wide range of applications.

[S3-P3A]

A Parameterless Line Segment and Elliptical Arc Detector with Enhanced Ellipse Fitting

Viorica Patraucean, Pierre Gurdjos, and Rafael Grompone von Gioi

We propose a combined line segment and elliptical arc detector, which formally guarantees the control of the number of false positives and requires no parameter tuning. The accuracy of the detected elliptical features is improved by using a novel non-iterative ellipse fitting technique, which merges the algebraic distance with the gradient orientation. The performance of the detector is evaluated on computer-generated images and on natural images.

[S3-P4A]

Detecting and Reconstructing 3D Mirror Symmetric Objects

Sudipta N. Sinha, Krishnan Ramnath, and Richard Szeliski

We present a system that detects 3D mirror-symmetric objects in images and then reconstructs their visible symmetric parts. Our detection stage is based on matching mirror symmetric feature points and descriptors and then estimating the symmetry direction using RANSAC. We enhance this step by augmenting feature descriptors with their affine-deformed versions and matching these extended sets of descriptors. The reconstruction stage uses a novel edge matching algorithm that matches symmetric pairs of curves that are likely to be counterparts. This allows the algorithm to reconstruct lightly textured objects, which are problematic for traditional feature-based and intensity-based stereo matchers.

[S3-P5A]

3D Reconstruction of Dynamic Scenes with Multiple Handheld Cameras

Hanqing Jiang, Haomin Liu, Ping Tan, Guofeng Zhang, and Hujun Bao

Accurate dense 3D reconstruction of dynamic scenes from natural images is still very challenging. Most previous methods rely on a large number of fixed cameras to obtain good results. Some of these methods further require separation of static and dynamic points, which are usually restricted to scenes with known background. We propose a novel dense depth estimation method which can automatically recover accurate and consistent depth maps from the synchronized video sequences taken by a few handheld cameras. Unlike fixed camera arrays, our data capturing setup is much more flexible and easier to use. Our algorithm simultaneously solves bilayer segmentation and depth estimation in a unified energy minimization framework, which combines different spatio-temporal constraints for effective depth optimization and segmentation of static and dynamic points. A variety of examples demonstrate the effectiveness of the proposed framework.

[S3-P6A]

Joint Face Alignment: Rescue Bad Alignments with Good Ones by Regularized Re-fitting

Xiaowei Zhao, Xiujuan Chai, and Shiguang Shan

Nowadays, more and more applications need to jointly align a set of facial images from one specific person, which forms the so-called joint face alignment problem. To address this problem, in this paper, starting from an initial face alignment results, we propose to enhance the alignments by a fundamentally novel idea: rescuing the bad alignments with their well-aligned neighbors. In our method, a discriminative alignment evaluator is well designed to assess the initial face alignments and separate the well-aligned images from the badly-aligned ones. To correct the bad ones, a robust regularized re-fitting algorithm is proposed by exploiting the appearance consistency between the badly-aligned image and its k well-aligned nearest neighbors. Experiments conducted on faces in the wild demonstrate that our method greatly improves the initial face alignment results of an off-the-shelf facial landmark locator. In addition, the effectiveness of our method is validated through comparing with other state-of-the-art methods in joint face alignment under complex conditions.

[S3-P7A]

Dynamic Facial Expression Recognition Using Longitudinal Facial Expression Atlases

Yimo Guo, Guoying Zhao, and Matti Pietikäinen

In this paper, we propose a new scheme to formulate the dynamic facial expression recognition problem as a longitudinal atlases construction and deformable groupwise image registration problem. The main contributions of this method include: 1) We model human facial feature changes during the facial expression process by a diffeomorphic image registration framework; 2) The subject-specific longitudinal change information of each facial expression is captured by building an expression growth model; 3) Longitudinal atlases of each facial expression are constructed by performing groupwise registration among all the corresponding expression image sequences of different subjects. The constructed atlases can reflect overall facial feature changes of each expression among the population, and can suppress the bias due to inter-personal variations. The proposed method was extensively evaluated on the Cohn-Kanade, MMI, and Oulu-CASIA VIS dynamic facial expression databases and was compared with several state-of-the-art facial expression recognition approaches. Experimental results demonstrate that our method consistently achieves the highest recognition accuracies among other methods under comparison on all the databases.

[S3-P8A]

Crosstalk Cascades for Frame-Rate Pedestrian Detection

Piotr Dollár, Ron Appel, and Wolf Kienzle

Cascades help make sliding window object detection fast, nevertheless, computational demands remain prohibitive for numerous applications. Currently, evaluation of adjacent windows proceeds independently; this is suboptimal as detector responses at nearby locations and scales are correlated. We propose to exploit these correlations by tightly coupling detector evaluation of nearby windows. We introduce two opposing mechanisms: detector excitation of promising neighbors and inhibition of inferior neighbors. By enabling neighboring detectors to communicate, crosstalk cascades achieve major gains (4-30× speedup) over cascades evaluated independently at each image location. Combined with recent advances in fast multi-scale feature computation, for which we provide an optimized implementation, our approach runs at 35-65 fps on 640×480 images while attaining state-of-the-art accuracy.

[S3-P9A]

Query Specific Fusion for Image Retrieval

Shaoting Zhang, Ming Yang, Timothee Cour, Kai Yu, and Dimitris N. Metaxas

Recent image retrieval algorithms based on local features indexed by a vocabulary tree and holistic features indexed by compact hashing codes both demonstrate excellent scalability. However, their retrieval precision may vary dramatically among queries. This motivates us to investigate how to fuse the ordered retrieval sets given by multiple retrieval methods, to further enhance the retrieval precision. Thus, we propose a graph-based query specific fusion approach where multiple retrieval sets are merged and reranked by conducting a link analysis on a fused graph. The retrieval quality of an individual method is measured by the consistency of the top candidates' nearest neighborhoods. Hence, the proposed method is capable of adaptively integrating the strengths of the retrieval methods using local or holistic features for different queries without any supervision. Extensive experiments demonstrate competitive performance on 4 public datasets, i.e., the UKbench, Corel-5K, Holidays and San Francisco Landmarks datasets.

[S3-P10A]

Size Matters: Exhaustive Geometric Verification for Image Retrieval

Henrik Stewénius, Steinar H. Gunderson, and Julien Pilet

The overreaching goals in large-scale image retrieval are bigger, better and cheaper. For systems based on local features we show how to get both efficient geometric verification of every match and unprecedented speed for the low sparsity situation. Large-scale systems based on quantized local features usually process the index one term at a time, forcing two separate scoring steps: First, a scoring step to find candidates with enough matches, and then a geometric verification step where a subset of the candidates are checked. Our method searches through the index a document at a time, verifying the geometry of every candidate in a single pass. We study the behavior of several algorithms with respect to index density -- a key element for large-scale databases. In order to further improve the efficiency we also introduce a new new data structure, called the counting min-tree, which outperforms other approaches when working with low database density, a necessary condition for very large-scale systems. We demonstrate the effectiveness of our approach with a proof of concept system that can match an image against a database of more than 90 billion images in just a few seconds.

[S3-P11A]

Scene Aligned Pooling for Complex Video Recognition

Liangliang Cao, Yadong Mu, Apostol Natsev, Shih-Fu Chang, Gang Hua, and John R. Smith

Real-world videos often contain dynamic backgrounds and evolving people activities, especially for those web videos generated by users in unconstrained scenarios. This paper proposes a new visual representation, namely scene aligned pooling, for the task of event recognition in complex videos. Based on the observation that a video clip is often composed with shots of different scenes, the key idea of scene aligned pooling is to decompose any video features into concurrent scene components, and to construct classification models adaptive to different scenes. The experiments on two large scale real-world datasets including the TRECVID Multimedia Event Detection 2011 and the Human Motion Recognition Databases (HMDB) show that our new visual representation can consistently improve various kinds of visual features such as different low-level color and texture features, or middle-level histogram of local descriptors such as SIFT, or space-time interest points, and high level semantic model features, by a significant margin. For example, we improve the-state-of-the-art accuracy on HMDB dataset by 20% in terms of accuracy.

[S3-P12A]

Discovering Latent Domains for Multisource Domain Adaptation

Judy Hoffman, Brian Kulis, Trevor Darrell, and Kate Saenko

Recent domain adaptation methods successfully learn cross-domain transforms to map points between source and target domains. Yet, these methods are either restricted to a single training domain, or assume that the separation into source domains is known a priori. However, most available training data contains multiple unknown domains. In this paper, we present both a novel domain transform mixture model which outperforms a single transform model when multiple domains are present, and a novel constrained clustering method that successfully discovers latent domains. Our discovery method is based on a novel hierarchical clustering technique that uses available object category information to constrain the set of feasible domain separations. To illustrate the effectiveness of our approach we present experiments on two commonly available image datasets with and without known domain labels: in both cases our method outperforms baseline techniques which use no domain adaptation or domain adaptation methods that presume a single underlying domain shift.

[S3-P13A]

Visual Recognition Using Local Quantized Patterns

Sibt ul Hussain and Bill Triggs

Features such as Local Binary Patterns (LBP) and Local Ternary Patterns (LTP) have been very successful in a number of areas including texture analysis, face recognition and object detection. They are based on the idea that small patterns of qualitative local gray-level differences contain a great deal of information about higher-level image content. Current local pattern features use hand-specified codings that are limited to small spatial supports and coarse graylevel comparisons. We introduce Local Quantized Patterns (LQP), a generalization that uses lookup-table-based vector quantization to code larger or deeper patterns. LQP inherits some of the flexibility and power of visual word representations without sacrificing the run-time speed and simplicity of local pattern ones. We show that it outperforms well-established features including HOG, LBP and LTP and their combinations on a range of challenging object detection and texture classification problems.

[S3-P14A]

Randomized Spatial Partition for Scene Recognition

Yuning Jiang, Junsong Yuan, and Gang Yu

The spatial layout of images plays a critical role in natural scene analysis. Despite previous work, e.g., spatial pyramid matching, how to design optimal spatial layout for scene classification remains an open problem due to the large variations of scene categories. This paper presents a novel image representation method, with the objective to characterize the image layout by various patterns, in the form of randomized spatial partition (RSP). The RSP-based image representation makes it possible to mine the most descriptive image layout pattern for each category of scenes, and then combine them by training a discriminative classifier, i.e., the proposed ORSP classifier. Besides RSP image representation, another powerful classifier, called the BRSP classifier, is also proposed. By weighting and boosting a sequence of various partition patterns, the BRSP classifier is more robust to the intra-class variations hence leads to a more accurate classification. Both RSP-based classifiers are tested on three publicly available scene datasets. The experimental results highlight the effectiveness of the proposed methods.

[S3-P15A]

Nested Sparse Quantization for Efficient Feature Coding

Xavier Boix, Gemma Roig, Christian Leistner, and Luc Van Gool

Many state-of-the-art methods in object recognition extract features from an image and encode them, followed by a pooling step and classification. Within this processing pipeline, often the encoding step is the bottleneck, for both computational efficiency and performance. We present a novel assignment-based encoding formulation. It allows for the fusion of assignment-based encoding and sparse coding into one formulation. We also use this to design a new, very efficient, encoding. At the heart of our formulation lies a quantization into a set of k-sparse vectors, which we denote as sparse quantization. We design the new encoding as two nested, sparse quantizations. Its efficiency stems from leveraging bit-wise representations. In a series of experiments on standard recognition benchmarks, namely Caltech 101, PASCAL VOC 07 and ImageNet, we demonstrate that our method achieves results that are competitive with the state-of-the-art, and requires orders of magnitude less time and memory. Our method is able to encode one million images using 4 CPUs in a single day, while maintaining a good performance.

[S3-P16A]

Comparative Evaluation of Binary Features

Jared Heinly, Enrique Dunn, and Jan-Michael Frahm

Performance evaluation of salient features has a long-standing tradition in computer vision. In this paper, we fill the gap of evaluation for the recent wave of binary feature descriptors, which aim to provide robustness while achieving high computational efficiency. We use established metrics to embed our assessment into the body of existing evaluations, allowing us to provide a novel taxonomy unifying both traditional and novel binary features. Moreover, we analyze the performance of different detector and descriptor pairings, which are often used in practice but have been infrequently analyzed. Additionally, we complement existing datasets with novel data testing for illumination change, pure camera rotation, pure scale change, and the variety present in photo-collections. Our performance analysis clearly demonstrates the power of the new class of features. To benefit the community, we also provide a website for the automatic testing of new description methods using our provided metrics and datasets (www.cs.unc.edu/feature-evaluation).

[S3-P17A]

Negative Evidences and Co-occurences in Image Retrieval: The Benefit of PCA and Whitening

Hervé Jégou and Ondrej Chum

The paper addresses large scale image retrieval with short vector representations. We study dimensionality reduction by Principal Component Analysis (PCA) and propose improvements to its different phases. We show and explicitly exploit relations between i) mean subtraction and the negative evidence, i.e., a visual word that is mutually missing in two descriptions being compared, and ii) the axis de-correlation and the co-occurrences phenomenon. Finally, we propose an effective way to alleviate the quantization artifacts through a joint dimensionality reduction of multiple vocabularies. The proposed techniques are simple, yet significantly and consistently improve over the state of the art on compact image representations. Complementary experiments in image classification show that the methods are generally applicable.

[S3-P18A]

W$\boldsymbol{\alpha}$SH: Weighted $\boldsymbol{\alpha}$-Shapes for Local Feature Detection

Christos Varytimidis, Konstantinos Rapantzikos, and Yannis Avrithis

Depending on the application, local feature detectors should comply with properties that are often contradictory, e.g. distinctiveness vs. robustness. Providing a good balance is a standing problem in the field. In this direction, we propose a novel approach for local feature detection starting from sampled edges. The detector is based on shape stability measures across the weighted α-filtration, a computational geometry construction that captures the shape of a non-uniform set of points. The extracted features are blob-like and include non-extremal regions as well as regions determined by cavities of boundary shape. Overall, the approach provides distinctive regions, while achieving high robustness in terms of repeatability and matching score, as well as competitive performance in a large scale image retrieval application.

[S3-P19A]

Sparselet Models for Efficient Multiclass Object Detection

Hyun Oh Song, Stefan Zickler, Tim Althoff, Ross Girshick, Mario Fritz, Christopher Geyer, Pedro Felzenszwalb, and Trevor Darrell

We develop an intermediate representation for deformable part models and show that this representation has favorable performance characteristics for multi-class problems when the number of classes is high. Our model uses sparse coding of part filters to represent each filter as a sparse linear combination of shared dictionary elements. This leads to a universal set of parts that are shared among all object classes. Reconstruction of the original part filter responses via sparse matrix-vector product reduces computation relative to conventional part filter convolutions. Our model is well suited to a parallel implementation, and we report a new GPU DPM implementation that takes advantage of sparse coding of part filters. The speed-up offered by our intermediate representation and parallel computation enable real-time DPM detection of 20 different object classes on a laptop computer.

[S3-P20A]

Nested Pictorial Structures

Steve Gu, Ying Zheng, and Carlo Tomasi

We propose a theoretical construct coined nested pictorial structure to represent an object by parts that are recursively nested. Three innovative ideas are proposed: First, the nested pictorial structure finds a part configuration that is allowed to be deformed in geometric arrangement, while being confined to be topologically nested. Second, we define nested features which lend themselves to better, more detailed accounting of pixel data cost and describe occlusion in a principled way. Third, we develop the concept of constrained distance transform, a variation of the generalized distance transform, to guarantee the topological nesting relations and to further enforce that parts have no overlap with each other. We show that matching an optimal nested pictorial structure of K parts on an image of N pixels takes O(NK) time using dynamic programming and constrained distance transform. In our MATLAB/C++ implementation, it takes less than 0.1 seconds to do the global optimal matching when $K = 10$ and $N=400\times400$. We demonstrate the usefulness of nested pictorial structures in the matching of objects of nested patterns, objects in occlusion, and objects that live in a context.

[S3-P21A]

Performance Capture of Interacting Characters with Handheld Kinects

Genzhi Ye, Yebin Liu, Nils Hasler, Xiangyang Ji, Qionghai Dai, and Christian Theobalt

We present an algorithm for marker-less performance capture of interacting humans using only three hand-held Kinect cameras. Our method reconstructs human skeletal poses, deforming surface geometry and camera poses for every time step of the depth video. Skeletal configurations and camera poses are found by solving a joint energy minimization problem which optimizes the alignment of RGBZ data from all cameras, as well as the alignment of human shape templates to the Kinect data. The energy function is based on a combination of geometric correspondence finding, implicit scene segmentation, and correspondence finding using image features. Only the combination of geometric and photometric correspondences and the integration of human pose and camera pose estimation enables reliable performance capture with only three sensors. As opposed to previous performance capture methods, our algorithm succeeds on general uncontrolled indoor scenes with potentially dynamic background, and it succeeds even if the cameras are moving.

[S3-P22A]

Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition

Stefan Mathe and Cristian Sminchisescu

Systems based on bag-of-words models operating on image features collected at maxima of sparse interest point operators have been extremely successful for both computer-based visual object and action recognition tasks. While the sparse, interest-point based approach to recognition is not inconsistent with visual processing in biological systems that operate in ″saccade and fixate″ regimes, the knowledge, methodology, and emphasis in the human and the computer vision communities remains sharply distinct. Here, we make three contributions aiming to bridge this gap. First, we complement existing state-of-the art large-scale dynamic computer vision datasets like Hollywood-2[1] and UCF Sports[2] with human eye movements collected under the ecological constraints of the visual action recognition task. To our knowledge these are the first massive human eye tracking datasets of significant size to be collected for video (497,107 frames, each viewed by 16 subjects), unique in terms of their (a) large scale and computer vision relevance, (b) dynamic, video stimuli, (c) task control, as opposed to free-viewing. Second, we introduce novel dynamic consistency and alignment models, which underline the remarkable stability of patterns of visual search among subjects. Third, we leverage the massive amounts of collected data in order to pursue studies and build automatic, end-to-end trainable computer vision systems based on human eye movements. Our studies not only shed light on the differences between computer vision spatio-temporal interest point image sampling strategies and human fixations, as well as their impact for visual recognition performance, but also demonstrate that human fixations can be accurately predicted, and when used in an end-to-end automatic system, leveraging some of the most advanced computer vision practice, can lead to state of the art results.

[S3-P1B]

Coherent Filtering: Detecting Coherent Motions from Crowd Clutters

Bolei Zhou, Xiaoou Tang, and Xiaogang Wang

Coherent motions, which describe the collective movements of individuals in crowd, widely exist in physical and biological systems. Understanding their underlying priors and detecting various coherent motion patterns from background clutters have both scientific values and a wide range of practical applications, especially for crowd motion analysis. In this paper, we propose and study a prior of coherent motion called Coherent Neighbor Invariance, which characterizes the local spatiotemporal relationships of individuals in coherent motion. Based on the coherent neighbor invariance, a general technique of detecting coherent motion patterns from noisy time-series data called Coherent Filtering is proposed. It can be effectively applied to data with different distributions at different scales in various real-world problems, where the environments could be sparse or extremely crowded with heavy noise. Experimental evaluation and comparison on synthetic and real data show the existence of Coherence Neighbor Invariance and the effectiveness of our Coherent Filtering.

[S3-P2B]

Robust 3D Action Recognition with Random Occupancy Patterns

Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu

We study the problem of action recognition from depth sequences captured by depth cameras, where noise and occlusion are common problems because they are captured with a single commodity camera. In order to deal with these issues, we extract semi-local features called random occupancy pattern (ROP) features, which employ a novel sampling scheme that effectively explores an extremely large sampling space. We also utilize a sparse coding approach to robustly encode these features. The proposed approach does not require careful parameter tuning. Its training is very fast due to the use of the high-dimensional integral image, and it is robust to the occlusions. Our technique is evaluated on two datasets captured by commodity depth cameras: an action dataset and a hand gesture dataset. Our classification results are superior to those obtained by the state of the art approaches on both datasets.

[S3-P3B]

Directional Space-Time Oriented Gradients for 3D Visual Pattern Analysis

Ehsan Norouznezhad, Mehrtash T. Harandi, Abbas Bigdeli, Mahsa Baktash, Adam Postula, and Brian C. Lovell

Various visual tasks such as the recognition of human actions, gestures, facial expressions, and classification of dynamic textures require modeling and the representation of spatio-temporal information. In this paper, we propose representing space-time patterns using directional spatio-temporal oriented gradients. In the proposed approach, a 3D video patch is represented by a histogram of oriented gradients over nine symmetric spatio-temporal planes. Video comparison is achieved through a positive definite similarity kernel that is learnt by multiple kernel learning. A rich spatio-temporal descriptor with a simple trade-off between discriminatory power and invariance properties is thereby obtained. To evaluate the proposed approach, we consider three challenging visual recognition tasks, namely the classification of dynamic textures, human gestures and human actions. Our evaluations indicate that the proposed approach attains significant classification improvements in recognition accuracy in comparison to state-of-the-art methods such as LBP-TOP, 3D-SIFT, HOG3D, tensor canonical correlation analysis, and dynamical fractal analysis.

[S3-P4B]

Polynomial Regression on Riemannian Manifolds

Jacob Hinkle, Prasanna Muralidharan, P. Thomas Fletcher, and Sarang Joshi

In this paper we develop the theory of parametric polynomial regression in Riemannian manifolds. The theory enables parametric analysis in a wide range of applications, including rigid and non-rigid kinematics as well as shape change of organs due to growth and aging. We show application of Riemannian polynomial regression to shape analysis in Kendall shape space. Results are presented, showing the power of polynomial regression on the classic rat skull growth data of Bookstein and the analysis of the shape changes associated with aging of the corpus callosum from the OASIS Alzheimer's study.

[S3-P5B]

Geodesic Saliency Using Background Priors

Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun

Generic object level saliency detection is important for many vision tasks. Previous approaches are mostly built on the prior that "appearance contrast between objects and backgrounds is high". Although various computational models have been developed, the problem remains challenging and huge behavioral discrepancies between previous approaches can be observed. This suggest that the problem may still be highly ill-posed by using this prior only. In this work, we tackle the problem from a different viewpoint: we focus more on the background instead of the object. We exploit two common priors about backgrounds in natural images, namely boundary and connectivity priors, to provide more clues for the problem. Accordingly, we propose a novel saliency measure called geodesic saliency. It is intuitive, easy to interpret and allows fast implementation. Furthermore, it is complementary to previous approaches, because it benefits more from background priors while previous approaches do not. Evaluation on two databases validates that geodesic saliency achieves superior results and outperforms previous approaches by a large margin, in both accuracy and speed (2 ms per image). This illustrates that appropriate prior exploitation is helpful for the ill-posed saliency detection problem.

[S3-P6B]

Joint Face Alignment with Non-parametric Shape Models

Brandon M. Smith and Li Zhang

We present a joint face alignment technique that takes a set of images as input and produces a set of shape- and appearance-consistent face alignments as output. Our method is an extension of the recent localization method of Belhumeur et al. [1], which combines the output of local detectors with a non-parametric set of face shape models. We are inspired by the recent joint alignment method of Zhao et al. [20], which employs a modified Active Appearance Model (AAM) approach to align a batch of images. We introduce an approach for simultaneously optimizing both a local appearance constraint, which couples the local estimates between multiple images, and a global shape constraint, which couples landmarks and images across the image set. In video sequences, our method greatly improves the temporal stability of landmark estimates without compromising accuracy relative to ground truth.

[S3-P7B]

Discriminative Bayesian Active Shape Models

Pedro Martins, Rui Caseiro, João F. Henriques, and Jorge Batista

This work presents a simple and very efficient solution to align facial parts in unseen images. Our solution relies on a Point Distribution Model (PDM) face model and a set of discriminant local detectors, one for each facial landmark. The patch responses can be embedded into a Bayesian inference problem, where the posterior distribution of the global warp is inferred in a maximum a posteriori (MAP) sense. However, previous formulations do not model explicitly the covariance of the latent variables, which represents the confidence in the current solution. In our Discriminative Bayesian Active Shape Model (DBASM) formulation, the MAP global alignment is inferred by a Linear Dynamical System (LDS) that takes this information into account. The Bayesian paradigm provides an effective fitting strategy, since it combines in the same framework both the shape prior and multiple sets of patch alignment classifiers to further improve the accuracy. Extensive evaluations were performed on several datasets including the challenging Labeled Faces in the Wild (LFW). Face parts descriptors were also evaluated, including the recently proposed Minimum Output Sum of Squared Error (MOSSE) filter. The proposed Bayesian optimization strategy improves on the state-of-the-art while using the same local detectors. We also show that MOSSE filters further improve on these results.

[S3-P8B]

Patch Based Synthesis for Single Depth Image Super-Resolution

Oisin Mac Aodha, Neill D.F. Campbell, Arun Nair, and Gabriel J. Brostow

We present an algorithm to synthetically increase the resolution of a solitary depth image using only a generic database of local patches. Modern range sensors measure depths with non-Gaussian noise and at lower starting resolutions than typical visible-light cameras. While patch based approaches for upsampling intensity images continue to improve, this is the first exploration of patching for depth images. We match against the height field of each low resolution input depth patch, and search our database for a list of appropriate high resolution candidate patches. Selecting the right candidate at each location in the depth image is then posed as a Markov random field labeling problem. Our experiments also show how important further depth-specific processing, such as noise removal and correct patch normalization, dramatically improves our results. Perhaps surprisingly, even better results are achieved on a variety of real test scenes by providing our algorithm with only synthetic training depth data.

[S3-P9B]

Annotation Propagation in Large Image Databases via Dense Image Correspondence

Michael Rubinstein, Ce Liu, and William T. Freeman

Our goal is to automatically annotate many images with a set of word tags and a pixel-wise map showing where each word tag occurs. Most previous approaches rely on a corpus of training images where each pixel is labeled. However, for large image databases, pixel labels are expensive to obtain and are often unavailable. Furthermore, when classifying multiple images, each image is typically solved for independently, which often results in inconsistent annotations across similar images. In this work, we incorporate dense image correspondence into the annotation model, allowing us to make do with significantly less labeled data and to resolve ambiguities by propagating inferred annotations from images with strong local visual evidence to images with weaker local evidence. We establish a large graphical model spanning all labeled and unlabeled images, then solve it to infer annotations, enforcing consistent annotations over similar visual patterns. Our model is optimized by efficient belief propagation algorithms embedded in an expectation-maximization (EM) scheme. Extensive experiments are conducted to evaluate the performance on several standard large-scale image datasets, showing that the proposed framework outperforms state-of-the-art methods.

[S3-P10B]

Numerically Stable Optimization of Polynomial Solvers for Minimal Problems

Yubin Kuang and Kalle Åström

Numerous geometric problems in computer vision involve the solution of systems of polynomial equations. This is particularly true for so called minimal problems, but also for finding stationary points for overdetermined problems. The state-of-the-art is based on the use of numerical linear algebra on the large but sparse coefficient matrix that represents the original equations multiplied with a set of monomials. The key observation in this paper is that the speed and numerical stability of the solver depends heavily on (i) what multiplication monomials are used and (ii) the set of so called permissible monomials from which numerical linear algebra routines choose the basis of a certain quotient ring. In the paper we show that optimizing with respect to these two factors can give both significant improvements to numerical stability as compared to the state of the art, as well as highly compact solvers, while still retaining numerical stability. The methods are validated on several minimal problems that have previously been shown to be challenging with improvement over the current state of the art.

[S3-P11B]

Has My Algorithm Succeeded? An Evaluator for Human Pose Estimators

Nataraj Jammalamadaka, Andrew Zisserman, Marcin Eichner, Vittorio Ferrari, and C.V. Jawahar

Most current vision algorithms deliver their output 'as is', without indicating whether it is correct or not. In this paper we propose evaluator algorithms that predict if a vision algorithm has succeeded. We illustrate this idea for the case of Human Pose Estimation (HPE). We describe the stages required to learn and test an evaluator, including the use of an annotated ground truth dataset for training and testing the evaluator (and we provide a new dataset for the HPE case), and the development of auxiliary features that have not been used by the (HPE) algorithm, but can be learnt by the evaluator to predict if the output is correct or not. Then an evaluator is built for each of four recently developed HPE algorithms using their publicly available implementations: Eichner and Ferrari [5], Sapp et al. [16], Andriluka et al. [2] and Yang and Ramanan [22]. We demonstrate that in each case the evaluator is able to predict if the algorithm has correctly estimated the pose or not.

[S3-P12B]

Group Tracking: Exploring Mutual Relations for Multiple Object Tracking

Genquan Duan, Haizhou Ai, Song Cao, and Shihong Lao

In this paper, we propose to track multiple previously unseen objects in unconstrained scenes. Instead of considering objects individually, we model objects in mutual context with each other to benefit robust and accurate tracking. We introduce a unified framework to combine both Individual Object Models (IOMs) and Mutual Relation Models (MRMs). The MRMs consist of three components, the relational graph to indicate related objects, the mutual relation vectors calculated within related objects to show the interactions, and the relational weights to balance all interactions and IOMs. As MRMs are varying along temporal sequences, we propose online algorithms to make MRMs adapt to current situations. We update relational graphs through analyzing object trajectories and cast the relational weight learning task as an online latent SVM problem. Extensive experiments on challenging real world video sequences demonstrate the efficiency and effectiveness of our framework.

[S3-P13B]

A Discrete Chain Graph Model for 3d+t Cell Tracking with High Misdetection Robustness

Bernhard X. Kausler, Martin Schiegg, Bjoern Andres, Martin Lindner, Ullrich Koethe, Heike Leitte, Jochen Wittbrodt, Lars Hufnagel, and Fred A. Hamprecht

Tracking by assignment is well suited for tracking a varying number of divisible cells, but suffers from false positive detections. We reformulate tracking by assignment as a chain graph–a mixed directed-undirected probabilistic graphical model–and obtain a tracking simultaneously over all time steps from the maximum a-posteriori configuration. The model is evaluated on two challenging four-dimensional data sets from developmental biology. Compared to previous work, we obtain improved tracks due to an increased robustness against false positive detections and the incorporation of temporal domain knowledge.

[S3-P14B]

Robust Tracking with Weighted Online Structured Learning

Rui Yao, Qinfeng Shi, Chunhua Shen, Yanning Zhang, and Anton van den Hengel

Robust visual tracking requires constant update of the target appearance model, but without losing track of previous appearance information. One of the difficulties with the online learning approach to this problem has been a lack of flexibility in the modelling of the inevitable variations in target and scene appearance over time. The traditional online learning approach to the problem treats each example equally, which leads to previous appearances being forgotten too quickly and a lack of emphasis on the most current observations. Through analysis of the visual tracking problem, we develop instead a novel weighted form of online risk which allows more subtlety in its representation. However, the traditional online learning framework does not accommodate this weighted form. We thus also propose a principled approach to weighted online learning using weighted reservoir sampling and provide a weighted regret bound as a theoretical guarantee of performance. The proposed novel online learning framework can handle examples with different importance weights for binary, multiclass, and even structured output labels in both linear and non-linear kernels. Applying the method to tracking results in an algorithm which is both efficient and accurate even in the presence of severe appearance changes. Experimental results show that the proposed tracker outperforms the current state-of-the-art.

[S3-P15B]

Fast Regularization of Matrix-Valued Images

Guy Rosman, Yu Wang, Xue-Cheng Tai, Ron Kimmel, and Alfred M. Bruckstein

Regularization of images with matrix-valued data is important in medical imaging, motion analysis and scene understanding. We propose a novel method for fast regularization of matrix group-valued images. Using the augmented Lagrangian framework we separate total- variation regularization of matrix-valued images into a regularization and a projection steps. Both steps are computationally efficient and easily parallelizable, allowing real-time regularization of matrix valued images on a graphic processing unit. We demonstrate the effectiveness of our method for smoothing several group-valued image types, with applications in directions diffusion, motion analysis from depth sensors, and DT-MRI denoising.

[S3-P16B]

Blind Correction of Optical Aberrations

Christian J. Schuler, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf

Camera lenses are a critical component of optical imaging systems, and lens imperfections compromise image quality. While traditionally, sophisticated lens design and quality control aim at limiting optical aberrations, recent works [1,2,3] promote the correction of optical flaws by computational means. These approaches rely on elaborate measurement procedures to characterize an optical system, and perform image correction by non-blind deconvolution. In this paper, we present a method that utilizes physically plausible assumptions to estimate non-stationary lens aberrations blindly, and thus can correct images without knowledge of specifics of camera and lens. The blur estimation features a novel preconditioning step that enables fast deconvolution. We obtain results that are competitive with state-of-the-art non-blind approaches.

[S3-P17B]

Inverse Rendering of Faces on a Cloudy Day

Oswald Aldrian and William A.P. Smith

In this paper we consider the problem of inverse rendering faces under unknown environment illumination using a morphable model. In contrast to previous approaches, we account for global illumination effects by incorporating statistical models for ambient occlusion and bent normals into our image formation model. We show that solving for ambient occlusion and bent normal parameters as part of the fitting process improves the accuracy of the estimated texture map and illumination environment. We present results on challenging data, rendered under complex natural illumination with both specular reflectance and occlusion of the illumination environment.

[S3-P18B]

On Tensor-Based PDEs and Their Corresponding Variational Formulations with Application to Color Image Denoising

Freddie Åström, George Baravdish, and Michael Felsberg

The case when a partial differential equation (PDE) can be considered as an Euler-Lagrange (E-L) equation of an energy functional, consisting of a data term and a smoothness term is investigated. We show the necessary conditions for a PDE to be the E-L equation for a corresponding functional. This energy functional is applied to a color image denoising problem and it is shown that the method compares favorably to current state-of-the-art color image denoising techniques.

[S3-P19B]

Kernelized Temporal Cut for Online Temporal Segmentation and Recognition

Dian Gong, Gérard Medioni, Sikai Zhu, and Xuemei Zhao

We address the problem of unsupervised online segmenting human motion sequences into different actions. Kernelized Temporal Cut (KTC), is proposed to sequentially cut the structured sequential data into different regimes. KTC extends previous works on online change-point detection by incorporating Hilbert space embedding of distributions to handle the nonparametric and high dimensionality issues. Based on KTC, a realtime online algorithm and a hierarchical extension are proposed for detecting both action transitions and cyclic motions at the same time. We evaluate and compare the approach to state-of-the-art methods on motion capture data, depth sensor data and videos. Experimental results demonstrate the effectiveness of our approach, which yields realtime segmentation, and produces higher action segmentation accuracy. Furthermore, by combining with sequence matching algorithms, we can online recognize actions of an arbitrary person from an arbitrary viewpoint, given realtime depth sensor input.

[S3-P20B]

Grain Segmentation of 3D Superalloy Images Using Multichannel EWCVT under Human Annotation Constraints

Yu Cao, Lili Ju, and Song Wang

Grain segmentation on 3D superalloy images provides superalloy's micro-structures, based on which many physical and mechanical properties can be evaluated. This is a challenging problem in senses of (1) the number of grains in a superalloy sample could be thousands or even more; (2) the intensity within a grain may not be homogeneous; and (3) superalloy images usually contains carbides and noises. Recently, the Multichannel Edge-Weighted Centroid Voronoi Tessellation (MCEWCVT) algorithm [1] was developed for grain segmentation and showed better performance than many widely used image segmentation algorithms. However, as a general-purpose clustering algorithm, MCEWCVT does not consider possible prior knowledge from material scientists in the process of grain segmentation. In this paper, we address this issue by defining an energy minimization problem which subject to certain constraints. Then we develop a Constrained Multichannel Edge-Weighted Centroid Voronoi Tessellation (CMEWCVT) algorithm to effectively solve this constrained minimization problem. In particular, manually annotated segmentation on a very small set of 2D slices are taken as constraints and incorporated into the whole clustering process. Experimental results demonstrate that the proposed CMEWCVT algorithm significantly improve the previous grain-segmentation performance.

[S3-P21B]

Hough Regions for Joining Instance Localization and Segmentation

Hayko Riemenschneider, Sabine Sternig, Michael Donoser, Peter M. Roth, and Horst Bischof

Object detection and segmentation are two challenging tasks in computer vision, which are usually considered as independent steps. In this paper, we propose a framework which jointly optimizes for both tasks and implicitly provides detection hypotheses and corresponding segmentations. Our novel approach is attachable to any of the available generalized Hough voting methods. We introduce Hough Regions by formulating the problem of Hough space analysis as Bayesian labeling of a random field. This exploits provided classifier responses, object center votes and low-level cues like color consistency, which are combined into a global energy term. We further propose a greedy approach to solve this energy minimization problem providing a pixel-wise assignment to background or to a specific category instance. This way we bypass the parameter sensitive non-maximum suppression that is required in related methods. The experimental evaluation demonstrates that state-of-the-art detection and segmentation results are achieved and that our method is inherently able to handle overlapping instances and an increased range of articulations, aspect ratios and scales.

[S3-P22B]

Learning to Segment a Video to Clips Based on Scene and Camera Motion

Adarsh Kowdle and Tsuhan Chen

In this paper, we present a novel learning-based algorithm for temporal segmentation of a video into clips based on both camera and scene motion, in particular, based on combinations of static vs. dynamic camera and static vs. dynamic scene. Given a video, we first perform shot boundary detection to segment the video to shots. We enforce temporal continuity by constructing a Markov Random Field (MRF) over the frames of each video shot with edges between consecutive frames and cast the segmentation problem as a frame level discrete labeling problem. Using manually labeled data we learn classifiers exploiting cues from optical flow to provide evidence for the different labels, and infer the best labeling over the frames. We show the effectiveness of the approach using user videos and full-length movies. Using sixty full-length movies spanning 50 years, we show that the proposed algorithm of grouping frames purely based on motion cues can aid computational applications such as recovering depth from a video and also reveal interesting trends in movies, which finds itself interesting novel applications in video analysis (time-stamping archive movies) and film studies.

[S3-P23B]

Evaluation of Image Segmentation Quality by Adaptive Ground Truth Composition

Bo Peng and Lei Zhang

Segmenting an image is an important step in many computer vision applications. However, image segmentation evaluation is far from being well-studied in contrast to the extensive studies on image segmentation algorithms. In this paper, we propose a framework to quantitatively evaluate the quality of a given segmentation with multiple ground truth segmentations. Instead of comparing directly the given segmentation to the ground truths, we assume that if a segmentation is "good", it can be constructed by pieces of the ground truth segmentations. Then for a given segmentation, we construct adaptively a new ground truth which can be locally matched to the segmentation as much as possible and preserve the structural consistency in the ground truths. The quality of the segmentation can then be evaluated by measuring its distance to the adaptively composite ground truth. To the best of our knowledge, this is the first work that provides a framework to adaptively combine multiple ground truths for quantitative segmentation evaluation. Experiments are conducted on the benchmark Berkeley segmentation database, and the results show that the proposed method can faithfully reflect the perceptual qualities of segmentations.

ORAL SESSION 3 DETECTION AND ATTRIBUTES

Tuesday, October 9
11:20 - 13:00

[S3-O1]

Exact Acceleration of Linear Object Detectors

Charles Dubout and François Fleuret

We describe a general and exact method to considerably speed up linear object detection systems operating in a sliding, multi-scale window fashion, such as the individual part detectors of part-based models. The main bottleneck of many of those systems is the computational cost of the convolutions between the multiple rescalings of the image to process, and the linear filters. We make use of properties of the Fourier transform and of clever implementation strategies to obtain a speedup factor proportional to the filters' sizes. The gain in performance is demonstrated on the well known Pascal VOC benchmark, where we accelerate the speed of said convolutions by an order of magnitude.

[S3-O2]

Latent Hough Transform for Object Detection

Nima Razavi, Juergen Gall, Pushmeet Kohli, and Luc van Gool

Hough transform based methods for object detection work by allowing image features to vote for the location of the object. While this representation allows for parts observed in different training instances to support a single object hypothesis, it also produces false positives by accumulating votes that are consistent in location but inconsistent in other properties like pose, color, shape or type. In this work, we propose to augment the Hough transform with latent variables in order to enforce consistency among votes. To this end, only votes that agree on the assignment of the latent variable are allowed to support a single hypothesis. For training a Latent Hough Transform (LHT) model, we propose a learning scheme that exploits the linearity of the Hough transform based methods. Our experiments on two datasets including the challenging PASCAL VOC 2007 benchmark show that our method outperforms traditional Hough transform based methods leading to state-of-the-art performance on some categories.

[S3-O3]

Using Linking Features in Learning Non-parametric Part Models

Leonid Karlinsky and Shimon Ullman

We present an approach to the detection of parts of highly deformable objects, such as the human body. Instead of using kinematic constraints on relative angles used by most existing approaches for modeling part-to-part relations, we learn and use special observed 'linking' features that support particular pairwise part configurations. In addition to modeling the appearance of individual parts, the current approach adds modeling of the appearance of part-linking, which is shown to provide useful information. For example, configurations of the lower and upper arms are supported by observing corresponding appearances of the elbow or other relevant features. The proposed model combines the support from all the linking features observed in a test image to infer the most likely joint configuration of all the parts of interest. The approach is trained using images with annotated parts, but no a-priori known part connections or connection parameters are assumed, and the linking features are discovered automatically during training. We evaluate the performance of the proposed approach on two challenging human body parts detection datasets, and obtain performance comparable, and in some cases superior, to the state-of-the-art. In addition, the approach generality is shown by applying it without modification to part detection on datasets of animal parts and of facial fiducial points.

[S3-O4]

Diagnosing Error in Object Detectors

Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai

This paper shows how to analyze the influences of object characteristics on detection performance and the frequency and impact of different types of false positives. In particular, we examine effects of occlusion, size, aspect ratio, visibility of parts, viewpoint, localization error, and confusion with semantically similar objects, other labeled objects, and background. We analyze two classes of detectors: the Vedaldi et al. multiple kernel learning detector and different versions of the Felzenszwalb et al. detector. Our study shows that sensitivity to size, localization error, and confusion with similar objects are the most impactful forms of error. Our analysis also reveals that many different kinds of improvement are necessary to achieve large gains, making more detailed analysis essential for the progress of recognition research. By making our software and annotations available, we make it effortless for future researchers to perform similar analysis.

[S3-O5]

Attributes for Classifier Feedback

Amar Parkash and Devi Parikh

Traditional active learning allows a (machine) learner to query the (human) teacher for labels on examples it finds confusing. The teacher then provides a label for only that instance. This is quite restrictive. In this paper, we propose a learning paradigm in which the learner communicates its belief (i.e. predicted label) about the actively chosen example to the teacher. The teacher then confirms or rejects the predicted label. More importantly, if rejected, the teacher communicates an explanation for why the learner's belief was wrong. This explanation allows the learner to propagate the feedback provided by the teacher to many unlabeled images. This allows a classifier to better learn from its mistakes, leading to accelerated discriminative learning of visual concepts even with few labeled images. In order for such communication to be feasible, it is crucial to have a language that both the human supervisor and the machine learner understand. Attributes provide precisely this channel. They are human-interpretable mid-level visual concepts shareable across categories e.g. ''furry'', ''spacious'', etc. We advocate the use of attributes for a supervisor to provide feedback to a classifier and directly communicate his knowledge of the world. We employ a straightforward approach to incorporate this feedback in the classifier, and demonstrate its power on a variety of visual recognition scenarios such as image classification and annotation. This application of attributes for providing classifiers feedback is very powerful, and has not been explored in the community. It introduces a new mode of supervision, and opens up several avenues for future research.

[S3-O6]

Constrained Semi-Supervised Learning Using Attributes and Comparative Attributes

Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta

We consider the problem of semi-supervised bootstrap learning for scene categorization. Existing semi-supervised approaches are typically unreliable and face semantic drift because the learning task is under-constrained. This is primarily because they ignore the strong interactions that often exist between scene categories, such as the common attributes shared across categories as well as the attributes which make one scene different from another. The goal of this paper is to exploit these relationships and constrain the semi-supervised learning problem. For example, the knowledge that an image is an auditorium can improve labeling of amphitheaters by enforcing constraint that an amphitheater image should have more circular structures than an auditorium image. We propose constraints based on mutual exclusion, binary attributes and comparative attributes and show that they help us to constrain the learning problem and avoid semantic drift. We demonstrate the effectiveness of our approach through extensive experiments, including results on a very large dataset of one million images.

POSTER SESSION 4

Tuesday, October 9
14:30 - 17:00

[S4-P1A]

Renormalization Returns: Hyper-renormalization and Its Applications

Kenichi Kanatani, Ali Al-Sharadqah, Nikolai Chernov, and Yasuyuki Sugaya

The technique of "renormalization" for geometric estimation attracted much attention when it was proposed in early 1990s for having higher accuracy than any other then known methods. Later, it was replaced by minimization of the reprojection error. This paper points out that renormalization can be modified so that it outperforms reprojection error minimization. The key fact is that renormalization directly specifies equations to solve, just as the "estimation equation" approach in statistics, rather than minimizing some cost. Exploiting this fact, we modify the problem so that the solution has zero bias up to high order error terms; we call the resulting scheme hyper-renormalization. We apply it to ellipse fitting to demonstrate that it indeed surpasses reprojection error minimization. We conclude that it is the best method available today.

[S4-P2A]

Scale Robust Multi View Stereo

Christian Bailer, Manuel Finckh, and Hendrik P.A. Lensch

We present a Multi View Stereo approach for huge unstructured image datasets that can deal with large variations in surface sampling rate of single images. Our method reconstructs surface parts always in the best available resolution. It considers scaling not only for large scale differences, but also between arbitrary small ones for a weighted merging of the best partial reconstructions. We create depth maps with our GPU based depth map algorithm, that also performs normal optimization. It matches several images that are found with a heuristic image selection method, to a reference image. We remove outliers by comparing depth maps against each other with a fast but reliable GPU approach. Then, we merge the different reconstructions from depth maps in 3D space by selecting the best points and optimizing them with not selected points. Finally, we create the surface by using a Delaunay graph cut.

[S4-P3A]

Laplacian Meshes for Monocular 3D Shape Recovery

Jonas Östlund, Aydin Varol, Dat Tien Ngo, and Pascal Fua

We show that by extending the Laplacian formalism, which was first introduced in the Graphics community to regularize 3D meshes, we can turn the monocular 3D shape reconstruction of a deformable surface given correspondences with a reference image into a well-posed problem. Furthermore, this does not require any training data and eliminates the need to pre-align the reference shape with the one to be reconstructed, as was done in earlier methods.

[S4-P4A]

Soft Inextensibility Constraints for Template-Free Non-rigid Reconstruction

Sara Vicente and Lourdes Agapito

In this paper, we exploit an inextensibility prior as a way to better constrain the highly ambiguous problem of non-rigid reconstruction from monocular views. While this widely applicable prior has been used before combined with the strong assumption of a known 3D-template, our work achieves template-free reconstruction using only inextensibility constraints. We show how to formulate an energy function that includes soft inextensibility constraints and rely on existing discrete optimisation methods to minimise it. Our method has all of the following advantages: (i) it can be applied to two tasks that have been so far considered independently – template based reconstruction and non-rigid structure from motion – producing comparable or better results than the state-of-the art methods; (ii) it can perform template-free reconstruction from as few as two images; and (iii) it does not require post-processing stitching or surface smoothing.

[S4-P5A]

Spatiotemporal Descriptor for Wide-Baseline Stereo Reconstruction of Non-rigid and Ambiguous Scenes

Eduard Trulls, Alberto Sanfeliu, and Francesc Moreno-Noguer

This paper studies the use of temporal consistency to match appearance descriptors and handle complex ambiguities when computing dynamic depth maps from stereo. Previous attempts have designed 3D descriptors over the spacetime volume and have been mostly used for monocular action recognition, as they cannot deal with perspective changes. Our approach is based on a state-of-the-art 2D dense appearance descriptor which we extend in time by means of optical flow priors, and can be applied to wide-baseline stereo setups. The basic idea behind our approach is to capture the changes around a feature point in time instead of trying to describe the spatiotemporal volume. We demonstrate its effectiveness on very ambiguous synthetic video sequences with ground truth data, as well as real sequences.

[S4-P6A]

Elevation Angle from Reflectance Monotonicity: Photometric Stereo for General Isotropic Reflectances

Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi

This paper exploits the monotonicity of general isotropic reflectances for estimating elevation angles of surface normal given the azimuth angles. With an assumption that the reflectance includes at least one lobe that is a monotonic function of the angle between the surface normal and half-vector (bisector of lighting and viewing directions), we prove that elevation angles can be uniquely determined when the surface is observed under varying directional lights densely and uniformly distributed over the hemisphere. We evaluate our method by experiments using synthetic and real data to show its wide applicability, even when the assumption does not strictly hold. By combining an existing method for azimuth angle estimation, our method derives complete surface normal estimates for general isotropic reflectances.

[S4-P7A]

Local Log-Euclidean Covariance Matrix (L^2ECM) for Image Representation and Its Applications

Peihua Li and Qilong Wang

This paper presents Local Log-Euclidean Covariance Matrix (L^2ECM) to represent neighboring image properties by capturing correlation of various image cues. Our work is inspired by the structure tensor which computes the second-order moment of image gradients for representing local image properties, and the Diffusion Tensor Imaging which produces tensor-valued image characterizing the local tissue structure. Our approach begins with extraction of raw features consisting of multiple image cues. For each pixel we compute a covariance matrix in its neighboring region, producing a tensor-valued image. The covariance matrices are symmetric and positive-definite (SPD) which forms a Riemannian manifold. In the Log-Euclidean framework, the SPD matrices form a Lie group equipped with Euclidean space structure, which enables common Euclidean operations in the logarithm domain. Hence, we compute the covariance matrix logarithm, obtaining the pixel-wise symmetric matrix. After half-vectorization we obtain the vector-valued L^2ECM image, which can be flexibly handled with Euclidean operations while preserving the geometric structure of SPD matrices. The L^2ECM features can be used in diverse image or vision tasks. We demonstrate some applications of its statistical modeling by simple second-order central moment and achieve promising performance.

[S4-P8A]

Ensemble Partitioning for Unsupervised Image Categorization

Dengxin Dai, Mukta Prasad, Christian Leistner, and Luc Van Gool

While the quality of object recognition systems can strongly benefit from more data, human annotation and labeling can hardly keep pace. This motivates the usage of autonomous and unsupervised learning methods. In this paper, we present a simple, yet effective method for unsupervised image categorization, which relies on discriminative learners. Since automatically obtaining error-free labeled training data for the learners is infeasible, we propose the concept of weak training (WT) set. WT sets have various deficiencies, but still carry useful information. Training on a single WT set cannot result in good performance, thus we design a random walk sampling scheme to create a series of diverse WT sets. This naturally allows our categorization learning to leverage ensemble learning techniques. In particular, for each WT set, we train a max-margin classifier to further partition the whole dataset to be categorized. By doing so, each WT set leads to a base partitioning of the dataset and all the base partitionings are combined into an ensemble proximity matrix. The final categorization is completed by feeding this proximity matrix into a spectral clustering algorithm. Experiments on a variety of challenging datasets show that our method outperforms competing methods by a considerable margin.

[S4-P9A]

Set Based Discriminative Ranking for Recognition

Yang Wu, Michihiko Minoh, Masayuki Mukunoki, and Shihong Lao

Recently both face recognition and body-based person re-identification have been extended from single-image based scenarios to video-based or even more generally image-set based problems. Set-based recognition brings new research and application opportunities while at the same time raises great modeling and optimization challenges. How to make the best use of the available multiple samples for each individual while at the same time not be disturbed by the great within-set variations is considered by us to be the major issue. Due to the difficulty of designing a global optimal learning model, most existing solutions are still based on unsupervised matching, which can be further categorized into three groups: a) set-based signature generation, b) direct set-to-set matching, and c) between-set distance finding. The first two count on good feature representation while the third explores data set structure and set-based distance measurement. The main shortage of them is the lack of learning-based discrimination ability. In this paper, we propose a set-based discriminative ranking model (SBDR), which iterates between set-to-set distance finding and discriminative feature space projection to achieve simultaneous optimization of these two. Extensive experiments on widely-used face recognition and person re-identification datasets not only demonstrate the superiority of our approach, but also shed some light on its properties and application domain.

[S4-P10A]

A Global Hypotheses Verification Method for 3D Object Recognition

Aitor Aldoma, Federico Tombari, Luigi Di Stefano, and Markus Vincze

We propose a novel approach for verifying model hypotheses in cluttered and heavily occluded 3D scenes. Instead of verifying one hypothesis at a time, as done by most state-of-the-art 3D object recognition methods, we determine object and pose instances according to a global optimization stage based on a cost function which encompasses geometrical cues. Peculiar to our approach is the inherent ability to detect significantly occluded objects without increasing the amount of false positives, so that the operating point of the object recognition algorithm can nicely move toward a higher recall without sacrificing precision. Our approach outperforms state-of-the-art on a challenging dataset including 35 household models obtained with the Kinect sensor, as well as on the standard 3D object recognition benchmark dataset.

[S4-P11A]

Are You Really Smiling at Me? Spontaneous versus Posed Enjoyment Smiles

Hamdi Dibeklioglu, Albert Ali Salah, and Theo Gevers

Smiling is an indispensable element of nonverbal social interaction. Besides, automatic distinction between spontaneous and posed expressions is important for visual analysis of social signals. Therefore, in this paper, we propose a method to distinguish between spontaneous and posed enjoyment smiles by using the dynamics of eyelid, cheek, and lip corner movements. The discriminative power of these movements, and the effect of different fusion levels are investigated on multiple databases. Our results improve the state-of-the-art. We also introduce the largest spontaneous/posed enjoyment smile database collected to date, and report new empirical and conceptual findings on smile dynamics. The collected database consists of 1240 samples of 400 subjects. Moreover, it has the unique property of having an age range from 8 to 76 years. Large scale experiments on the new database indicate that eyelid dynamics are highly relevant for smile classification, and there are age-related differences in smile dynamics.

[S4-P12A]

Efficient Monte Carlo Sampler for Detecting Parametric Objects in Large Scenes

Yannick Verdié and Florent Lafarge

Point processes have demonstrated efficiency and competitiveness when addressing object recognition problems in vision. However, simulating these mathematical models is a difficult task, especially on large scenes. Existing samplers suffer from average performances in terms of computation time and stability. We propose a new sampling procedure based on a Monte Carlo formalism. Our algorithm exploits Markovian properties of point processes to perform the sampling in parallel. This procedure is embedded into a data-driven mechanism such that the points are non-uniformly distributed in the scene. The performances of the sampler are analyzed through a set of experiments on various object recognition problems from large scenes, and through comparisons to the existing algorithms.

[S4-P13A]

Supervised Geodesic Propagation for Semantic Label Transfer

Xiaowu Chen, Qing Li, Yafei Song, Xin Jin, and Qinping Zhao

In this paper we propose a novel semantic label transfer method using supervised geodesic propagation (SGP). We use supervised learning to guide the seed selection and the label propagation. Given an input image, we first retrieve its similar image set from annotated databases. A Joint Boost model is learned on the similar image set of the input image. Then the recognition proposal map of the input image is inferred by this learned model. The initial distance map is defined by the proposal map: the higher probability, the smaller distance. In each iteration step of the geodesic propagation, the seed is selected as the one with the smallest distance from the undetermined superpixels. We learn a classifier as an indicator to indicate whether to propagate labels between two neighboring superpixels. The training samples of the indicator are annotated neighboring pairs from the similar image set. The geodesic distances of its neighbors are updated according to the combination of the texture and boundary features and the indication value. Experiments on three datasets show that our method outperforms the traditional learning based methods and the previous label transfer method for the semantic segmentation work.

[S4-P14A]

Bayesian Face Revisited: A Joint Formulation

Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun

In this paper, we revisit the classical Bayesian face recognition method by Baback Moghaddam et al. and propose a new joint formulation. The classical Bayesian method models the appearance difference between two faces. We observe that this "difference" formulation may reduce the separability between classes. Instead, we model two faces jointly with an appropriate prior on the face representation. Our joint formulation leads to an EM-like model learning at the training time and an efficient, closed-formed computation at the test time. On extensive experimental evaluations, our method is superior to the classical Bayesian face and many other supervised approaches. Our method achieved 92.4% test accuracy on the challenging Labeled Face in Wild (LFW) dataset. Comparing with current best commercial system, we reduced the error rate by 10%.

[S4-P15A]

Beyond Bounding-Boxes: Learning Object Shape by Model-Driven Grouping

Antonio Monroy and Björn Ommer

Visual recognition requires to learn object models from training data. Commonly, training samples are annotated by marking only the bounding-box of objects, since this appears to be the best trade-off between labeling information and effectiveness. However, objects are typically not box-shaped. Thus, the usual parametrization of object hypotheses by only their location, scale and aspect ratio seems inappropriate since the box contains a significant amount of background clutter. Most important, however, is that object shape becomes only explicit once objects are segregated from the background. Segmentation is an ill-posed problem and so we propose an approach for learning object models for detection while, simultaneously, learning to segregate objects from clutter and extracting their overall shape. For this purpose, we exclusively use bounding-box annotated training data. The approach groups fragmented object regions using the Multiple Instance Learning (MIL) framework to obtain a meaningful representation of object shape which, at the same time, crops away distracting background clutter to improve the appearance representation.

[S4-P16A]

In Defence of Negative Mining for Annotating Weakly Labelled Data

Parthipan Siva, Chris Russell, and Tao Xiang

We propose a novel approach to annotating weakly labelled data. In contrast to many existing approaches that perform annotation by seeking clusters of self-similar exemplars (minimising intra-class variance), we perform image annotation by selecting exemplars that have never occurred before in the much larger, and strongly annotated, negative training set (maximising inter-class variance). Compared to existing methods, our approach is fast, robust, and obtains state of the art results on two challenging data-sets – voc2007 (all poses), and the msr2 action data-set, where we obtain a 10% increase. Moreover, this use of negative mining complements existing methods, that seek to minimize the intra-class variance, and can be readily integrated with many of them.

[S4-P17A]

Describing Clothing by Semantic Attributes

Huizhong Chen, Andrew Gallagher, and Bernd Girod

Describing clothing appearance with semantic attributes is an appealing technique for many important applications. In this paper, we propose a fully automated system that is capable of generating a list of nameable attributes for clothes on human body in unconstrained images. We extract low-level features in a pose-adaptive manner, and combine complementary features for learning attribute classifiers. Mutual dependencies between the attributes are then explored by a Conditional Random Field to further improve the predictions from independent classifiers. We validate the performance of our system on a challenging clothing attribute dataset, and introduce a novel application of dressing style analysis that utilizes the semantic attributes produced by our system.

[S4-P18A]

Graph Matching via Sequential Monte Carlo

Yumin Suh, Minsu Cho, and Kyoung Mu Lee

Graph matching is a powerful tool for computer vision and machine learning. In this paper, a novel approach to graph matching is developed based on the sequential Monte Carlo framework. By constructing a sequence of intermediate target distributions, the proposed algorithm sequentially performs a sampling and importance resampling to maximize the graph matching objective. Through the sequential sampling procedure, the algorithm effectively collects potential matches under one-to-one matching constraints to avoid the adverse effect of outliers and deformation. Experimental evaluations on synthetic graphs and real images demonstrate its higher robustness to deformation and outliers.

[S4-P19A]

Jet-Based Local Image Descriptors

Anders Boesen Lindbo Larsen, Sune Darkner, Anders Lindbjerg Dahl, and Kim Steenstrup Pedersen

We present a general novel image descriptor based on higherorder differential geometry and investigate the effect of common descriptor choices. Our investigation is twofold in that we develop a jet-based descriptor and perform a comparative evaluation with current state-of-the-art descriptors on the recently released DTU Robot dataset. We demonstrate how the use of higher-order image structures enables us to reduce the descriptor dimensionality while still achieving very good performance. The descriptors are tested in a variety of scenarios including large changes in scale, viewing angle and lighting. We show that the proposed jet-based descriptor is superior to state-of-the-art for DoG interest points and show competitive performance for the other tested interest points.

[S4-P20A]

Abnormal Object Detection by Canonical Scene-Based Contextual Model

Sangdon Park, Wonsik Kim, and Kyoung Mu Lee

Contextual modeling is a critical issue in scene understanding. Object detection accuracy can be improved by exploiting tendencies that are common among object configurations. However, conventional contextual models only exploit the tendencies of normal objects; abnormal objects that do not follow the same tendencies are hard to detect through contextual model. This paper proposes a novel generative model that detects abnormal objects by meeting four proposed criteria of success. This model generates normal as well as abnormal objects, each following their respective tendencies. Moreover, this generation is controlled by a latent scene variable. All latent variables of the proposed model are predicted through optimization via population-based Markov Chain Monte Carlo, which has a relatively short convergence time. We present a new abnormal dataset classified into three categories to thoroughly measure the accuracy of the proposed model for each category; the results demonstrate the superiority of our proposed approach over existing methods.

[S4-P21A]

Shapecollage: Occlusion-Aware, Example-Based Shape Interpretation

Forrester Cole, Phillip Isola, William T. Freeman, Frédo Durand, and Edward H. Adelson

This paper presents an example-based method to interpret a 3D shape from a single image depicting that shape. A major difficulty in applying an example-based approach to shape interpretation is the combinatorial explosion of shape possibilities that occur at occluding contours. Our key technical contribution is a new shape patch representation and corresponding pairwise compatibility terms that allow for flexible matching of overlapping patches, avoiding the combinatorial explosion by allowing patches to explain only the parts of the image they best fit. We infer the best set of localized shape patches over a graph of keypoints at multiple scales to produce a discontinuous shape representation we term a shape collage. To reconstruct a smooth result, we fit a surface to the collage using the predicted confidence of each shape patch. We demonstrate the method on shapes depicted in line drawing, diffuse and glossy shading, and textured styles.

[S4-P22A]

Interactive Facial Feature Localization

Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S. Huang

We address the problem of interactive facial feature localization from a single image. Our goal is to obtain an accurate segmentation of facial features on high-resolution images under a variety of pose, expression, and lighting conditions. Although there has been significant work in facial feature localization, we are addressing a new application area, namely to facilitate intelligent high-quality editing of portraits, that brings requirements not met by existing methods. We propose an improvement to the Active Shape Model that allows for greater independence among the facial components and improves on the appearance fitting step by introducing a Viterbi optimization process that operates along the facial contours. Despite the improvements, we do not expect perfect results in all cases. We therefore introduce an interaction model whereby a user can efficiently guide the algorithm towards a precise solution. We introduce the Helen Facial Feature Dataset consisting of annotated portrait images gathered from Flickr that are more diverse and challenging than currently existing datasets. We present experiments that compare our automatic method to published results, and also a quantitative evaluation of the effectiveness of our interactive method.

[S4-P1B]

Propagative Hough Voting for Human Activity Recognition

Gang Yu, Junsong Yuan, and Zicheng Liu

Hough-transform based voting has been successfully applied to both object and activity detections. However, most current Hough voting methods will suffer when insufficient training data is provided. To address this problem, we propose propagative Hough voting for activity analysis. Instead of letting local features vote individually, we perform feature voting using random projection trees (RPT) which leverage the low-dimension manifold structure to match feature points in the high-dimensional feature space. Our RPT can index the unlabeled feature points in an unsupervised way. After the trees are constructed, the label and spatial-temporal configuration information are propagated from the training samples to the testing data via RPT. The proposed activity recognition method does not rely on human detection and tracking, and can well handle the scale and intra-class variations of the activity patterns. The superior performances on two benchmarked activity datasets validate that our method outperforms the state-of-the-art techniques not only when there is sufficient training data such as in activity recognition, but also when there is limited training data such as in activity search with one query example.

[S4-P2B]

Spatio-Temporal Phrases for Activity Recognition

Yimeng Zhang, Xiaoming Liu, Ming-Ching Chang, Weina Ge, and Tsuhan Chen

The local feature based approaches have become popular for activity recognition. A local feature captures the local movement and appearance of a local region in a video, and thus can be ambiguous; e.g., it cannot tell whether a movement is from a person's hand or foot, when the camera is far away from the person. To better distinguish different types of activities, people have proposed using the combination of local features to encode the relationships of local movements. Due to the computation limit, previous work only creates a combination from neighboring features in space and/or time. In this paper, we propose an approach that efficiently identifies both local and long-range motion interactions; taking the "push" activity as an example, our approach can capture the combination of the hand movement of one person and the foot response of another person, the local features of which are both spatially and temporally far away from each other. Our computational complexity is in linear time to the number of local features in a video. The extensive experiments show that our approach is generically effective for recognizing a wide variety of activities and activities spanning a long term, compared to a number of state-of-the-art methods.

[S4-P3B]

Complex Events Detection Using Data-Driven Concepts

Yang Yang and Mubarak Shah

Automatic event detection in a large collection of unconstrained videos is a challenging and important task. The key issue is to describe long complex video with high level semantic descriptors, which should find the regularity of events in the same category while distinguish those from different categories. This paper proposes a novel unsupervised approach to discover data-driven concepts from multi-modality signals (audio, scene and motion) to describe high level semantics of videos. Our methods consists of three main components: we first learn the low-level features separately from three modalities. Secondly we discover the data-driven concepts based on the statistics of learned features mapped to a low dimensional space using deep belief nets (DBNs). Finally, a compact and robust sparse representation is learned to jointly model the concepts from all three modalities. Extensive experimental results on large in-the-wild dataset show that our proposed method significantly outperforms state-of-the-art methods.

[S4-P4B]

Learning to Recognize Unsuccessful Activities Using a Two-Layer Latent Structural Model

Qiang Zhou and Gang Wang

In this paper, we propose to recognize unsuccessful activities (e.g., one tries to dress himself but fails), which have much more complex temporal structures, as we don't know when the activity performer fails (which is called the point of failure in this paper). We develop a two-layer latent structural SVM model to tackle this problem: the first layer specifies the point of failure, and the second layer specifies the temporal positions of a number of key stages accordingly. The stages before the point of failure are successful stages, while the stages after the point of failure are background stages. Given weakly labeled training data, our training algorithm alternates between inferring the two-layer latent structure and updating the structural SVM parameters. In recognition, our method can not only recognize unsuccessful activities, but also infer the latent structure. We demonstrate the effectiveness of our proposed method on several newly collected datasets.

[S4-P5B]

Action Recognition Using Subtensor Constraint

Qiguang Liu and Xiaochun Cao

Human action recognition from videos draws tremendous interest in the past many years. In this work, we first find that the trifocal tensor resides in a twelve dimensional subspace of the original space if the first two views are already matched and the fundamental matrix between them is known, which we refer to as subtensor. Then we use the subtensor to perform the task of action recognition under three views. We find that treating the two template views separately or not considering the correspondence relation already known between the first two views omits a lot of useful information. Experiments and datasets are designed to demonstrate the effectiveness and improved performance of the proposed approach.

[S4-P6B]

Approximate Gaussian Mixtures for Large Scale Vocabularies

Yannis Avrithis and Yannis Kalantidis

We introduce a clustering method that combines the flexibility of Gaussian mixtures with the scaling properties needed to construct visual vocabularies for image retrieval. It is a variant of expectation-maximization that can converge rapidly while dynamically estimating the number of components. We employ approximate nearest neighbor search to speed-up the E-step and exploit its iterative nature to make search incremental, boosting both speed and precision. We achieve superior performance in large scale retrieval, being as fast as the best known approximate k-means.

[S4-P7B]

Globally Optimal Closed-Surface Segmentation for Connectomics

Bjoern Andres, Thorben Kroeger, Kevin L. Briggman, Winfried Denk, Natalya Korogod, Graham Knott, Ullrich Koethe, and Fred A. Hamprecht

We address the problem of partitioning a volume image into a previously unknown number of segments, based on a likelihood of merging adjacent supervoxels. Towards this goal, we adapt a higher-order probabilistic graphical model that makes the duality between supervoxels and their joint faces explicit and ensures that merging decisions are consistent and surfaces of final segments are closed. First, we propose a practical cutting-plane approach to solve the MAP inference problem to global optimality despite its NP-hardness. Second, we apply this approach to challenging large-scale 3D segmentation problems for neural circuit reconstruction (Connectomics), demonstrating the advantage of this higher-order model over independent decisions and finite-order approximations.

[S4-P8B]

Reduced Analytical Dependency Modeling for Classifier Fusion

Andy Jinhua Ma and Pong Chi Yuen

This paper addresses the independent assumption issue in classifier fusion process. In the last decade, dependency modeling techniques were developed under some specific assumptions which may not be valid in practical applications. In this paper, using analytical functions on posterior probabilities of each feature, we propose a new framework to model dependency without those assumptions. With the analytical dependency model (ADM), we give an equivalent condition to the independent assumption from the properties of marginal distributions, and show that the proposed ADM can model dependency. Since ADM may contain infinite number of undetermined coefficients, we further propose a reduced form of ADM, based on the convergent properties of analytical functions. Finally, under the regularized least square criterion, an optimal Reduced Analytical Dependency Model (RADM) is learned by approximating posterior probabilities such that all training samples are correctly classified. Experimental results show that the proposed RADM outperforms existing classifier fusion methods on Digit, Flower, Face and Human Action databases.

[S4-P9B]

Learning to Match Appearances by Correlations in a Covariance Metric Space

Slawomir Bak, Guillaume Charpiat, Etienne Corvée, François Brémond, and Monique Thonnat

This paper addresses the problem of appearance matching across disjoint camera views. Significant appearance changes, caused by variations in view angle, illumination and object pose, make the problem challenging. We propose to formulate the appearance matching problem as the task of learning a model that selects the most descriptive features for a specific class of objects. Learning is performed in a covariance metric space using an entropy-driven criterion. Our main idea is that different regions of the object appearance ought to be matched using various strategies to obtain a distinctive representation. The proposed technique has been successfully applied to the person re-identification problem, in which a human appearance has to be matched across non-overlapping cameras. We demonstrate that our approach improves state of the art performance in the context of pedestrian recognition.

[S4-P10B]

On the Convergence of Graph Matching: Graduated Assignment Revisited

Yu Tian, Junchi Yan, Hequan Zhang, Ya Zhang, Xiaokang Yang, and Hongyuan Zha

We focus on the problem of graph matching that is fundamental in computer vision and machine learning. Many state-of-the-arts frequently formulate it as integer quadratic programming, which incorporates both unary and second-order terms. This formulation is in general NP-hard thus obtaining an exact solution is computationally intractable. Therefore most algorithms seek the approximate optimum by relaxing techniques. This paper commences with the finding of the "circular" character of solution chain obtained by the iterative Gradient Assignment (via Hungarian method) in the discrete domain, and proposes a method for guiding the solver converging to a fixed point, resulting a convergent algorithm for graph matching in discrete domain. Furthermore, we extend the algorithms to their counterparts in continuous domain, proving the classical graduated assignment algorithm will converge to a double-circular solution chain, and the proposed Soft Constrained Graduated Assignment (SCGA) method will converge to a fixed (discrete) point, both under wild conditions. Competitive performances are reported in both synthetic and real experiments.

[S4-P11B]

Image Annotation Using Metric Learning in Semantic Neighbourhoods

Yashaswi Verma and C.V. Jawahar

Automatic image annotation aims at predicting a set of textual labels for an image that describe its semantics. These are usually taken from an annotation vocabulary of few hundred labels. Because of the large vocabulary, there is a high variance in the number of images corresponding to different labels ("class-imbalance"). Additionally, due to the limitations of manual annotation, a significant number of available images are not annotated with all the relevant labels ("weak-labelling"). These two issues badly affect the performance of most of the existing image annotation models. In this work, we propose 2PKNN, a two-step variant of the classical K-nearest neighbour algorithm, that addresses these two issues in the image annotation task. The first step of 2PKNN uses "image-to-label" similarities, while the second step uses "image-to-image" similarities; thus combining the benefits of both. Since the performance of nearest-neighbour based methods greatly depends on how features are compared, we also propose a metric learning framework over 2PKNN that learns weights for multiple features as well as distances together. This is done in a large margin set-up by generalizing a well-known (single-label) classification metric learning algorithm for multi-label prediction. For scalability, we implement it by alternating between stochastic sub-gradient descent and projection steps. Extensive experiments demonstrate that, though conceptually simple, 2PKNN alone performs comparable to the current state-of-the-art on three challenging image annotation datasets, and shows significant improvements after metric learning.

[S4-P12B]

Dynamic Programming for Approximate Expansion Algorithm

Olga Veksler

Expansion algorithm is a popular optimization method for labeling problems. For many common energies, each expansion step can be optimally solved with a min-cut/max flow algorithm. While the observed performance of max-flow for the expansion algorithm is fast, its theoretical time complexity is worse than linear in the number of pixels. Recently, Dynamic Programming (DP) was shown to be useful for 2D labeling problems via a ''tiered labeling'' algorithm, although the structure of allowed (tiered) is quite restrictive. We show another use of DP in a 2D labeling case. Namely, we use DP for an approximate expansion step. Our expansion-like moves are more limited in the structure than the max-flow expansion moves. In fact, our moves are more restrictive than the tiered labeling structure, but their complexity is linear in the number of pixels, making them extremely efficient in practice. We illustrate the performance of our DP-expansion on the Potts energy, but our algorithm can be used for any pairwise energies. We achieve better efficiency with almost the same energy compared to the max-flow expansion moves.

[S4-P13B]

Real-Time Compressive Tracking

Kaihua Zhang, Lei Zhang, and Ming-Hsuan Yang

It is a challenging task to develop effective and efficient appearance models for robust object tracking due to factors such as pose variation, illumination change, occlusion, and motion blur. Existing online tracking algorithms often update models with samples from observations in recent frames. While much success has been demonstrated, numerous issues remain to be addressed. First, while these adaptive appearance models are data-dependent, there does not exist sufficient amount of data for online algorithms to learn at the outset. Second, online tracking algorithms often encounter the drift problems. As a result of self-taught learning, these mis-aligned samples are likely to be added and degrade the appearance models. In this paper, we propose a simple yet effective and efficient tracking algorithm with an appearance model based on features extracted from the multi-scale image feature space with data-independent basis. Our appearance model employs non-adaptive random projections that preserve the structure of the image feature space of objects. A very sparse measurement matrix is adopted to efficiently extract the features for the appearance model. We compress samples of foreground targets and the background using the same sparse measurement matrix. The tracking task is formulated as a binary classification via a naive Bayes classifier with online update in the compressed domain. The proposed compressive tracking algorithm runs in real-time and performs favorably against state-of-the-art algorithms on challenging sequences in terms of efficiency, accuracy and robustness.

[S4-P14B]

Tracking Feature Points in Uncalibrated Images with Radial Distortion

Miguel Lourenço and João Pedro Barreto

The appearance of moving features in the field-of-view (FoV) of the camera may substantially change due to different camera poses. Typical solutions for tracking image points involve the assumption of an image motion model and the estimation of the motion parameters using image alignment techniques. While for conventional cameras this suffices, the radial distortion that arises in cameras with wide FoV lenses makes the standard motion models inaccurate. In this paper, we propose a set of motion models that implicitly encompass the distortion effect arising in this type of imaging devices. The proposed motion models are included in a standard image alignment framework for performing feature tracking in cameras presenting significant distortion. Consolidation experiments in repeatability and structure-from-motion scenarios show that the proposed RD-KLT trackers significantly improve the tracking performance in images presenting radial distortion, with minimal computational overhead when compared with a state-of-the-art KLT tracker.

[S4-P15B]

Divergence-Free Motion Estimation

Isabelle Herlin, Dominique Béréziat, Nicolas Mercier, and Sergiy Zhuk

This paper describes an innovative approach to estimate motion from image observations of divergence-free flows. Unlike most state-of-the-art methods, which only minimize the divergence of the motion field, our approach utilizes the vorticity-velocity formalism in order to construct a motion field in the subspace of divergence free functions. A 4DVAR-like image assimilation method is used to generate an estimate of the vorticity field given image observations. Given that vorticity estimate, the motion is obtained solving the Poisson equation. Results are illustrated on synthetic image observations and compared to those obtained with state-of-the-art methods, in order to quantify the improvements brought by the presented approach. The method is then applied to ocean satellite data to demonstrate its performance on the real images.

[S4-P16B]

Visual Tracking via Adaptive Tracker Selection with Multiple Features

Ju Hong Yoon, Du Yong Kim, and Kuk-Jin Yoon

In this paper, a robust visual tracking method is proposed to track an object in dynamic conditions that include motion blur, illumination changes, pose variations, and occlusions. To cope with these challenges, multiple trackers with different feature descriptors are utilized, and each of which shows different level of robustness to certain changes in an object's appearance. To fuse these independent trackers, we propose two configurations, tracker selection and interaction. The tracker interaction is achieved based on a transition probability matrix (TPM) in a probabilistic manner. The tracker selection extracts one tracking result from among multiple tracker outputs by choosing the tracker that has the highest tracker probability. According to various changes in an object's appearance, the TPM and tracker probability are updated in a recursive Bayesian form by evaluating each tracker's reliability, which is measured by a robust tracker likelihood function (TLF). When the tracking in each frame is completed, the estimated object's state is obtained and fed into the reference update via the proposed learning strategy, which retains the robustness and adaptability of the TLF and multiple trackers. The experimental results demonstrate that our proposed method is robust in various benchmark scenarios.

[S4-P17B]

Image Enhancement Using Calibrated Lens Simulations

Yichang Shih, Brian Guenter, and Neel Joshi

All lenses have optical aberrations which reduce image sharpness. These aberrations can be reduced by deconvolving an image using the lens point spread function (PSF). However, fully measuring a PSF is laborious and prohibitive. Alternatively, one can simulate the PSF if the lens model is known. However, due to manufacturing tolerances lenses differ subtly from their models, so often a simulated PSF is a poor match to measured data. We present an algorithm that uses a PSF measurement at a single depth to calibrate the nominal lens model to the measured PSF. The calibrated model can then be used to compute the PSF for any desired setting of lens parameters for any scene depth, without additional measurements or calibration. The calibrated model gives deconvolution results comparable to measurement but is much more compact and require hundreds of times fewer calibration images.

[S4-P18B]

Color Constancy, Intrinsic Images, and Shape Estimation

Jonathan T. Barron and Jitendra Malik

We present SIRFS (shape, illumination, and reflectance from shading), the first unified model for recovering shape, chromatic illumination, and reflectance from a single image. Our model is an extension of our previous work [1], which addressed the achromatic version of this problem. Dealing with color requires a modified problem formulation, novel priors on reflectance and illumination, and a new optimization scheme for dealing with the resulting inference problem. Our approach outperforms all previously published algorithms for intrinsic image decomposition and shape-from-shading on the MIT intrinsic images dataset [1, 2] and on our own ''naturally'' illuminated version of that dataset.

[S4-P19B]

A Fast Illumination and Deformation Insensitive Image Comparison Algorithm Using Wavelet-Based Geodesics

Anne Jorstad, David Jacobs, and Alain Trouvé

We present a fast image comparison algorithm for handling variations in illumination and moderate amounts of deformation using an efficient geodesic framework. As the geodesic is the shortest path between two images on a manifold, it is a natural choice to use the length of the geodesic to determine the image similarity. Distances on the manifold are defined by a metric that is insensitive to changes in scene lighting. This metric is described in the wavelet domain where it is able to handle moderate amounts of deformation, and can be calculated extremely fast (less than 3ms per image comparison). We demonstrate the similarity between our method and the illumination insensitivity achieved by the Gradient Direction. Strong results are presented on the AR Face Database.

[S4-P20B]

Large-Scale Gaussian Process Classification with Flexible Adaptive Histogram Kernels

Erik Rodner, Alexander Freytag, Paul Bodesheim, and Joachim Denzler

We present how to perform exact large-scale multi-class Gaussian process classification with parameterized histogram intersection kernels. In contrast to previous approaches, we use a full Bayesian model without any sparse approximation techniques, which allows for learning in sub-quadratic and classification in constant time. To handle the additional model flexibility induced by parameterized kernels, our approach is able to optimize the parameters with large-scale training data. A key ingredient of this optimization is a new efficient upper bound of the negative Gaussian process log-likelihood. Experiments with image categorization tasks exhibit high performance gains with flexible kernels as well as learning within a few minutes and classification in microseconds for databases, where exact Gaussian process inference was not possible before.

[S4-P21B]

Background Subtraction with Dirichlet Processes

Tom S.F. Haines and Tao Xiang

Background subtraction is an important first step for video analysis, where it is used to discover the objects of interest for further processing. Such an algorithm often consists of a background model and a regularisation scheme. The background model determines a per-pixel measure of if a pixel belongs to the background or the foreground, whilst the regularisation brings in information from adjacent pixels. A new method is presented that uses a Dirichlet process Gaussian mixture model to estimate a per-pixel background distribution, which is followed by probabilistic regularisation. Key advantages include inferring the per-pixel mode count, such that it accurately models dynamic backgrounds, and that it updates its model continuously in a principled way.

[S4-P22B]

Mobile Product Image Search by Automatic Query Object Extraction

Xiaohui Shen, Zhe Lin, Jonathan Brandt, and Ying Wu

Mobile product image search aims at identifying a product, or retrieving similar products from a database based on a photo captured from a mobile phone camera. Application of traditional image retrieval methods (e.g. bag-of-words) to mobile visual search has been shown to be effective in identifying duplicate/near-duplicate photos, near-planar and textured objects such as landmarks, books/cd covers. However, retrieving more general product categories is still a challenging research problem due to variations in viewpoint, illumination, scale, the existence of blur and background clutter in the query image, etc. In this paper, we propose a new approach that can simultaneously extract the product instance from the query, identify the instance, and retrieve visually similar product images. Based on the observation that good query segmentation helps improve retrieval accuracy and good search results provide good priors for segmentation, we formulate our approach in an iterative scheme to improve both query segmentation and retrieval accuracy. To this end, a weighted object mask voting algorithm is proposed based on a spatially-constrained model, which allows robust localization and segmentation of the query object, and achieves significantly better retrieval accuracy than previous methods. We show the effectiveness of our approach by applying it to a large, real-world product image dataset and a new object category dataset.

[S4-P23B]

Analyzing the Subspace Structure of Related Images: Concurrent Segmentation of Image Sets

Lopamudra Mukherjee, Vikas Singh, Jia Xu, and Maxwell D. Collins

We develop new algorithms to analyze and exploit the joint subspace structure of a set of related images to facilitate the process of concurrent segmentation of a large set of images. Most existing approaches for this problem are either limited to extracting a single similar object across the given image set or do not scale well to a large number of images containing multiple objects varying at different scales. One of the goals of this paper is to show that various desirable properties of such an algorithm (ability to handle multiple images with multiple objects showing arbitary scale variations) can be cast elegantly using simple constructs from linear algebra: this significantly extends the operating range of such methods. While intuitive, this formulation leads to a hard optimization problem where one must perform the image segmentation task together with appropriate constraints which enforce desired algebraic regularity (e.g., common subspace structure). We propose efficient iterative algorithms (with small computational requirements) whose key steps reduce to objective functions solvable by max-flow and/or nearly closed form identities. We study the qualitative, theoretical, and empirical properties of the method, and present results on benchmark datasets.

[S4-P24B]

Artistic Image Classification: An Analysis on the PRINTART Database

Gustavo Carneiro, Nuno Pinho da Silva, Alessio Del Bue, and João Paulo Costeira

Artistic image understanding is an interdisciplinary research field of increasing importance for the computer vision and the art history communities. For computer vision scientists, this problem offers challenges where new techniques can be developed; and for the art history community new automatic art analysis tools can be developed. On the positive side, artistic images are generally constrained by compositional rules and artistic themes. However, the low-level texture and color features exploited for photographic image analysis are not as effective because of inconsistent color and texture patterns describing the visual classes in artistic images. In this work, we present a new database of monochromatic artistic images containing 988 images with a global semantic annotation, a local compositional annotation, and a pose annotation of human subjects and animal types. In total, 75 visual classes are annotated, from which 27 are related to the theme of the art image, and 48 are visual classes that can be localized in the image with bounding boxes. Out of these 48 classes, 40 have pose annotation, with 37 denoting human subjects and 3 representing animal types. We also provide a complete evaluation of several algorithms recently proposed for image annotation and retrieval. We then present an algorithm achieving remarkable performance over the most successful algorithm hitherto proposed for this problem. Our main goal with this paper is to make this database, the evaluation process, and the benchmark results available for the computer vision community.

ORAL SESSION 4
ACTIONS AND ACTIVITIES

Tuesday, October 9
17:05 - 18:30

[S4-O1]

Detecting Actions, Poses, and Objects with Relational Phraselets

Chaitanya Desai and Deva Ramanan

We present a novel approach to modeling human pose, together with interacting objects, based on compositional models of local visual interactions and their relations. Skeleton models, while flexible enough to capture large articulations, fail to accurately model self-occlusions and interactions. Poselets and Visual Phrases address this limitation, but do so at the expense of requiring a large set of templates. We combine all three approaches with a compositional model that is flexible enough to model detailed articulations but still captures occlusions and object interactions. Unlike much previous work on action classification, we do not assume test images are labeled with a person, and instead present results for ''action detection'' in an unlabeled image. Notably, for each detection, our model reports back a detailed description including an action label, articulated human pose, object poses, and occlusion flags. We demonstrate that modeling occlusion is crucial for recognizing human-object interactions. We present results on the PASCAL Action Classification challenge that shows our unified model advances the state-of-the-art for detection, action classification, and articulated pose estimation.

[S4-O2]

Action Recognition with Exemplar Based 2.5D Graph Matching

Bangpeng Yao and Li Fei-Fei

This paper deals with recognizing human actions in still images. We make two key contributions. (1) We propose a novel, 2.5D representation of action images that considers both view-independent pose information and rich appearance information. A 2.5D graph of an action image consists of a set of nodes that are key-points of the human body, as well as a set of edges that are spatial relationships between the nodes. Each key-point is represented by view-independent 3D positions and local 2D appearance features. The similarity between two action images can then be measured by matching their corresponding 2.5D graphs. (2) We use an exemplar based action classification approach, where a set of representative images are selected for each action class. The selected images cover large within-action variations and carry discriminative information compared with the other classes. This exemplar based representation of action classes further makes our approach robust to pose variations and occlusions. We test our method on two publicly available datasets and show that it achieves very promising performance.

[S4-O3]

Cost-Sensitive Top-Down/Bottom-Up Inference for Multiscale Activity Recognition

Mohamed R. Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song-Chun Zhu

This paper addresses a new problem, that of multiscale activity recognition. Our goal is to detect and localize a wide range of activities, including individual actions and group activities, which may simultaneously co-occur in high-resolution video. The video resolution allows for digital zoom-in (or zoom-out) for examining fine details (or coarser scales), as needed for recognition. The key challenge is how to avoid running a multitude of detectors at all spatiotemporal scales, and yet arrive at a holistically consistent video interpretation. To this end, we use a three-layered AND-OR graph to jointly model group activities, individual actions, and participating objects. The AND-OR graph allows a principled formulation of efficient, cost-sensitive inference via an explore-exploit strategy. Our inference optimally schedules the following computational processes: 1) direct application of activity detectors – called α process; 2) bottom-up inference based on detecting activity parts – called β process; and 3) top-down inference based on detecting activity context – called γ process. The scheduling iteratively maximizes the log-posteriors of the resulting parse graphs. For evaluation, we have compiled and benchmarked a new dataset of high-resolution videos of group and individual activities co-occurring in a courtyard of the UCLA campus.

[S4-O4]

Activity Forecasting

Kris M. Kitani, Brian D. Ziebart, James Andrew Bagnell, and Martial Hebert

We address the task of inferring the future actions of people from noisy visual input. We denote this task activity forecasting. To achieve accurate activity forecasting, our approach models the effect of the physical environment on the choice of human actions. This is accomplished by the use of state-of-the-art semantic scene understanding combined with ideas from optimal control theory. Our unified model also integrates several other key elements of activity analysis, namely, destination forecasting, sequence smoothing and transfer learning. As proof-of-concept, we focus on the domain of trajectory-based activity analysis from visual input. Experimental results demonstrate that our model accurately predicts distributions over future actions of individuals. We show how the same techniques can improve the results of tracking algorithms by leveraging information about likely goals and trajectories.

[S4-O5]

A Unified Framework for Multi-target Tracking and Collective Activity Recognition

Wongun Choi and Silvio Savarese

We present a coherent, discriminative framework for simultaneously tracking multiple people and estimating their collective activities. Instead of treating the two problems separately, our model is grounded in the intuition that a strong correlation exists between a person's motion, their activity, and the motion and activities of other nearby people. Instead of directly linking the solutions to these two problems, we introduce a hierarchy of activity types that creates a natural progression that leads from a specific person's motion to the activity of the group as a whole. Our model is capable of jointly tracking multiple people, recognizing individual activities (atomic activities), the interactions between pairs of people (interaction activities), and finally the behavior of groups of people (collective activities). We also propose an algorithm for solving this otherwise intractable joint inference problem by combining belief propagation with a version of the branch and bound algorithm equipped with integer programming. Experimental results on challenging video datasets demonstrate our theoretical claims and indicate that our model achieves the best collective activity classification results to date.

POSTER SESSION 5

Wednesday, October 10
08:45 - 11:15

[S5-P1A]

Camera Pose Estimation Using First-Order Curve Differential Geometry

Ricardo Fabbri, Benjamin B. Kimia, and Peter J. Giblin

This paper considers and solves the problem of estimating camera pose given a pair of point-tangent correspondences between the 3D scene and the projected image. The problem arises when considering curve geometry as the basis of forming correspondences, computation of structure and calibration, which in its simplest form is a point augmented with the curve tangent. We show that while the standard resectioning problem is solved with a minimum of three points given the intrinsic parameters, when points are augmented with tangent information only two points are required, leading to substantial computational savings, e.g., when used as a minimal engine within ransac. In addition, computational algorithms are developed to find a practical and efficient solution shown to effectively recover camera pose using both synthetic and realistic datasets. The resolution of this problem is intended as a basic building block of future curve-based structure from motion systems, allowing new views to be incrementally registered to a core set of views for which relative pose has already been computed.

[S5-P2A]

Beyond Feature Points: Structured Prediction for Monocular Non-rigid 3D Reconstruction

Mathieu Salzmann and Raquel Urtasun

Existing approaches to non-rigid 3D reconstruction either are specifically designed for feature point correspondences, or require a good shape initialization to exploit more complex image likelihoods. In this paper, we formulate reconstruction as inference in a graphical model, where the variables encode the rotations and translations of the facets of a surface mesh. This lets us exploit complex likelihoods even in the absence of a good initialization. In contrast to existing approaches that set the weights of the likelihood terms manually, our formulation allows us to learn them from as few as a single training example. To improve efficiency, we combine our structured prediction formalism with a gradient-based scheme. Our experiments show that our approach yields tremendous improvement over state-of-the-art gradient-based methods.

[S5-P3A]

Learning Spatially-Smooth Mappings in Non-Rigid Structure From Motion

Onur C. Hamsici, Paulo F.U. Gotardo, and Aleix M. Martinez

Non-rigid structure from motion (NRSFM) is a classical underconstrained problem in computer vision. A common approach to make NRSFM more tractable is to constrain 3D shape deformation to be smooth over time. This constraint has been used to compress the deformation model and reduce the number of unknowns that are estimated. However, temporal smoothness cannot be enforced when the data lacks temporal ordering and its benefits are less evident when objects undergo abrupt deformations. This paper proposes a new NRSFM method that addresses these problems by considering deformations as spatial variations in shape space and then enforcing spatial, rather than temporal, smoothness. This is done by modeling each 3D shape coefficient as a function of its input 2D shape. This mapping is learned in the feature space of a rotation invariant kernel, where spatial smoothness is intrinsically defined by the mapping function. As a result, our model represents shape variations compactly using custom-built coefficient bases learned from the input data, rather than a pre-specified set such as the Discrete Cosine Transform. The resulting kernel-based mapping is a by-product of the NRSFM solution and leads to another fundamental advantage of our approach: for a newly observed 2D shape, its 3D shape is recovered by simply evaluating the learned function.

[S5-P4A]

In Defence of RANSAC for Outlier Rejection in Deformable Registration

Quoc-Huy Tran, Tat-Jun Chin, Gustavo Carneiro, Michael S. Brown, and David Suter

This paper concerns the robust estimation of non-rigid deformations from feature correspondences. We advance the surprising view that for many realistic physical deformations, the error of the mismatches (outliers) usually dwarfs the effects of the curvature of the manifold on which the correct matches (inliers) lie, to the extent that one can tightly enclose the manifold within the error bounds of a low-dimensional hyperplane for accurate outlier rejection. This justifies a simple RANSAC-driven deformable registration technique that is at least as accurate as other methods based on the optimisation of fully deformable models. We support our ideas with comprehensive experiments on synthetic and real data typical of the deformations examined in the literature.

[S5-P5A]

A Tensor Voting Approach for Multi-view 3D Scene Flow Estimation and Refinement

Jaesik Park, Tae Hyun Oh, Jiyoung Jung, Yu-Wing Tai, and In So Kweon

We introduce a framework to estimate and refine 3D scene flow which connects 3D structures of a scene across different frames. In contrast to previous approaches which compute 3D scene flow that connects depth maps from a stereo image sequence or from a depth camera, our approach takes advantage of full 3D reconstruction which computes the 3D scene flow that connects 3D point clouds from multi-view stereo system. Our approach uses a standard multi-view stereo and optical flow algorithm to compute the initial 3D scene flow. A unique two-stage refinement process regularizes the scene flow direction and magnitude sequentially. The scene flow direction is refined by utilizing 3D neighbor smoothness defined by tensor voting. The magnitude of the scene flow is refined by connecting the implicit surfaces across the consecutive 3D point clouds. Our estimated scene flow is temporally consistent. Our approach is efficient, model free, and it is effective in error corrections and outlier rejections. We tested our approach on both synthetic and real-world datasets. Our experimental results show that our approach out-performs previous algorithms quantitatively on synthetic dataset, and it improves the reconstructed 3D model from the refined 3D point cloud in real-world dataset.

[S5-P6A]

Two-View Underwater Structure and Motion for Cameras under Flat Refractive Interfaces

Lai Kang, Lingda Wu, and Yee-Hong Yang

In an underwater imaging system, a refractive interface is introduced when a camera looks into the water-based environment, resulting in distorted images due to refraction. Simply ignoring the refraction effect or using the lens radial distortion model causes erroneous 3D reconstruction. This paper deals with a general underwater imaging setup using two cameras, of which each camera is placed in a separate waterproof housing with a flat window. The impact of refraction is explicitly modeled in the refractive camera model. Based on two new concepts, namely the Ellipse of Refrax (EoR) and Refractive Depth (RD) of a scene point, we show that provably optimal underwater structure and motion under L_∞-norm can be estimated given known rotation. The constraint of known rotation is further relaxed by incorporating two-view geometry estimation into a new hybrid optimization framework. The experimental results using both synthetic and real images demonstrate that the proposed method can significantly improve the accuracy of camera motion and 3D structure estimation for underwater applications.

[S5-P7A]

Reading Ancient Coins: Automatically Identifying Denarii Using Obverse Legend Seeded Retrieval

Ognjen Arandjelovic

The aim of this paper is to automatically identify a Roman Imperial denarius from a single query photograph of its obverse and reverse. Such functionality has the potential to contribute greatly to various national schemes which encourage laymen to report their finds to local museums. Our work introduces a series of novelties: (i) this is the first paper which describes a method for extracting the legend of an ancient coin from a photograph; (ii) we are also the first to suggest the idea and propose a method for identifying a coin using a series of carefully engineered retrievals, each harnessed for further information using visual or meta-data processing; (iii) we show how in addition to a unique standard reference number for a query coin, the proposed system can be used to extract salient coin information (issuing authority, obverse and reverse descriptions, mint date) and retrieve images of other coins of the same type.

[S5-P8A]

Robust and Practical Face Recognition via Structured Sparsity

Kui Jia, Tsung-Han Chan, and Yi Ma

Sparse representation based classification (SRC) methods have recently drawn much attention in face recognition, due to their good performance and robustness against misalignment, illumination variation, and occlusion. They assume the errors caused by image variations can be modeled as pixel-wisely sparse. However, in many practical scenarios these errors are not truly pixel-wisely sparse but rather sparsely distributed with structures, i.e., they constitute contiguous regions distributed at different face positions. In this paper, we introduce a class of structured sparsity-inducing norms into the SRC framework, to model various corruptions in face images caused by misalignment, shadow (due to illumination change), and occlusion. For practical face recognition, we develop an automatic face alignment method based on minimizing the structured sparsity norm. Experiments on benchmark face datasets show improved performance over SRC and other alternative methods.

[S5-P9A]

Recognizing Materials from Virtual Examples

Wenbin Li and Mario Fritz

Due to the strong impact of machine learning methods on visual recognition, performance on many perception task is driven by the availability of sufficient training data. A promising direction which has gained new relevance in recent years is the generation of virtual training examples by means of computer graphics methods in order to provide richer training sets for recognition and detection on real data. Success stories of this paradigm have been mostly reported for the synthesis of shape features and 3D depth maps. Therefore we investigate in this paper if and how appearance descriptors can be transferred from the virtual world to real examples. We study two popular appearance descriptors on the task of material categorization as it is a pure appearance-driven task. Beyond this initial study, we also investigate different approach of combining and adapting virtual and real data in order to bridge the gap between rendered and real-data. Our study is carried out using a new database of virtual materials VIPS that complements the existing KTH-TIPS material database.

[S5-P10A]

Scene Recognition on the Semantic Manifold

Roland Kwitt, Nuno Vasconcelos, and Nikhil Rasiwasia

A new architecture, denoted spatial pyramid matching on the semantic manifold (SPMSM), is proposed for scene recognition. SPMSM is based on a recent image representation on a semantic probability simplex, which is now augmented with a rough encoding of spatial information. A connection between the semantic simplex and a Riemmanian manifold is established, so as to equip the architecture with a similarity measure that respects the manifold structure of the semantic space. It is then argued that the closed-form geodesic distance between two manifold points is a natural measure of similarity between images. This leads to a conditionally positive definite kernel that can be used with any SVM classifier. An approximation of the geodesic distance reveals connections to the well-known Bhattacharyya kernel, and is explored to derive an explicit feature embedding for this kernel, by simple square-rooting. This enables a low-complexity SVM implementation, using a linear SVM on the embedded features. Several experiments are reported, comparing SPMSM to state-of-the-art recognition methods. SPMSM is shown to achieve the best recognition rates in the literature for two large datasets (MIT Indoor and SUN) and rates equivalent or superior to the state-of-the-art on a number of smaller datasets. In all cases, the resulting SVM also has much smaller dimensionality and requires much fewer support vectors than previous classifiers. This guarantees much smaller complexity and suggests improved generalization beyond the datasets considered.

[S5-P11A]

Unsupervised Temporal Commonality Discovery

Wen-Sheng Chu, Feng Zhou, and Fernando De la Torre

Unsupervised discovery of commonalities in images has recently attracted much interest due to the need to find correspondences in large amounts of visual data. A natural extension, and a relatively unexplored problem, is how to discover common semantic temporal patterns in videos. That is, given two or more videos, find the subsequences that contain similar visual content in an unsupervised manner. We call this problem Temporal Commonality Discovery (TCD). The naive exhaustive search approach to solve the TCD problem has a computational complexity quadratic with the length of each sequence, making it impractical for regular-length sequences. This paper proposes an efficient branch and bound (B&B) algorithm to tackle the TCD problem. We derive tight bounds for classical distances between temporal bag of words of two segments, including l_1, intersection and X^2. Using these bounds the B&B algorithm can efficiently find the global optimal solution. Our algorithm is general, and it can be applied to any feature that has been quantified into histograms. Experiments on finding common facial actions in video and human actions in motion capture data demonstrate the benefits of our approach. To the best of our knowledge, this is the first work that addresses unsupervised discovery of common events in videos.

[S5-P12A]

Finding People Using Scale, Rotation and Articulation Invariant Matching

Hao Jiang

A scale, rotation and articulation invariant method is proposed to match human subjects in images. Different from the widely used pictorial structure scheme, the proposed method directly matches body parts to image regions which are obtained from object independent proposals and successively merged superpixels. Body part region matching is formulated as a graph matching problem. We globally assign a body part candidate to each node on the model graph so that the overall configuration satisfies the spatial layout of a human body plan, part regions have small overlap, and the part coverage follows proper area ratios. The proposed graph model is non-tree and contains high order hyper-edges. We propose an efficient method that finds global optimal solution to the matching problem with a sequence of branch and bound procedures. The experiments show that the proposed method is able to handle arbitrary scale, rotation, articulation and match human subjects in cluttered images.

[S5-P13A]

Measuring Image Distances via Embedding in a Semantic Manifold

Chen Fang and Lorenzo Torresani

In this work we introduce novel image metrics that can be used with distance-based classifiers or directly to decide whether two input images belong to the same class. While most prior image distances rely purely on comparisons of low-level features extracted from the inputs, our metrics use a large database of labeled photos as auxiliary data to draw semantic relationships between the two images, beyond those computable from simple visual features. In a preprocessing stage our approach derives a semantic image graph from the labeled dataset, where the nodes are the labeled images and the edges connect pictures with related labels. The graph can be viewed as modeling a semantic image manifold, and it enables the use of graph distances to approximate semantic distances. Thus, we reformulate the task of measuring the semantic distance between two unlabeled pictures as the problem of embedding the two input images in the semantic graph. We propose and evaluate several embedding schemes and graph distance metrics. Our results on Caltech101, Caltech256 and ImageNet show that our distances consistently match or outperform the state-of-the-art in this field.

[S5-P14A]

Efficient Point-to-Subspace Query in l^1 with Application to Robust Face Recognition

Ju Sun, Yuqian Zhang, and John Wright

Motivated by vision tasks such as robust face and object recognition, we consider the following general problem: given a collection of low-dimensional linear subspaces in a high-dimensional ambient (image) space, and a query point (image), efficiently determine the nearest subspace to the query in l^1 distance. We show in theory this problem can be solved with a simple two-stage algorithm: (1) random Cauchy projection of query and subspaces into low-dimensional spaces followed by efficient distance evaluation (l^1 regression); (2) getting back to the high-dimensional space with very few candidates and performing exhaustive search. We present preliminary experiments on robust face recognition to corroborate our theory.

[S5-P15A]

Recognizing Complex Events Using Large Margin Joint Low-Level Event Model

Hamid Izadinia and Mubarak Shah

In this paper we address the challenging problem of complex event recognition by using low-level events. In this problem, each complex event is captured by a long video in which several low-level events happen. The dataset contains several videos and due to the large number of videos and complexity of the events, the available annotation for the low-level events is very noisy which makes the detection task even more challenging. To tackle these problems we model the joint relationship between the low-level events in a graph where we consider a node for each low-level event and whenever there is a correlation between two low-level events the graph has an edge between the corresponding nodes. In addition, for decreasing the effect of weak and/or irrelevant low-level event detectors we consider the presence/absence of low-level events as hidden variables and learn a discriminative model by using latent SVM formulation. Using our learned model for the complex event recognition, we can also apply it for improving the detection of the low-level events in video clips which enables us to discover a conceptual description of the video. Thus our model can do complex event recognition and explain a video in terms of low-level events in a single framework. We have evaluated our proposed method over the most challenging multimedia event detection dataset. The experimental results reveals that the proposed method performs well compared to the baseline method. Further, our results of conceptual description of video shows that our model is learned quite well to handle the noisy annotation and surpass the low-level event detectors which are directly trained on the raw features.

[S5-P16A]

Multi-component Models for Object Detection

Chunhui Gu, Pablo Arbeláez, Yuanqing Lin, Kai Yu, and Jitendra Malik

In this paper, we propose a multi-component approach for object detection. Rather than attempting to represent an object category with a monolithic model, or pre-defining a reduced set of aspects, we form visual clusters from the data that are tight in appearance and configuration spaces. We train individual classifiers for each component, and then learn a second classifier that operates at the category level by aggregating responses from multiple components. In order to reduce computation cost during detection, we adopt the idea of object window selection, and our segmentation-based selection mechanism produces fewer than 500 windows per image while preserving high object recall. When compared to the leading methods in the challenging VOC PASCAL 2010 dataset, our multi-component approach obtains highly competitive results. Furthermore, unlike monolithic detection methods, our approach allows the transfer of finer-grained semantic information from the components, such as keypoint location and segmentation masks.

[S5-P17A]

Discriminative Decorrelation for Clustering and Classification

Bharath Hariharan, Jitendra Malik, and Deva Ramanan

Object detection has over the past few years converged on using linear SVMs over HOG features. Training linear SVMs however is quite expensive, and can become intractable as the number of categories increase. In this work we revisit a much older technique, viz. Linear Discriminant Analysis, and show that LDA models can be trained almost trivially, and with little or no loss in performance. The covariance matrices we estimate capture properties of natural images. Whitening HOG features with these covariances thus removes naturally occuring correlations between the HOG features. We show that these whitened features (which we call WHO) are considerably better than the original HOG features for computing similarities, and prove their usefulness in clustering. Finally, we use our findings to produce an object detection system that is competitive on PASCAL VOC 2007 while being considerably easier to train and test.

[S5-P18A]

Beyond Spatial Pyramids: A New Feature Extraction Framework with Dense Spatial Sampling for Image Classification

Shengye Yan, Xinxing Xu, Dong Xu, Stephen Lin, and Xuelong Li

We introduce a new framework for image classification that extends beyond the window sampling of fixed spatial pyramids to include a comprehensive set of windows densely sampled over location, size and aspect ratio. To effectively deal with this large set of windows, we derive a concise high-level image feature using a two-level extraction method. At the first level, window-based features are computed from local descriptors (e.g., SIFT, spatial HOG, LBP) in a process similar to standard feature extractors. Then at the second level, the new image feature is determined from the window-based features in a manner analogous to the first level. This higher level of abstraction offers both efficient handling of dense samples and reduced sensitivity to misalignment. More importantly, our simple yet effective framework can readily accommodate a large number of existing pooling/coding methods, allowing them to extract features beyond the spatial pyramid representation. To effectively fuse the second level feature with a standard first level image feature for classification, we additionally propose a new learning algorithm, called Generalized Adaptive lp-norm Multiple Kernel Learning (GA-MKL), to learn an adapted robust classifier based on multiple base kernels constructed from image features and multiple sets of pre-learned classifiers of all the classes. Extensive evaluation on the object recognition (Caltech256) and scene recognition (15Scenes) benchmark datasets demonstrates that the proposed method outperforms state-of-the-art image classification algorithms under a broad range of settings.

[S5-P19A]

Subspace Learning in Krein Spaces: Complete Kernel Fisher Discriminant Analysis with Indefinite Kernels

Stefanos Zafeiriou

Positive definite kernels, such as Gaussian Radial Basis Functions (GRBF), have been widely used in computer vision for designing feature extraction and classification algorithms. In many cases non-positive definite (npd) kernels and non metric similarity/dissimilarity measures naturally arise (e.g., Hausdorff distance, Kullback Leibler Divergences and Compact Support (CS) Kernels). Hence, there is a practical and theoretical need to properly handle npd kernels within feature extraction and classification frameworks. Recently, classifiers such as Support Vector Machines (SVMs) with npd kernels, Indefinite Kernel Fisher Discriminant Analysis (IKFDA) and Indefinite Kernel Quadratic Analysis (IKQA) were proposed. In this paper we propose feature extraction methods using indefinite kernels. In particular, first we propose an Indefinite Kernel Principal Component Analysis (IKPCA). Then, we properly define optimization problems that find discriminant projections with indefinite kernels and propose a Complete Indefinite Kernel Fisher Discriminant Analysis (CIKFDA) that solves the proposed problems. We show the power of the proposed frameworks in a fully automatic face recognition scenario.

[S5-P20A]

A Novel Material-Aware Feature Descriptor for Volumetric Image Registration in Diffusion Tensor Space

Shuai Li, Qinping Zhao, Shengfa Wang, Tingbo Hou, Aimin Hao, and Hong Qin

This paper advocates a novel material-aware feature descriptor for volumetric image registration. We rigorously formulate a novel probability density function (PDF) based distance metric to devise a compact local feature descriptor supporting invariance of full 3D orientation and isometric deformation. The central idea is to employ anisotropic heat diffusion to characterize the detected local volumetric features. It is achieved by the elegant unification of diffusion tensor (DT) space construction based on local Hessian eigen-system, multi-scale feature extraction based on DT-weighted dyadic wavelet transform, and local distance definition based on PDF formulated in DT space. The diffusion, intrinsic structure-aware nature makes our volumetric feature descriptor more robust to noise. With volumetric images registration as verifiable application, various experiments on different volumetric images demonstrate the superiority of our descriptor.

[S5-P21A]

Efficient Closed-Form Solution to Generalized Boundary Detection

Marius Leordeanu, Rahul Sukthankar, and Cristian Sminchisescu

Boundary detection is essential for a variety of computer vision tasks such as segmentation and recognition. We propose a unified formulation for boundary detection, with closed-form solution, which is applicable to the localization of different types of boundaries, such as intensity edges and occlusion boundaries from video and RGB-D cameras. Our algorithm simultaneously combines low- and mid-level image representations, in a single eigenvalue problem, and we solve over an infinite set of putative boundary orientations. Moreover, our method achieves state of the art results at a significantly lower computational cost than current methods. We also propose a novel method for soft-segmentation that can be used in conjunction with our boundary detection algorithm and improve its accuracy at a negligible extra computational cost.

[S5-P22A]

Attribute Learning for Understanding Unstructured Social Activity

Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong

The rapid development of social video sharing platforms has created a huge demand for automatic video classification and annotation techniques, in particular for videos containing social activities of a group of people (e.g. YouTube video of a wedding reception). Recently, attribute learning has emerged as a promising paradigm for transferring learning to sparsely labelled classes in object or single-object short action classification. In contrast to existing work, this paper for the first time, tackles the problem of attribute learning for understanding group social activities with sparse labels. This problem is more challenging because of the complex multi-object nature of social activities, and the unstructured nature of the activity context. To solve this problem, we (1) contribute an unstructured social activity attribute (USAA) dataset with both visual and audio attributes, (2) introduce the concept of semi-latent attribute space and (3) propose a novel model for learning the latent attributes which alleviate the dependence of existing models on exact and exhaustive manual specification of the attribute-space. We show that our framework is able to exploit latent attributes to outperform contemporary approaches for addressing a variety of realistic multi-media sparse data learning tasks including: multi-task learning, N-shot transfer learning, learning with label noise and importantly zero-shot learning.

[S5-P1B]

Statistical Inference of Motion in the Invisible

Haroon Idrees, Imran Saleemi, and Mubarak Shah

This paper focuses on the unexplored problem of inferring motion of objects that are invisible to all cameras in a multiple camera setup. As opposed to methods for learning relationships between disjoint cameras, we take the next step to actually infer the exact spatiotemporal behavior of objects while they are invisible. Given object trajectories within disjoint cameras' FOVs (field-of-view), we introduce constraints on the behavior of objects as they travel through the unobservable areas that lie in between. These constraints include vehicle following (the trajectories of vehicles adjacent to each other at entry and exit are time-shifted relative to each other), collision avoidance (no two trajectories pass through the same location at the same time) and temporal smoothness (restricts the allowable movements of vehicles based on physical limits). The constraints are embedded in a generalized, global cost function for the entire scene, incorporating influences of all objects, followed by a bounded minimization using an interior point algorithm, to obtain trajectory representations of objects that define their exact dynamics and behavior while invisible. Finally, a statistical representation of motion in the entire scene is estimated to obtain a probabilistic distribution representing individual behaviors, such as turns, constant velocity motion, deceleration to a stop, and acceleration from rest for evaluation and visualization. Experiments are reported on real world videos from multiple disjoint cameras in NGSIM data set, and qualitative as well as quantitative analysis confirms the validity of our approach.

[S5-P2B]

Going with the Flow: Pedestrian Efficiency in Crowded Scenes

Louis Kratz and Ko Nishino

Video analysis of crowded scenes is challenging due to the complex motion of individual people in the scene. The collective motion of pedestrians form a crowd flow, but individuals often largely deviate from it as they anticipate and react to each other. Deviations from the crowd decreases the pedestrian's efficiency: a sociological concept that measures the difference of actual motion from the intended speed and direction. In this paper, we derive a novel method for estimating pedestrian efficiency from videos. We first introduce a novel crowd motion model that encodes the temporal evolution of local motion patterns represented with directional statistics distributions. This model is then used to estimate the intended motion of pedestrians at every space-time location, which enables visual measurement of the pedestrian efficiency. We demonstrate the use of this pedestrian efficiency to detect unusual events and to track individuals in crowded scenes. Experimental results show that the use of pedestrian efficiency leads to state-of-the-art accuracy in these critical applications.

[S5-P3B]

Reconstructing 3D Human Pose from 2D Image Landmarks

Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh

Reconstructing an arbitrary configuration of 3D points from their projection in an image is an ill-posed problem. When the points hold semantic meaning, such as anatomical landmarks on a body, human observers can often infer a plausible 3D configuration, drawing on extensive visual memory. We present an activity-independent method to recover the 3D configuration of a human figure from 2D locations of anatomical landmarks in a single image, leveraging a large motion capture corpus as a proxy for visual memory. Our method solves for anthropometrically regular body pose and explicitly estimates the camera via a matching pursuit algorithm operating on the image projections. Anthropometric regularity (i.e., that limbs obey known proportions) is a highly informative prior, but directly applying such constraints is intractable. Instead, we enforce a necessary condition on the sum of squared limb-lengths that can be solved for in closed form to discourage implausible configurations in 3D. We evaluate performance on a wide variety of human poses captured from different viewpoints and show generalization to novel 3D configurations and robustness to missing data.

[S5-P4B]

Fast Tiered Labeling with Topological Priors

Ying Zheng, Steve Gu, and Carlo Tomasi

We consider labeling an image with multiple tiers. Tiers, one on top of another, enforce a strict vertical order among objects (e.g. sky is above the ground). Two new ideas are explored: First, under a simplification of the general tiered labeling framework proposed by Felzenszwalb and Veksler [1], we design an efficient O(KN) algorithm for the approximate optimal labeling of an image of N pixels with K tiers. Our algorithm runs in over 100 frames per second on images of VGA resolutions when K is less than 6. When K = 3, our solution overlaps with the globally optimal one by Felzenszwalb and Veksler in over 99% of all pixels but runs 1000 times faster. Second, we define a topological prior that specifies the number of local extrema in the tier boundaries, and give an O(NM) algorithm to find a single, optimal tier boundary with exactly M local maxima and minima. These two extensions enrich the general tiered labeling framework and enable fast computation. The proposed topological prior further improves the accuracy in labeling details.

[S5-P5B]

TreeCANN - k-d Tree Coherence Approximate Nearest Neighbor Algorithm

Igor Olonetsky and Shai Avidan

TreeCANN is a fast algorithm for approximately matching all patches between two images. It does so by following the established convention of finding an initial set of matching patch candidates between the two images and then propagating good matches to neighboring patches in the image plane. TreeCANN accelerates each of these components substantially leading to an algorithm that is ×3 to ×5 faster than existing methods. Seed matching is achieved using a properly tuned k-d tree on a sparse grid of patches. In particular, we show that a sequence of key design decisions can make k-d trees run as fast as recently proposed state-of-the-art methods, and because of image coherency it is enough to consider only a sparse grid of patches across the image plane. We then develop a novel propagation step that is based on the integral image, which drastically reduces the computational load that is dominated by the need to repeatedly measure similarity between pairs of patches. As a by-product we give an optimal algorithm for exact matching that is based on the integral image. The proposed exact algorithm is faster than previously reported results and depends only on the size of the images and not on the size of the patches. We report results on large and varied data sets and show that TreeCANN is orders of magnitude faster than exact NN search yet produces matches that are within 1% error, compared to the exact NN search.

[S5-P6B]

Robust Regression

Dong Huang, Ricardo Silveira Cabral, and Fernando De la Torre

Discriminative methods (e.g., kernel regression, SVM) have been extensively used to solve problems such as object recognition, image alignment and pose estimation from images. Regression methods typically map image features (X) to continuous (e.g., pose) or discrete (e.g., object category) values. A major drawback of existing regression methods is that samples are directly projected onto a subspace and hence fail to account for outliers which are common in realistic training sets due to occlusion, specular reflections or noise. It is important to notice that in existing regression methods, and discriminative methods in general, the regressor variables X are assumed to be noise free. Due to this assumption, discriminative methods experience significant degrades in performance when gross outliers are present. Despite its obvious importance, the problem of robust discriminative learning has been relatively unexplored in computer vision. This paper develops the theory of Robust Regression (RR) and presents an effective convex approach that uses recent advances on rank minimization. The framework applies to a variety of problems in computer vision including robust linear discriminant analysis, multi-label classification and head pose estimation from images. Several synthetic and real world examples are used to illustrate the benefits of RR.

[S5-P7B]

Domain Adaptive Dictionary Learning

Qiang Qiu, Vishal M. Patel, Pavan Turaga, and Rama Chellappa

Many recent efforts have shown the effectiveness of dictionary learning methods in solving several computer vision problems. However, when designing dictionaries, training and testing domains may be different, due to different view points and illumination conditions. In this paper, we present a function learning framework for the task of transforming a dictionary learned from one visual domain to the other, while maintaining a domain-invariant sparse representation of a signal. Domain dictionaries are modeled by a linear or non-linear parametric function. The dictionary function parameters and domain-invariant sparse codes are then jointly learned by solving an optimization problem. Experiments on real datasets demonstrate the effectiveness of our approach for applications such as face recognition, pose alignment and pose estimation.

[S5-P8B]

A Robust and Efficient Doubly Regularized Metric Learning Approach

Meizhu Liu and Baba C. Vemuri

A proper distance metric is fundamental in many computer vision and pattern recognition applications such as classification, image retrieval, face recognition and so on. However, it is usually not clear what metric is appropriate for specific applications, therefore it becomes more reliable to learn a task oriented metric. Over the years, many metric learning approaches have been reported in literature. A typical one is to learn a Mahalanobis distance which is parameterized by a positive semidefinite (PSD) matrix M. An efficient method of estimating M is to treat M as a linear combination of rank-one matrices that can be learned using a boosting type approach. However, such approaches have two main drawbacks. First, the weight change across the training samples may be non-smooth. Second, the learned rank-one matrices might be redundant. In this paper, we propose a doubly regularized metric learning algorithm, termed by DRMetric, which imposes two regularizations on the conventional metric learning method. First, a regularization is applied on the weight of the training examples, which prevents unstable change of the weights and also prevents outlier examples from being weighed too much. Besides, a regularization is applied on the rank-one matrices to make them independent. This greatly reduces the redundancy of the rank-one matrices. We present experiments depicting the performance of the proposed method on a variety of datasets for various applications.

[S5-P9B]

A Discriminative Data-Dependent Mixture-Model Approach for Multiple Instance Learning in Image Classification

Qifan Wang, Luo Si, and Dan Zhang

Multiple Instance Learning (MIL) has been widely used in various applications including image classification. However, existing MIL methods do not explicitly address the multi-target problem where the distributions of positive instances are likely to be multi-modal. This strongly limits the performance of multiple instance learning in many real world applications. To address this problem, this paper proposes a novel discriminative data-dependent mixture-model method for multiple instance learning (MM-MIL) approach in image classification. The new method explicitly handles the multi-target problem by introducing a data-dependent mixture model, which allows positive instances to come from different clusters in a flexible manner. Furthermore, the kernelized representation of the proposed model allows effective and efficient learning in high dimensional feature space. An extensive set of experimental results demonstrate that the proposed new MM-MIL approach substantially outperforms several state-of-art MIL algorithms on benchmark datasets.

[S5-P10B]

No Bias Left behind: Covariate Shift Adaptation for Discriminative 3D Pose Estimation

Makoto Yamada, Leonid Sigal, and Michalis Raptis

Discriminative, or (structured) prediction, methods have proved effective for variety of problems in computer vision; a notable example is 3D monocular pose estimation. All methods to date, however, relied on an assumption that training (source) and test (target) data come from the same underlying joint distribution. In many real cases, including standard datasets, this assumption is flawed. In presence of training set bias, the learning results in a biased model whose performance degrades on the (target) test set. Under the assumption of covariate shift we propose an unsupervised domain adaptation approach to address this problem. The approach takes the form of training instance re-weighting, where the weights are assigned based on the ratio of training and test marginals evaluated at the samples. Learning with the resulting weighted training samples, alleviates the bias in the learned models. We show the efficacy of our approach by proposing weighted variants of Kernel Regression (KR) and Twin Gaussian Processes (TGP). We show that our weighted variants outperform their un-weighted counterparts and improve on the state-of-the-art performance in the public (HumanEva) dataset.

[S5-P11B]

Labeling Images by Integrating Sparse Multiple Distance Learning and Semantic Context Modeling

Chuanjun Ji, Xiangdong Zhou, Lan Lin, and Weidong Yang

Recent progress on Automatic Image Annotation (AIA) is achieved by either exploiting low level visual features or high level semantic context. Integrating these two paradigms to further leverage the performance of AIA is promising. However, very few previous works have studied this issue in a unified framework. In this paper, we propose a unified model based on Conditional Random Fields (CRF), which establishes tight interaction between visual features and semantic context. In particular, Kernelized Logistic Regression (KLR) with multiple visual distance learning is embedded into the CRF framework. We introduce L1 and L2 regularization terms into the unified learning process for the distance learning and the parameters penalty respectively. The experiments are conducted on two benchmarks: Corel and TRECVID-2005 data sets for evaluation. The experimental results show that, compared with the state-of-the-art methods, the unified model achieves significant improvement on annotation performance and shows more robustness with increasing number of various visual features.

[S5-P12B]

Exploiting the Circulant Structure of Tracking-by-Detection with Kernels

João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista

Recent years have seen greater interest in the use of discriminative classifiers in tracking systems, owing to their success in object detection. They are trained online with samples collected during tracking. Unfortunately, the potentially large number of samples becomes a computational burden, which directly conflicts with real-time requirements. On the other hand, limiting the samples may sacrifice performance. Interestingly, we observed that, as we add more and more samples, the problem acquires circulant structure. Using the well-established theory of Circulant matrices, we provide a link to Fourier analysis that opens up the possibility of extremely fast learning and detection with the Fast Fourier Transform. This can be done in the dual space of kernel machines as fast as with linear classifiers. We derive closed-form solutions for training and detection with several types of kernels, including the popular Gaussian and polynomial kernels. The resulting tracker achieves performance competitive with the state-of-the-art, can be implemented with only a few lines of code and runs at hundreds of frames-per-second. MATLAB code is provided in the paper (see Algorithm 1).

[S5-P13B]

Online Spatio-temporal Structural Context Learning for Visual Tracking

Longyin Wen, Zhaowei Cai, Zhen Lei, Dong Yi, and Stan Z. Li

Visual tracking is a challenging problem, because the target frequently change its appearance, randomly move its location and get occluded by other objects in unconstrained environments. The state changes of the target are temporally and spatially continuous, in this paper therefore, a robust Spatio-Temporal structural context based Tracker (STT) is presented to complete the tracking task in unconstrained environments. The temporal context capture the historical appearance information of the target to prevent the tracker from drifting to the background in a long term tracking. The spatial context model integrates contributors, which are the key-points automatically discovered around the target, to build a supporting field. The supporting field provides much more information than appearance of the target itself so that the location of the target will be predicted more precisely. Extensive experiments on various challenging databases demonstrate the superiority of our proposed tracker over other state-of-the-art trackers.

[S5-P14B]

Automatic Tracking of a Large Number of Moving Targets in 3D

Ye Liu, Hui Li, and Yan Qiu Chen

This paper addresses the problem of tracking a large number of targets moving in 3D space using multiple calibrated video cameras. Most visual details of the targets are lost in the captured images because of limited image resolution, and the remainder can be easily corrupted due to frequent occlusion, which makes it difficult to determine both across-view and temporal correspondences. We propose a fully automatic tracking system that is capable of detecting and tracking a large number of flying targets in a 3D volume. The system includes a 3D tracking method in the framework of particle filter. Different from previous 2D tracking methods, the proposed method models the 3D attributes of targets and furthest collects weak visual information from multiple views, which makes the tracker robust against occlusion and distraction. The ambiguities in stereo matching when initializing trackers are handled by an effective multiple hypothesis generation and verification mechanism. The whole system is fully automatic in dealing with variable number of targets and robust against detection and matching errors. Our system has successfully been used by biologists to recover the 3D trajectories of hundreds of fruit flies flying freely in a 3D volume.

[S5-P15B]

Towards Optimal Non-rigid Surface Tracking

Martin Klaudiny, Chris Budd, and Adrian Hilton

This paper addresses the problem of optimal alignment of non-rigid surfaces from multi-view video observations to obtain a temporally consistent representation. Conventional non-rigid surface tracking performs frame-to-frame alignment which is subject to the accumulation of errors resulting in drift over time. Recently, non-sequential tracking approaches have been introduced which re-order the input data based on a dissimilarity measure. One or more input sequences are represented in a tree with reducing alignment path length. This limits drift and increases robustness to large non-rigid deformations. However, jumps may occur in the aligned mesh sequence where tree branches meet due to independent error accumulation. Optimisation of the tree for non-sequential tracking is proposed to minimise the errors in temporal consistency due to both the drift and jumps. A novel cluster tree enforces sequential tracking in local segments of the sequence while allowing global non-sequential traversal among these segments. This provides a mechanism to create a tree structure which reduces the number of jumps between branches and limits the length of branches. Comprehensive evaluation is performed on a variety of challenging non-rigid surfaces including faces, cloth and people. This demonstrates that the proposed cluster tree achieves better temporal consistency than the previous sequential and non-sequential tracking approaches. Quantitative ground-truth comparison on a synthetic facial performance shows reduced error with the cluster tree.

[S5-P16B]

Full Body Performance Capture under Uncontrolled and Varying Illumination: A Shading-Based Approach

Chenglei Wu, Kiran Varanasi, and Christian Theobalt

This paper presents a marker-less method for full body human performance capture by analyzing shading information from a sequence of multi-view images, which are recorded under uncontrolled and changing lighting conditions. Both the articulated motion of the limbs and then the fine-scale surface detail are estimated in a temporally coherent manner. In a temporal framework, differential 3D human pose-changes from the previous time-step are expressed in terms of constraints on the visible image displacements derived from shading cues, estimated albedo and estimated scene illumination. The incident illumination at each frame are estimated jointly with pose, by assuming the Lambertian model of reflectance. The proposed method is independent of image silhouettes and training data, and is thus applicable in cases where background segmentation cannot be performed or a set of training poses is unavailable. We show results on challenging cases for pose-tracking such as changing backgrounds, occlusions and changing lighting conditions.

[S5-P17B]

Automatic Exposure Correction of Consumer Photographs

Lu Yuan and Jian Sun

We study the problem of automatically correcting the exposure of an input image. Generic auto-exposure correction methods usually fail in individual over-/under-exposed regions. Interactive corrections may fix this issue, but adjusting every photograph requires skill and time. This paper will automate the interactive correction technique by estimating the image specific S-shaped non-linear tone curve that best fits the input image. Our first contribution is a new Zone-based region-level optimal exposure evaluation, which would consider both the visibility of individual regions and relative contrast between regions. Then a detail-preserving S-curve adjustment is applied based on the optimal exposure to obtain the final output. We show that our approach enables better corrections comparing with popular image editing tools and other automatic methods.

[S5-P18B]

Image Guided Tone Mapping with Locally Nonlinear Model

Huxiang Gu, Ying Wang, Shiming Xiang, Gaofeng Meng, and Chunhong Pan

In this paper, we propose an effective locally nonlinear tone mapping algorithm for compressing the High Dynamic Range (HDR) images. Instead of linearly scaling the luminance of pixels, our core idea is to introduce local gamma correction with adaptive parameters on small overlapping patches over the entire input image. A framework for HDR image compression is then introduced, in which the global optimization problem is deduced and two guided images are adopted to induct the optimum solution. The optimal compression can finally be achieved by solving the optimization problem which can be transformed to a sparse linear equation. Extensive experimental results on a variety of HDR images and a carefully designed perceptually evaluation have demonstrated that our approach can achieve better performances than the state-of-the-art approaches.

[S5-P19B]

A Comparison of the Statistical Properties of IQA Databases Relative to a Set of Newly Captured High-Definition Images

Javier Silvestre-Blanes, Ian van der Linde, and Rubén Pérez-Lloréns

A broad range of image processing applications require image databases during development and testing. Whilst some image databases have been assembled with specific applications in mind, others are intended for more general use, with image content that is purposefully not application-specific. General-purpose image databases are in frequent use in the development of new compression algorithms, including in the evaluation of the efficacy of lossy compression techniques via statistical and human (perceptual) image quality assessment methods. The question of how the images featuring in standard image databases are selected is important, but is rarely quantitatively justified. In this article, we describe the compilation of a new image database of high-definition color images. We present statistical analyzes both of the images that feature in the most widely used extant databases, and the new database that we have compiled, in order to evaluate how broad a range of the statistics measured each database spans.

[S5-P20B]

Supervised Assessment of Segmentation Hierarchies

Jordi Pont-Tuset and Ferran Marques

This paper addresses the problem of the supervised assessment of hierarchical region-based image representations. Given the large amount of partitions represented in such structures, the supervised assessment approaches in the literature are based on selecting a reduced set of representative partitions and evaluating their quality. Assessment results, therefore, depend on the partition selection strategy used. Instead, we propose to find the partition in the tree that best matches the ground-truth partition, that is, the upper-bound partition selection. We show that different partition selection algorithms can lead to different conclusions regarding the quality of the assessed trees and that the upper-bound partition selection provides the following advantages: 1) it does not limit the assessment to a reduced set of partitions, and 2) it better discriminates the random trees from actual ones, which reflects a better qualitative behavior. We model the problem as a Linear Fractional Combinatorial Optimization (LFCO) problem, which makes the upper-bound selection feasible and efficient.

[S5-P21B]

Image Labeling on a Network: Using Social-Network Metadata for Image Classification

Julian McAuley and Jure Leskovec

Large-scale image retrieval benchmarks invariably consist of images from the Web. Many of these benchmarks are derived from online photo sharing networks, like Flickr, which in addition to hosting images also provide a highly interactive social community. Such communities generate rich metadata that can naturally be harnessed for image classification and retrieval. Here we study four popular benchmark datasets, extending them with social-network metadata, such as the groups to which each image belongs, the comment thread associated with the image, who uploaded it, their location, and their network of friends. Since these types of data are inherently relational, we propose a model that explicitly accounts for the interdependencies between images sharing common properties. We model the task as a binary labeling problem on a network, and use structured learning techniques to learn model parameters. We find that social-network metadata are useful in a variety of classification tasks, in many cases outperforming methods based on image content.

[S5-P22B]

Segmentation Based Particle Filtering for Real-Time 2D Object Tracking

Vasileios Belagiannis, Falk Schubert, Nassir Navab, and Slobodan Ilic

We address the problem of visual tracking of arbitrary objects that undergo significant scale and appearance changes. The classical tracking methods rely on the bounding box surrounding the target object. Regardless of the tracking approach, the use of bounding box quite often introduces background information. This information propagates in time and its accumulation quite often results in drift and tracking failure. This is particularly the case with the particle filtering approach that is often used for visual tracking. However, it always uses a bounding box around the object to compute features of the particle samples. Since this causes the drift, we propose to use segmentation for sampling. Relying on segmentation and computing the colour and gradient orientation histograms from these segmented particle samples allows the tracker to easily adapt to the object's deformations, occlusions, orientation, scale and appearance changes. We propose two particle sampling strategies based on segmentation. In the first, segmentation is done for every propagated particle sample, while in the second only the strongest particle sample is segmented. Depending on this decision there is obviously a trade-off between speed and performance. We perform an exhaustive quantitative evaluation on a number of challenging sequences and compare our method with the number of state-of-the-art methods previously evaluated on those sequences. The results we obtain outperform majority of the related work, both in terms of the performance and speed.

[S5-P23B]

Online Video Segmentation by Bayesian Split-Merge Clustering

Juho Lee, Suha Kwak, Bohyung Han, and Seungjin Choi

We present an online video segmentation algorithm based on a novel nonparametric Bayesian clustering method called Bayesian Split-Merge Clustering (BSMC). BSMC can efficiently cluster dynamically changing data through split and merge processes at each time step, where the decision for splitting and merging is made by approximate posterior distributions over partitions with Dirichlet Process (DP) priors. Moreover, BSMC sidesteps the difficult problem of finding the proper number of clusters by virtue of the flexibility of nonparametric Bayesian models. We naturally apply BSMC to online video segmentation, which is composed of three steps—pixel clustering, histogram-based merging and temporal matching. We demonstrate the performance of our algorithm on complex real video sequences compared to other existing methods.

[S5-P24B]

Joint Classification-Regression Forests for Spatially Structured Multi-object Segmentation

Ben Glocker, Olivier Pauly, Ender Konukoglu, and Antonio Criminisi

In many segmentation scenarios, labeled images contain rich structural information about spatial arrangement and shapes of the objects. Integrating this rich information into supervised learning techniques is promising as it generates models which go beyond learning class association, only. This paper proposes a new supervised forest model for joint classification-regression which exploits both class and structural information. Training our model is achieved by optimizing a joint objective function of pixel classification and shape regression. Shapes are represented implicitly via signed distance maps obtained directly from ground truth label maps. Thus, we can associate each image point not only with its class label, but also with its distances to object boundaries, and this at no additional cost regarding annotations. The regression component acts as spatial regularization learned from data and yields a predictor with both class and spatial consistency. In the challenging context of simultaneous multi-organ segmentation, we demonstrate the potential of our approach through experimental validation on a large dataset of 80 three-dimensional CT scans.

ORAL SESSION 5
MRFs AND EARLY VISION

Wednesday, October 10
11:20 - 13:00

[S5-O1]

Diverse M-Best Solutions in Markov Random Fields

Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich

Much effort has been directed at algorithms for obtaining the highest probability (MAP) configuration in probabilistic (random field) models. In many situations, one could benefit from additional high-probability solutions. Current methods for computing the M most probable configurations produce solutions that tend to be very similar to the MAP solution and each other. This is often an undesirable property. In this paper we propose an algorithm for the Diverse M-Best problem, which involves finding a diverse set of highly probable solutions under a discrete probabilistic model. Given a dissimilarity function measuring closeness of two solutions, our formulation involves maximizing a linear combination of the probability and dissimilarity to previous solutions. Our formulation generalizes the M-Best MAP problem and we show that for certain families of dissimilarity functions we can guarantee that these solutions can be found as easily as the MAP solution.

[S5-O2]

Generic Cuts: An Efficient Algorithm for Optimal Inference in Higher Order MRF-MAP

Chetan Arora, Subhashis Banerjee, Prem Kalra, and S.N. Maheshwari

We propose a new algorithm called Generic Cuts for computing optimal solutions to 2 label MRF-MAP problems with higher order clique potentials satisfying submodularity. The algorithm runs in time $O(2^k n^3)$ in the worst case (k is clique order and n is the number of pixels). A special gadget is introduced to model flows in a high order clique and a technique for building a flow graph is specified. Based on the primal dual structure of the optimization problem the notions of capacity of an edge and cut are generalized to define a flow problem. We show that in this flow graph max flow is equal to min cut which also is the optimal solution to the problem when potentials are submodular. This is in contrast to all prevalent techniques of optimizing Boolean energy functions involving higher order potentials including those based on reductions to quadratic potential functions which provide only approximate solutions even for submodular functions. We show experimentally that our implementation of the Generic Cuts algorithm is more than an order of magnitude faster than all algorithms including reduction based whose outputs on submodular potentials are near optimal.

[S5-O3]

Filter-Based Mean-Field Inference for Random Fields with Higher-Order Terms and Product Label-Spaces

Vibhav Vineet, Jonathan Warrell, and Philip H.S. Torr

Recently, a number of cross bilateral filtering methods have been proposed for solving multi-label problems in computer vision, such as stereo, optical flow and object class segmentation that show an order of magnitude improvement in speed over previous methods. These methods have achieved good results despite using models with only unary and/or pairwise terms. However, previous work has shown the value of using models with higher-order terms e.g. to represent label consistency over large regions, or global co-occurrence relations. We show how these higher-order terms can be formulated such that filter-based inference remains possible. We demonstrate our techniques on joint stereo and object labeling problems, as well as object class segmentation, showing in addition for joint object-stereo labeling how our method provides an efficient approach to inference in product label-spaces. We show that we are able to speed up inference in these models around 10-30 times with respect to competing graph-cut/move-making methods, as well as maintaining or improving accuracy in all cases. We show results on PascalVOC-10 for object class segmentation, and Leuven for joint object-stereo labeling.

[S5-O4]

Continuous Markov Random Fields for Robust Stereo Estimation

Koichiro Yamaguchi, Tamir Hazan, David McAllester, and Raquel Urtasun

In this paper we present a novel slanted-plane model which reasons jointly about occlusion boundaries as well as depth. We formulate the problem as one of inference in a hybrid MRF composed of both continuous (i.e., slanted 3D planes) and discrete (i.e., occlusion boundaries) random variables. This allows us to define potentials encoding the ownership of the pixels that compose the boundary between segments, as well as potentials encoding which junctions are physically possible. Our approach outperforms the state-of-the-art on Middlebury high resolution imagery [1] as well as in the more challenging KITTI dataset [2], while being more efficient than existing slanted plane MRF methods, taking on average 2 minutes to perform inference on high resolution imagery.

[S5-O5]

Good Regions to Deblur

Zhe Hu and Ming-Hsuan Yang

The goal of single image deblurring is to recover both a latent clear image and an underlying blur kernel from one input blurred image. Recent works focus on exploiting natural image priors or additional image observations for deblurring, but pay less attention to the influence of image structures on estimating blur kernels. What is the useful image structure and how can one select good regions for deblurring? We formulate the problem of learning good regions for deblurring within the Conditional Random Field framework. To better compare blur kernels, we develop an effective similarity metric for labeling training samples. The learned model is able to predict good regions from an input blurred image for deblurring without user guidance. Qualitative and quantitative evaluations demonstrate that good regions can be selected by the proposed algorithms for effective image deblurring.

[S5-O6]

Patch Complexity, Finite Pixel Correlations and Optimal Denoising

Anat Levin, Boaz Nadler, Fredo Durand, and William T. Freeman

Image restoration tasks are ill-posed problems, typically solved with priors. Since the optimal prior is the exact unknown density of natural images, actual priors are only approximate and typically restricted to small patches. This raises several questions: How much may we hope to improve current restoration results with future sophisticated algorithms? And more fundamentally, even with perfect knowledge of natural image statistics, what is the inherent ambiguity of the problem? In addition, since most current methods are limited to finite support patches or kernels, what is the relation between the patch complexity of natural images, patch size, and restoration errors? Focusing on image denoising, we make several contributions. First, in light of computational constraints, we study the relation between denoising gain and sample size requirements in a non parametric approach. We present a law of diminishing return, namely that with increasing patch size, rare patches not only require a much larger dataset, but also gain little from it. This result suggests novel adaptive variable-sized patch schemes for denoising. Second, we study absolute denoising limits, regardless of the algorithm used, and the converge rate to them as a function of patch size. Scale invariance of natural images plays a key role here and implies both a strictly positive lower bound on denoising and a power law convergence. Extrapolating this parametric law gives a ballpark estimate of the best achievable denoising, suggesting that some improvement, although modest, is still possible.

POSTER SESSION 6

Wednesday, October 10
14:30 - 17:00

[S6-P1A]

Guaranteed Ellipse Fitting with the Sampson Distance

Zygmunt L. Szpak, Wojciech Chojnacki, and Anton van den Hengel

When faced with an ellipse fitting problem, practitioners frequently resort to algebraic ellipse fitting methods due to their simplicity and efficiency. Currently, practitioners must choose between algebraic methods that guarantee an ellipse fit but exhibit high bias, and geometric methods that are less biased but may no longer guarantee an ellipse solution. We address this limitation by proposing a method that strikes a balance between these two objectives. Specifically, we propose a fast stable algorithm for fitting a guaranteed ellipse to data using the Sampson distance as a data-parameter discrepancy measure. We validate the stability, accuracy, and efficiency of our method on both real and synthetic data. Experimental results show that our algorithm is a fast and accurate approximation of the computationally more expensive orthogonal-distance-based ellipse fitting method. In view of these qualities, our method may be of interest to practitioners who require accurate and guaranteed ellipse estimates.

[S6-P2A]

A Locally Linear Regression Model for Boundary Preserving Regularization in Stereo Matching

Shengqi Zhu, Li Zhang, and Hailin Jin

We propose a novel regularization model for stereo matching that uses large neighborhood windows. The model is based on the observation that in a local neighborhood there exists a linear relationship between pixel values and disparities. Compared to the traditional boundary preserving regularization models that use adjacent pixels, the proposed model is robust to image noise and captures higher level interactions. We develop a globally optimized stereo matching algorithm based on this regularization model. The algorithm alternates between finding a quadratic upper bound of the relaxed energy function and solving the upper bound using iterative reweighted least squares. To reduce the chance of being trapped in local minima, we propose a progressive convex-hull filter to tighten the data cost relaxation. Our evaluation on the Middlebury datasets shows the effectiveness of our method in preserving boundary sharpness while keeping regions smooth. We also evaluate our method on a wide range of challenging real-world videos. Experimental results show that our method outperforms existing methods in temporal consistency.

[S6-P3A]

A Novel Fast Method for L_∞ Problems in Multiview Geometry

Zhijun Dai, Yihong Wu, Fengjun Zhang, and Hongan Wang

Optimization using the L_∞ norm is an increasingly important area in multiview geometry. Previous work has shown that globally optimal solutions can be computed reliably using the formulation of generalized fractional programming, in which algorithms solve a sequence of convex problems independently to approximate the optimal L_∞ norm error. We found the sequence of convex problems are highly related and we propose a method to derive a Newton-like step from any given point. In our method, the feasible region of the current involved convex problem is contracted gradually along with the Newton-like steps, and the updated point locates on the boundary of the new feasible region. We propose an effective strategy to make the boundary point become an interior point through one dimension augmentation and relaxation. Results are presented and compared to the state of the art algorithms on simulated and real data for some multiview geometry problems with improved performance on both runtime and Newton-like iterations.

[S6-P4A]

Visibility Probability Structure from SfM Datasets and Applications

Siddharth Choudhary and P.J. Narayanan

Large scale reconstructions of camera matrices and point clouds have been created using structure from motion from community photo collections. Such a dataset is rich in information; it represents a sampling of the geometry and appearance of the underlying space. In this paper, we encode the visibility information between and among points and cameras as visibility probabilities. The conditional visibility probability of a set of points on a point (or a set of cameras on a camera) can rank points (or cameras) based on their mutual dependence. We combine the conditional probability with a distance measure to prioritize points for fast guided search for the image localization problem. We define dual problem of feature triangulation as finding the 3D coordinates of a given image feature point. We use conditional visibility probability to quickly identify a subset of cameras in which a feature is visible.

[S6-P5A]

A Generative Model for Online Depth Fusion

Oliver J. Woodford and George Vogiatzis

We present a probabilistic, online, depth map fusion framework, whose generative model for the sensor measurement process accurately incorporates both long-range visibility constraints and a spatially varying, probabilistic outlier model. In addition, we propose an inference algorithm that updates the state variables of this model in linear time each frame. Our detailed evaluation compares our approach against several others, demonstrating and explaining the improvements that this model offers, as well as highlighting a problem with all current methods: systemic bias.

[S6-P6A]

Depth Recovery Using an Adaptive Color-Guided Auto-Regressive Model

Jingyu Yang, Xinchen Ye, Kun Li, and Chunping Hou

This paper proposes an adaptive color-guided auto-regressive (AR) model for high quality depth recovery from low quality measurements captured by depth cameras. We formulate the depth recovery task into a minimization of AR prediction errors subject to measurement consistency. The AR predictor for each pixel is constructed according to both the local correlation in the initial depth map and the nonlocal similarity in the accompanied high quality color image. Experimental results show that our method outperforms existing state-of-the-art schemes, and is versatile for both mainstream depth sensors: ToF camera and Kinect.

[S6-P7A]

Learning Hybrid Part Filters for Scene Recognition

Yingbin Zheng, Yu-Gang Jiang, and Xiangyang Xue

This paper introduces a new image representation for scene recognition, where an image is described based on the response maps of object part filters. The part filters are learned from existing datasets with object location annotations, using deformable part-based models trained by latent SVM [1]. Since different objects may contain similar parts, we describe a method that uses a semantic hierarchy to automatically determine and merge filters shared by multiple objects. The merged hybrid filters are then applied to new images. Our proposed representation, called Hybrid-Parts, is generated by pooling the response maps of the hybrid filters. Contrast to previous scene recognition approaches that adopted object-level detections as feature inputs, we harness filter responses of object parts, which enable a richer and finer-grained representation. The use of the hybrid filters is important towards a more compact representation, compared to directly using all the original part filters. Through extensive experiments on several scene recognition benchmarks, we demonstrate that Hybrid-Parts outperforms recent state-of-the-arts, and combining it with standard low-level features such as the GIST descriptor can lead to further improvements.

[S6-P8A]

Parametric Manifold of an Object under Different Viewing Directions

Xiaozheng Zhang, Yongsheng Gao, and Terry Caelli

The appearance of a 3D object depends on both the viewing directions and illumination conditions. It is proven that all n-pixel images of a convex object with Lambertian surface under variable lighting from infinity form a convex polyhedral cone (called illumination cone) in n-dimensional space. This paper tries to answer the other half of the question: What is the set of images of an object under all viewing directions? A novel image representation is proposed, which transforms any n-pixel image of a 3D object to a vector in a 2n-dimensional pose space. In such a pose space, we prove that the transformed images of a 3D object under all viewing directions form a parametric manifold in a 6-dimensional linear subspace. With in-depth rotations along a single axis in particular, this manifold is an ellipse. Furthermore, we show that this parametric pose manifold of a convex object can be estimated from a few images in different poses and used to predict object's appearances under unseen viewing directions. These results immediately suggest a number of approaches to object recognition, scene detection, and 3D modelling. Experiments on both synthetic data and real images were reported, which demonstrates the validity of the proposed representation.

[S6-P9A]

Fast Approximations to Structured Sparse Coding and Applications to Object Classification

Arthur Szlam, Karol Gregor, and Yann LeCun

We describe a method for fast approximation of sparse coding. A given input vector is passed through a binary tree. Each leaf of the tree contains a subset of dictionary elements. The coefficients corresponding to these dictionary elements are allowed to be nonzero and their values are calculated quickly by multiplication with a precomputed pseudoinverse. The tree parameters, the dictionary, and the subsets of the dictionary corresponding to each leaf are learned. In the process of describing this algorithm, we discuss the more general problem of learning the groups in group structured sparse modeling. We show that our method creates good sparse representations by using it in the object recognition framework of [1,2]. Implementing our own fast version of the SIFT descriptor the whole system runs at 20 frames per second on 321 × 481 sized images on a laptop with a quad-core cpu, while sacrificing very little accuracy on the Caltech 101, Caltech 256, and 15 scenes benchmarks.

[S6-P10A]

Displacement Template with Divide-&-Conquer Algorithm for Significantly Improving Descriptor Based Face Recognition Approaches

Liang Chen, Ling Yan, Yonghuai Liu, Lixin Gao, and Xiaoqin Zhang

This paper proposes a displacement template structure for improving descriptor based face recognition approaches. With this template structure, a face is represented by a template consisting of a set of piled blocks; each block pile consists of a few heavily overlapped blocks from the face image. An ensemble of blocks, one from each pile, is taken as a candidate image of the face. When a descriptor based approach is used, we are able to generate a displacement description template for the face by replacing each block in the template with its local description, where a concatenation of the local descriptions of the blocks, one from each pile, is taken to be a candidate description of the face. Using the description template together with a divide-and-conquer algorithm for computing the similarities between description templates, we have demonstrated the significantly improved performance of LBP, TPLBP and FPLBP templates over original LBP, TPLBP and FPLBP approaches by the experiments on benchmark face databases.

[S6-P11A]

Latent Pyramidal Regions for Recognizing Scenes

Fereshteh Sadeghi and Marshall F. Tappen

In this paper we propose a simple but efficient image representation for solving the scene classification problem. Our new representation combines the benefits of spatial pyramid representation using nonlinear feature coding and latent Support Vector Machine (LSVM) to train a set of Latent Pyramidal Regions (LPR). Each of our LPRs captures a discriminative characteristic of the scenes and is trained by searching over all possible sub-windows of the images in a latent SVM training procedure. Each LPR is represented in a spatial pyramid and uses non-linear locality constraint coding for learning both shape and texture patterns of the scene. The final response of the LPRs form a single feature vector which we call the LPR representation and can be used for the classification task. We tested our proposed scene representation model in three datasets which contain a variety of scene categories (15-Scenes, UIUC-Sports and MIT-indoor). Our LPR representation obtains state-of-the-art results on all these datasets which shows that it can simultaneously model the global and local scene characteristics in a single framework and is general enough to be used for both indoor and outdoor scene classification.

[S6-P12A]

Augmented Attribute Representations

Viktoriia Sharmanska, Novi Quadrianto, and Christoph H. Lampert

We propose a new learning method to infer a mid-level feature representation that combines the advantage of semantic attribute representations with the higher expressive power of non-semantic features. The idea lies in augmenting an existing attribute-based representation with additional dimensions for which an autoencoder model is coupled with a large-margin principle. This construction allows a smooth transition between the zero-shot regime with no training example, the unsupervised regime with training examples but without class labels, and the supervised regime with training examples and with class labels. The resulting optimization problem can be solved efficiently, because several of the necessity steps have closed-form solutions. Through extensive experiments we show that the augmented representation achieves better results in terms of object categorization accuracy than the semantic representation alone.

[S6-P13A]

Exploring the Spatial Hierarchy of Mixture Models for Human Pose Estimation

Yuandong Tian, C. Lawrence Zitnick, and Srinivasa G. Narasimhan

Human pose estimation requires a versatile yet well-constrained spatial model for grouping locally ambiguous parts together to produce a globally consistent hypothesis. Previous works either use local deformable models deviating from a certain template, or use a global mixture representation in the pose space. In this paper, we propose a new hierarchical spatial model that can capture an exponential number of poses with a compact mixture representation on each part. Using latent nodes, it can represent high-order spatial relationship among parts with exact inference. Different from recent hierarchical models that associate each latent node to a mixture of appearance templates (like HoG), we use the hierarchical structure as a pure spatial prior avoiding the large and often confounding appearance space. We verify the effectiveness of this model in three ways. First, samples representing human-like poses can be drawn from our model, showing its ability to capture high-order dependencies of parts. Second, our model achieves accurate reconstruction of unseen poses compared to a nearest neighbor pose representation. Finally, our model achieves state-of-art performance on three challenging datasets, and substantially outperforms recent hierarchical models.

[S6-P14A]

People Orientation Recognition by Mixtures of Wrapped Distributions on Random Trees

Davide Baltieri, Roberto Vezzani, and Rita Cucchiara

The recognition of people orientation in single images is still an open issue in several real cases, when the image resolution is poor, body parts cannot be distinguished and localized or motion cannot be exploited. However, the estimation of a person orientation, even an approximated one, could be very useful to improve people tracking and re-identification systems, or to provide a coarse alignment of body models on the input images. In these situations, holistic features seem to be more effective and faster than model based 3D reconstructions. In this paper we propose to describe the people appearance with multi-level HoG feature sets and to classify their orientation using an array of Extremely Randomized Trees classifiers trained on quantized directions. The outputs of the classifiers are then integrated into a global continuous probability density function using a Mixture of Approximated Wrapped Gaussian distributions. Experiments on the TUD Multiview Pedestrians, the Sarc3D, and the 3DPeS datasets confirm the efficacy of the method and the improvement with respect to state of the art approaches.

[S6-P15A]

Hybrid Classifiers for Object Classification with a Rich Background

Margarita Osadchy, Daniel Keren, and Bella Fadida-Specktor

The majority of current methods in object classification use the one-against-rest training scheme. We argue that when applied to a large number of classes, this strategy is problematic: as the number of classes increases, the negative class becomes a very large and complicated collection of images. The resulting classification problem then becomes extremely unbalanced, and kernel SVM classifiers trained on such sets require long training time and are slow in prediction. To address these problems, we propose to consider the negative class as a background and characterize it by a prior distribution. Further, we propose to construct ″hybrid″ classifiers, which are trained to separate this distribution from the samples of the positive class. A typical classifier first projects (by a function which may be non-linear) the inputs to a one-dimensional space, and then thresholds this projection. Theoretical results and empirical evaluation suggest that, after projection, the background has a relatively simple distribution, which is much easier to parameterize and work with. Our results show that hybrid classifiers offer an advantage over SVM classifiers, both in performance and complexity, especially when the negative (background) class is large.

[S6-P16A]

Unsupervised and Supervised Visual Codes with Restricted Boltzmann Machines

Hanlin Goh, Nicolas Thome, Matthieu Cord, and Joo-Hwee Lim

Recently, the coding of local features (e.g. SIFT) for image categorization tasks has been extensively studied. Incorporated within the Bag of Words (BoW) framework, these techniques optimize the projection of local features into the visual codebook, leading to state-of-the-art performances in many benchmark datasets. In this work, we propose a novel visual codebook learning approach using the restricted Boltzmann machine (RBM) as our generative model. Our contribution is three-fold. Firstly, we steer the unsupervised RBM learning using a regularization scheme, which decomposes into a combined prior for the sparsity of each feature's representation as well as the selectivity for each codeword. The codewords are then fine-tuned to be discriminative through the supervised learning from top-down labels. Secondly, we evaluate the proposed method with the Caltech-101 and 15-Scenes datasets, either matching or outperforming state-of-the-art results. The codebooks are compact and inference is fast. Finally, we introduce an original method to visualize the codebooks and decipher what each visual codeword encodes.

[S6-P17A]

A New Biologically Inspired Color Image Descriptor

Jun Zhang, Youssef Barhomi, and Thomas Serre

We describe a novel framework for the joint processing of color and shape information in natural images. A hierarchical non-linear spatio-chromatic operator yields spatial and chromatic opponent channels, which mimics processing in the primate visual cortex. We extend two popular object recognition systems (i.e., the Hmax hierarchical model of visual processing and a sift-based bag-of-words approach) to incorporate color information along with shape information. We further use the framework in combination with the gist algorithm for scene categorization as well as the Berkeley segmentation algorithm. In all cases, the proposed approach is shown to outperform standard grayscale/shape-based descriptors as well as alternative color processing schemes on several datasets.

[S6-P18A]

Finding Correspondence from Multiple Images via Sparse and Low-Rank Decomposition

Zinan Zeng, Tsung-Han Chan, Kui Jia, and Dong Xu

We investigate the problem of finding the correspondence from multiple images, which is a challenging combinatorial problem. In this work, we propose a robust solution by exploiting the priors that the rank of the ordered patterns from a set of linearly correlated images should be lower than that of the disordered patterns, and the errors among the reordered patterns are sparse. This problem is equivalent to find a set of optimal partial permutation matrices for the disordered patterns such that the rearranged patterns can be factorized as a sum of a low rank matrix and a sparse error matrix. A scalable algorithm is proposed to approximate the solution by solving two sub-problems sequentially: minimization of the sum of nuclear norm and l^1 norm for solving relaxed partial permutation matrices, followed by a binary integer programming to project each relaxed partial permutation matrix to the feasible solution. We verify the efficacy and robustness of the proposed method with extensive experiments with both images and videos.

[S6-P19A]

Multidimensional Spectral Hashing

Yair Weiss, Rob Fergus, and Antonio Torralba

With the growing availability of very large image databases, there has been a surge of interest in methods based on ''semantic hashing'', i.e. compact binary codes of data-points so that the Hamming distance between codewords correlates with similarity. In reviewing and comparing existing methods, we show that their relative performance can change drastically depending on the definition of ground-truth neighbors. Motivated by this finding, we propose a new formulation for learning binary codes which seeks to reconstruct the affinity between datapoints, rather than their distances. We show that this criterion is intractable to solve exactly, but a spectral relaxation gives an algorithm where the bits correspond to thresholded eigenvectors of the affinity matrix, and as the number of datapoints goes to infinity these eigenvectors converge to eigenfunctions of Laplace-Beltrami operators, similar to the recently proposed Spectral Hashing (SH) method. Unlike SH whose performance may degrade as the number of bits increases, the optimal code using our formulation is guaranteed to faithfully reproduce the affinities as the number of bits increases. We show that the number of eigenfunctions needed may increase exponentially with dimension, but introduce a ''kernel trick'' to allow us to compute with an exponentially large number of bits but using only memory and computation that grows linearly with dimension. Experiments shows that MDSH outperforms the state-of-the art, especially in the challenging regime of small distance thresholds.

[S6-P20A]

What Makes a Good Detector? – Structured Priors for Learning from Few Examples

Tianshi Gao, Michael Stark, and Daphne Koller

Transfer learning can counter the heavy-tailed nature of the distribution of training examples over object classes. Here, we study transfer learning for object class detection. Starting from the intuition that "what makes a good detector" should manifest itself in the form of repeatable statistics over existing "good" detectors, we design a low-level feature model that can be used as a prior for learning new object class models from scarce training data. Our priors are structured, capturing dependencies both on the level of individual features and spatially neighboring pairs of features. We confirm experimentally the connection between the information captured by our priors and "good" detectors as well as the connection to transfer learning from sources of different quality. We give an in-depth analysis of our priors on a subset of the challenging PASCAL VOC 2007 data set and demonstrate improved average performance over all 20 classes, achieved without manual intervention.

[S6-P21A]

A Convolutional Treelets Binary Feature Approach to Fast Keypoint Recognition

Chenxia Wu, Jianke Zhu, Jiemi Zhang, Chun Chen, and Deng Cai

Fast keypoint recognition is essential to many vision tasks. In contrast to the classification-based approaches [1,2], we directly formulate the keypoint recognition as an image patch retrieval problem, which enjoys the merit of finding the matched keypoint and its pose simultaneously. A novel convolutional treelets approach is proposed to effectively extract the binary features from the patches. A corresponding sub-signature-based locality sensitive hashing scheme is employed for the fast approximate nearest neighbor search in patch retrieval. Experiments on both synthetic data and real-world images have shown that our method performs better than state-of-the-art descriptor-based and classification-based approaches.

[S6-P22A]

Categorizing Turn-Taking Interactions

Karthir Prabhakar and James M. Rehg

We address the problem of categorizing turn-taking interactions between individuals. Social interactions are characterized by turn-taking and arise frequently in real-world videos. Our approach is based on the use of temporal causal analysis to decompose a space-time visual word representation of video into co-occuring independent segments, called causal sets [1]. These causal sets then serves the input to a multiple instance learning framework to categorize turn-taking interactions. We introduce a new turn-taking interactions dataset consisting of social games and sports rallies. We demonstrate that our formulation of multiple instance learning (QP-MISVM) is better able to leverage the repetitive structure in turn-taking interactions and demonstrates superior performance relative to a conventional bag of words model.

[S6-P1B]

Local Expert Forest of Score Fusion for Video Event Classification

Jingchen Liu, Scott McCloskey, and Yanxi Liu

We address the problem of complicated event categorization from a large dataset of videos "in the wild", where multiple classifiers are applied independently to evaluate each video with a 'likelihood' score. The core contribution of this paper is a local expert forest model for meta-level score fusion for event detection under heavily imbalanced class distributions. Our motivation is to adapt to performance variations of the classifiers in different regions of the score space, using a divide-and-conquer technique. We propose a novel method to partition the likelihood-space, being sensitive to local label distributions in imbalanced data, and train a pair of locally optimized experts each time. Multiple pairs of experts based on different partitions ('trees') form a 'forest', balancing local adaptivity and over-fitting of the model. As a result, our model disregards classifiers in regions of the score space where their performance is bad, achieving both local source selection and fusion. We experiment with the TRECVID Multimedia Event Detection (MED) dataset, detecting 15 complicated events from around 34k video clips comprising more than 1000 hours, and demonstrate superior performance compared to other score-level fusion methods.

[S6-P2B]

View-Invariant Action Recognition Using Latent Kernelized Structural SVM

Xinxiao Wu and Yunde Jia

This paper goes beyond recognizing human actions from a fixed view and focuses on action recognition from an arbitrary view. A novel learning algorithm, called latent kernelized structural SVM, is proposed for the view-invariant action recognition, which extends the kernelized structural SVM framework to include latent variables. Due to the changing and frequently unknown positions of the camera, we regard the view label of action as a latent variable and implicitly infer it during both learning and inference. Motivated by the geometric correlation between different views and semantic correlation between different action classes, we additionally propose a mid-level correlation feature which describes an action video by a set of decision values from the pre-learned classifiers of all the action classes from all the views. Each decision value captures both geometric and semantic correlations between the action video and the corresponding action class from the corresponding view. After that, we combine the low-level visual cue, mid-level correlation description, and high-level label information into a novel nonlinear kernel under the latent kernelized structural SVM framework. Extensive experiments on multi-view IXMAS and MuHAVi action datasets demonstrate that our method generally achieves higher recognition accuracy than other state-of-the-art methods.

[S6-P3B]

Trajectory-Based Modeling of Human Actions with Motion Reference Points

Yu-Gang Jiang, Qi Dai, Xiangyang Xue, Wei Liu, and Chong-Wah Ngo

Human action recognition in videos is a challenging problem with wide applications. State-of-the-art approaches often adopt the popular bag-of-features representation based on isolated local patches or temporal patch trajectories, where motion patterns like object relationships are mostly discarded. This paper proposes a simple representation specifically aimed at the modeling of such motion relationships. We adopt global and local reference points to characterize motion information, so that the final representation can be robust to camera movement. Our approach operates on top of visual codewords derived from local patch trajectories, and therefore does not require accurate foreground-background separation, which is typically a necessary step to model object relationships. Through an extensive experimental evaluation, we show that the proposed representation offers very competitive performance on challenging benchmark datasets, and combining it with the bag-of-features representation leads to substantial improvement. On Hollywood2, Olympic Sports, and HMDB51 datasets, we obtain 59.5%, 80.6% and 40.7% respectively, which are the best reported results to date.

[S6-P4B]

PatchMatchGraph: Building a Graph of Dense Patch Correspondences for Label Transfer

Stephen Gould and Yuhang Zhang

We address the problem of semantic segmentation, or multi-class pixel labeling, by constructing a graph of dense overlapping patch correspondences across large image sets. We then transfer annotations from labeled images to unlabeled images using the established patch correspondences. Unlike previous approaches to non-parametric label transfer our approach does not require an initial image retrieval step. Moreover, we operate on a graph for computing mappings between images, which avoids the need for exhaustive pairwise comparisons. Consequently, we can leverage offline computation to enhance performance at test time. We conduct extensive experiments to analyze different variants of our graph construction algorithm and evaluate multi-class pixel labeling performance on several challenging datasets.

[S6-P5B]

A Unifying Theory of Active Discovery and Learning

Timothy M. Hospedales, Shaogang Gong, and Tao Xiang

For learning problems where human supervision is expensive, active query selection methods are often exploited to maximise the return of each supervision. Two problems where this has been successfully applied are active discovery – where the aim is to discover at least one instance of each rare class with few supervisions; and active learning – where the aim is to maximise a classifier's performance with least supervision. Recently, there has been interest in optimising these tasks jointly, i.e., active learning with undiscovered classes, to support efficient interactive modelling of new domains. Mixtures of active discovery and learning and other schemes have been exploited, but perform poorly due to heuristic objectives. In this study, we show with systematic theoretical analysis how the previously disparate tasks of active discovery and learning can be cleanly unified into a single problem, and hence are able for the first time to develop a unified query algorithm to directly optimise this problem. The result is a model which consistently outperforms previous attempts at active learning in the presence of undiscovered classes, with no need to tune parameters for different datasets.

[S6-P6B]

Extracting 3D Scene-Consistent Object Proposals and Depth from Stereo Images

Michael Bleyer, Christoph Rhemann, and Carsten Rother

This work combines two active areas of research in computer vision: unsupervised object extraction from a single image, and depth estimation from a stereo image pair. A recent, successful trend in unsupervised object extraction is to exploit so-called ‘‘3D scene-consistency’’, that is enforcing that objects obey underlying physical constraints of the 3D scene, such as occupancy of 3D space and gravity of objects. Our main contribution is to introduce the concept of 3D scene-consistency into stereo matching. We show that this concept is beneficial for both tasks, object extraction and depth estimation. In particular, we demonstrate that our approach is able to create a large set of 3D scene-consistent object proposals, by varying e.g. the prior on the number of objects. After automatically ranking the proposals we show experimentally that our results are considerably closer to ground truth than state-of-the-art techniques which either use stereo or monocular images. We envision that our method will build the front-end of a future object recognition system for stereo images.

[S6-P7B]

Repairing Sparse Low-Rank Texture

Xiao Liang, Xiang Ren, Zhengdong Zhang, and Yi Ma

In this paper, we show how to harness both low-rank and sparse structures in regular or near regular textures for image completion. Our method leverages the new convex optimization for low-rank and sparse signal recovery and can automatically correctly repair the global structure of a corrupted texture, even without precise information about the regions to be completed. Through extensive simulations, we show our method can complete and repair textures corrupted by errors with both random and contiguous supports better than existing low-rank matrix recovery methods. Through experimental comparisons with existing image completion systems (such as Photoshop) our method demonstrate significant advantage over local patch based texture synthesis techniques in dealing with large corruption, non-uniform texture, and large perspective deformation.

[S6-P8B]

Active Frame Selection for Label Propagation in Videos

Sudheendra Vijayanarasimhan and Kristen Grauman

Manually segmenting and labeling objects in video sequences is quite tedious, yet such annotations are valuable for learning-based approaches to object and activity recognition. While automatic label propagation can help, existing methods simply propagate annotations from arbitrarily selected frames (e.g., the first one) and so may fail to best leverage the human effort invested. We define an active frame selection problem: select k frames for manual labeling, such that automatic pixel-level label propagation can proceed with minimal expected error. We propose a solution that directly ties a joint frame selection criterion to the predicted errors of a flow-based random field propagation model. It selects the set of k frames that together minimize the total mislabeling risk over the entire sequence. We derive an efficient dynamic programming solution to optimize the criterion. Further, we show how to automatically determine how many total frames k should be labeled in order to minimize the total manual effort spent labeling and correcting propagation errors. We demonstrate our method's clear advantages over several baselines, saving hours of human effort per video.

[S6-P9B]

Non-causal Temporal Prior for Video Deblocking

Deqing Sun and Ce Liu

Real-world video sequences coded at low bit rates suffer from compression artifacts, which are visually disruptive and can cause problems to computer vision algorithms. Unlike the denoising problem where the high frequency components of the signal are present in the noisy observation, most high frequency details are lost during compression and artificial discontinuities arise across the coding block boundaries. In addition to sparse spatial priors that can reduce the blocking artifacts for a single frame, temporal information is needed to recover the lost spatial details. However, establishing accurate temporal correspondences from the compressed videos is challenging because of the loss of high frequency details and the increase of false blocking artifacts. In this paper, we propose a non-causal temporal prior model to reduce video compression artifacts by propagating information from adjacent frames and iterating between image reconstruction and motion estimation. Experimental results on real-world sequences demonstrate that the deblocked videos by the proposed system have marginal statistics of high frequency components closer to those of the original ones, and are better input for standard edge and corner detectors than the coded ones.

[S6-P10B]

Text Image Deblurring Using Text-Specific Properties

Hojin Cho, Jue Wang, and Seungyong Lee

State-of-the-art blind image deconvolution approaches have difficulties when dealing with text images, since they rely on natural image statistics which do not respect the special properties of text images. On the other hand, previous document image restoring systems and the recently proposed black-and-white document image deblurring method [1] are limited, and cannot handle large motion blurs and complex background. We propose a novel text image deblurring method which takes into account the specific properties of text images. Our method extends the commonly used optimization framework for image deblurring to allow domain-specific properties to be incorporated in the optimization process. Experimental results show that our method can generate higher quality deblurring results on text images than previous approaches.

[S6-P11B]

Sequential Spectral Learning to Hash with Multiple Representations

Saehoon Kim, Yoonseop Kang, and Seungjin Choi

Learning to hash involves learning hash functions from a set of images for embedding high-dimensional visual descriptors into a similarity-preserving low-dimensional Hamming space. Most of existing methods resort to a single representation of images, that is, only one type of visual descriptors is used to learn a hash function to assign binary codes to images. However, images are often described by multiple different visual descriptors (such as SIFT, GIST, HOG), so it is desirable to incorporate these multiple representations into learning a hash function, leading to multi-view hashing. In this paper we present a sequential spectral learning approach to multi-view hashing where a hash function is sequentially determined by solving the successive maximization of local variances subject to decorrelation constraints. We compute multi-view local variances by α-averaging view-specific distance matrices such that the best averaged distance matrix is determined by minimizing its α-divergence from view-specific distance matrices. We also present a scalable implementation, exploiting a fast approximate k-NN graph construction method, in which α-averaged distances computed in small partitions determined by recursive spectral bisection are gradually merged in conquer steps until whole examples are used. Numerical experiments on Caltech-256, CIFAR-20, and NUS-WIDE datasets confirm the high performance of our method, in comparison to single-view spectral hashing as well as existing multi-view hashing methods.

[S6-P12B]

Two-Granularity Tracking: Mediating Trajectory and Detection Graphs for Tracking under Occlusions

Katerina Fragkiadaki, Weiyu Zhang, Geng Zhang, and Jianbo Shi

We propose a tracking framework that mediates grouping cues from two levels of tracking granularities, detection tracklets and point trajectories, for segmenting objects in crowded scenes. Detection tracklets capture objects when they are mostly visible. They may be sparse in time, may miss partially occluded or deformed objects, or contain false positives. Point trajectories are dense in space and time. Their affinities integrate long range motion and 3D disparity information, useful for segmentation. Affinities may leak though across similarly moving objects, since they lack model knowledge. We establish one trajectory and one detection tracklet graph, encoding grouping affinities in each space and associations across. Two-granularity tracking is cast as simultaneous detection tracklet classification and clustering (cl^2) in the joint space of tracklets and trajectories. We solve cl^2 by explicitly mediating contradictory affinities in the two graphs: Detection tracklet classification modifies trajectory affinities to reflect object specific dis-associations. Non-accidental grouping alignment between detection tracklets and trajectory clusters boosts or rejects corresponding detection tracklets, changing accordingly their classification.We show our model can track objects through sparse, inaccurate detections and persistent partial occlusions. It adapts to the changing visibility masks of the targets, in contrast to detection based bounding box trackers, by effectively switching between the two granularities according to object occlusions, deformations and background clutter.

[S6-P13B]

Taking Mobile Multi-object Tracking to the Next Level: People, Unknown Objects, and Carried Items

Dennis Mitzel and Bastian Leibe

In this paper, we aim to take mobile multi-object tracking to the next level. Current approaches work in a tracking-by-detection framework, which limits them to object categories for which pre-trained detector models are available. In contrast, we propose a novel tracking-before-detection approach that can track both known and unknown object categories in very challenging street scenes. Our approach relies on noisy stereo depth data in order to segment and track objects in 3D. At its core is a novel, compact 3D representation that allows us to robustly track a large variety of objects, while building up models of their 3D shape online. In addition to improving tracking performance, this representation allows us to detect anomalous shapes, such as carried items on a person's body. We evaluate our approach on several challenging video sequences of busy pedestrian zones and show that it outperforms state-of-the-art approaches.

[S6-P14B]

Dynamic Context for Tracking behind Occlusions

Fei Xiong, Octavia I. Camps, and Mario Sznaier

Tracking objects in the presence of clutter and occlusion remains a challenging problem. Current approaches often rely on a priori target dynamics and/or use nearly rigid image context to determine the target position. In this paper, a novel algorithm is proposed to estimate the location of a target while it is hidden due to occlusion. The main idea behind the algorithm is to use contextual dynamical cues from multiple supporter features which may move with the target, move independently of the target, or remain stationary. These dynamical cues are learned directly from the data without making prior assumptions about the motions of the target and/or the support features. As illustrated through several experiments, the proposed algorithm outperforms state of the art approaches under long occlusions and severe camera motion.

[S6-P15B]

To Track or To Detect? An Ensemble Framework for Optimal Selection

Xu Yan, Xuqing Wu, Ioannis A. Kakadiaris, and Shishir K. Shah

This paper presents a novel approach for multi-target tracking using an ensemble framework that optimally chooses target tracking results from that of independent trackers and a detector at each time step. The ensemble model is designed to select the best candidate scored by a function integrating detection confidence, appearance affinity, and smoothness constraints imposed using geometry and motion information. Parameters of our association score function are discriminatively trained with a max-margin framework. Optimal selection is achieved through a hierarchical data association step that progressively associates candidates to targets. By introducing a second target classifier and using the ranking score from the pre-trained classifier as the detection confidence measure, we add additional robustness against unreliable detections. The proposed algorithm robustly tracks a large number of moving objects in complex scenes with occlusions. We evaluate our approach on a variety of public datasets and show promising improvements over state-of-the-art methods.

[S6-P16B]

Spatial and Angular Variational Super-Resolution of 4D Light Fields

Sven Wanner and Bastian Goldluecke

We present a variational framework to generate super-resolved novel views from 4D light field data sampled at low resolution, for example by a plenoptic camera. In contrast to previous work, we formulate the problem of view synthesis as a continuous inverse problem, which allows us to correctly take into account foreshortening effects caused by scene geometry transformations. High-accuracy depth maps for the input views are locally estimated using epipolar plane image analysis, which yields floating point depth precision without the need for expensive matching cost minimization. The disparity maps are further improved by increasing angular resolution with synthesized intermediate views. Minimization of the super-resolution model energy is performed with state of the art convex optimization algorithms within seconds.

[S6-P17B]

Blur-Kernel Estimation from Spectral Irregularities

Amit Goldstein and Raanan Fattal

We describe a new method for recovering the blur kernel in motion-blurred images based on statistical irregularities their power spectrum exhibits. This is achieved by a power-law that refines the one traditionally used for describing natural images. The new model better accounts for biases arising from the presence of large and strong edges in the image. We use this model together with an accurate spectral whitening formula to estimate the power spectrum of the blur. The blur kernel is then recovered using a phase retrieval algorithm with improved convergence and disambiguation capabilities. Unlike many existing methods, the new approach does not perform a maximum a posteriori estimation, which involves repeated reconstructions of the latent image, and hence offers attractive running times. We compare the new method with state-of-the-art methods and report various advantages, both in terms of efficiency and accuracy.

[S6-P18B]

Deconvolving PSFs for a Better Motion Deblurring Using Multiple Images

Xiang Zhu, Filip Šroubek, and Peyman Milanfar

Blind deconvolution of motion blur is hard, but it can be made easier if multiple images are available. This observation, and an algorithm using two differently-blurred images of a scene are the subject of this paper. While this idea is not new, existing methods have so far not delivered very practical results. In practice, the PSFs corresponding to the two given images are estimated and assumed to be close to the latent motion blurs. But in actual fact, these estimated blurs are often far from the truth, for a simple reason: They often share a common, and unidentified PSF that goes unaccounted for. That is, the estimated PSFs are themselves "blurry". While this can be due to any number of other blur sources including shallow depth of field, out of focus, lens aberrations, diffraction effects, and the like, it is also a mathematical artifact of the ill-posedness of the deconvolution problem. In this paper, instead of estimating the PSFs directly and only once from the observed images, we first generate a rough estimate of the PSFs using a robust multichannel deconvolution algorithm, and then "deconvolve the PSFs" to refine the outputs. Simulated and real data experiments show that this strategy works quite well for motion blurred images, producing state of the art results.

[S6-P19B]

Depth and Deblurring from a Spectrally-Varying Depth-of-Field

Ayan Chakrabarti and Todd Zickler

We propose modifying the aperture of a conventional color camera so that the effective aperture size for one color channel is smaller than that for the other two. This produces an image where different color channels have different depths-of-field, and from this we can computationally recover scene depth, reconstruct an all-focus image and achieve synthetic re-focusing, all from a single shot. These capabilities are enabled by a spatio-spectral image model that encodes the statistical relationship between gradient profiles across color channels. This approach substantially improves depth accuracy over alternative single-shot coded-aperture designs, and since it avoids introducing additional spatial distortions and is light efficient, it allows high-quality deblurring and lower exposure times. We demonstrate these benefits with comparisons on synthetic data, as well as results on images captured with a prototype lens.

[S6-P20B]

Segmentation over Detection by Coupled Global and Local Sparse Representations

Wei Xia, Zheng Song, Jiashi Feng, Loong-Fah Cheong, and Shuicheng Yan

Motivated by the rising performances of object detection algorithms, we investigate how to further precisely segment out objects within the output bounding boxes. The task is formulated as a unified optimization problem, pursuing a unique latent object mask in non-parametric manner. For a given test image, the objects are first detected by detectors. Then for each detected bounding box, the objects of the same category along with their object masks are extracted from the training set. The latent mask of the object within the bounding box is inferred based on three objectives: 1) the latent mask should be coherent, subject to sparse errors caused by within-category diversities, with the global bounding-box-level mask inferred by sparse representation over the bounding boxes of the same category within the training set; 2) the latent mask should be coherent with local patch-level mask inferred by sparse representation of the individual patch over all spatially nearby (handling local deformations) patches of the same category in the training set; and 3) mask property within each sufficiently small super-pixel should be consistent. All these three objectives are integrated into a unified optimization problem, and finally the sparse representation coefficients and the latent mask are alternately optimized based on Lasso optimization and smooth approximation followed by Accelerated Proximal Gradient method, respectively. Extensive experiments on the Pascal VOC object segmentation datasets, VOC2007 and VOC2010, show that our proposed algorithm achieves competitive results with the state-of-the-art learning based algorithms, and is superior over other detection based object segmentation algorithms.

[S6-P21B]

Moving Object Segmentation Using Motor Signals

Changhai Xu, Jingen Liu, and Benjamin Kuipers

Moving object segmentation from an image sequence is essential for a robot to interact with its environment. Traditional vision approaches appeal to pure motion analysis on videos without exploiting the source of the background motion. We observe, however, that the background motion (from the robot's egocentric view) has stronger correlation to the robot's motor signals than the foreground motion. We propose a novel approach to detecting moving objects by clustering features into background and foreground according to their motion consistency with motor signals. Specifically, our approach learns homography and fundamental matrices as functions of motor signals, and predict sparse feature locations from the learned matrices. The errors between the predictions and their actual tracked locations are used to label them into background and foreground. The labels of the sparse features are then propagated to all pixels. Our approach does not require building a dense mosaic background or searching for affine, homography, or fundamental matrix parameters for foreground separation. In addition, it does not need to explicitly model the intrinsic and extrinsic calibration parameters hence requires much less prior geometry knowledge. It works completely in 2D image space, and does not involve any complex analysis or computation in 3D space.

[S6-P22B]

Block-Sparse RPCA for Consistent Foreground Detection

Zhi Gao, Loong-Fah Cheong, and Mo Shan

Recent evaluation of representative background subtraction techniques demonstrated the drawbacks of these methods, with hardly any approach being able to reach more than 50% precision at recall level higher than 90%. Challenges in realistic environment include illumination change causing complex intensity variation, background motions (trees, waves, etc.) whose magnitude can be greater than the foreground, poor image quality under low light, camouflage etc. Existing methods often handle only part of these challenges; we address all these challenges in a unified framework which makes little specific assumption of the background. We regard the observed image sequence as being made up of the sum of a low-rank background matrix and a sparse outlier matrix and solve the decomposition using the Robust Principal Component Analysis method. We dynamically estimate the support of the foreground regions via a motion saliency estimation step, so as to impose spatial coherence on these regions. Unlike smoothness constraint such as MRF, our method is able to obtain crisply defined foreground regions, and in general, handles large dynamic background motion much better. Extensive experiments on benchmark and additional challenging datasets demonstrate that our method significantly outperforms the state-of-the-art approaches and works effectively on a wide range of complex scenarios.

[S6-P23B]

A Generative Model for Simultaneous Estimation of Human Body Shape and Pixel-Level Segmentation

Ingmar Rauschert and Robert T. Collins

This paper addresses pixel-level segmentation of a human body from a single image. The problem is formulated as a multi-region segmentation where the human body is constrained to be a collection of geometrically linked regions and the background is split into a small number of distinct zones. We solve this problem in a Bayesian framework for jointly estimating articulated body pose and the pixel-level segmentation of each body part. Using an image likelihood function that simultaneously generates and evaluates the image segmentation corresponding to a given pose, we robustly explore the posterior body shape distribution using a data-driven, coarse-to-fine Metropolis Hastings sampling scheme that includes a strongly data-driven proposal term.

[S6-P24B]

Visual Dictionary Learning for Joint Object Categorization and Segmentation

Aastha Jain, Luca Zappella, Patrick McClure, and René Vidal

Representing objects using elements from a visual dictionary is widely used in object detection and categorization. Prior work on dictionary learning has shown improvements in the accuracy of object detection and categorization by learning discriminative dictionaries. However none of these dictionaries are learnt for joint object categorization and segmentation. Moreover, dictionary learning is often done separately from classifier training, which reduces the discriminative power of the model. In this paper, we formulate the semantic segmentation problem as a joint categorization, segmentation and dictionary learning problem. To that end, we propose a latent conditional random field (CRF) model in which the observed variables are pixel category labels and the latent variables are visual word assignments. The CRF energy consists of a bottom-up segmentation cost, a top-down bag of (latent) words categorization cost, and a dictionary learning cost. Together, these costs capture relationships between image features and visual words, relationships between visual words and object categories, and spatial relationships among visual words. The segmentation, categorization, and dictionary learning parameters are learnt jointly using latent structural SVMs, and the segmentation and visual words assignments are inferred jointly using energy minimization techniques. Experiments on the Graz02 and CamVid datasets demonstrate the performance of our approach.

ORAL SESSION 6 GEOMETRY AND RECOGNITION

Wednesday, October 10
17:05 - 18:30

[S6-O1]

People Watching: Human Actions as a Cue for Single View Geometry

David F. Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A. Efros, Ivan Laptev, and Josef Sivic

We present an approach which exploits the coupling between human actions and scene geometry. We investigate the use of human pose as a cue for single-view 3D scene understanding. Our method builds upon recent advances in still-image pose estimation to extract functional and geometric constraints about the scene. These constraints are then used to improve state-of-the-art single-view 3D scene understanding approaches. The proposed method is validated on a collection of monocular time-lapse sequences collected from YouTube and a dataset of still images of indoor scenes. We demonstrate that observing people performing different actions can significantly improve estimates of 3D scene geometry.

[S6-O2]

Indoor Segmentation and Support Inference from RGBD Images

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus

We present an approach to interpret the major surfaces, objects, and support relations of an indoor scene from an RGBD image. Most existing work ignores physical interactions or is applied only to tidy rooms and hallways. Our goal is to parse typical, often messy, indoor scenes into floor, walls, supporting surfaces, and object regions, and to recover support relationships. One of our main interests is to better understand how 3D cues can best inform a structured 3D interpretation. We also contribute a novel integer programming formulation to infer physical support relations. We offer a new dataset of 1449 RGBD images, capturing 464 diverse indoor scenes, with detailed annotations. Our experiments demonstrate our ability to infer support relations in complex scenes and verify that our 3D scene cues and inferred support lead to better object segmentation.

[S6-O3]

Beyond the Line of Sight: Labeling the Underlying Surfaces

Ruiqi Guo and Derek Hoiem

Scene understanding requires reasoning about both what we can see and what is occluded. We offer a simple and general approach to infer labels of occluded background regions. Our approach incorporates estimates of visible surrounding background, detected objects, and shape priors from transferred training regions. We demonstrate the ability to infer the labels of occluded background regions in both the outdoor StreetScenes dataset and an indoor scene dataset using the same approach. Our experiments show that our method outperforms competent baselines.

[S6-O4]

Depth Extraction from Video Using Non-parametric Sampling

Kevin Karsch, Ce Liu, and Sing Bing Kang

We describe a technique that automatically generates plausible depth maps from videos using non-parametric depth sampling. We demonstrate our technique in cases where past methods fail (non-translating cameras and dynamic scenes). Our technique is applicable to single images as well as videos. For videos, we use local motion cues to improve the inferred depth maps, while optical flow is used to ensure temporal depth consistency. For training and evaluation, we use a Kinect-based system to collect a large dataset containing stereoscopic videos with known depths. We show that our depth estimation technique outperforms the state-of-the-art on benchmark databases. Our technique can be used to automatically convert a monoscopic video into stereo for 3D visualization, and we demonstrate this through a variety of visually pleasing results for indoor and outdoor scenes, including results from the feature film Charade.

[S6-O5]

Multiple View Object Cosegmentation Using Appearance and Stereo Cues

Adarsh Kowdle, Sudipta N. Sinha, and Richard Szeliski

We present an automatic approach to segment an object in calibrated images acquired from multiple viewpoints. Our system starts with a new piecewise planar layer-based stereo algorithm that estimates a dense depth map that consists of a set of 3D planar surfaces. The algorithm is formulated using an energy minimization framework that combines stereo and appearance cues, where for each surface, an appearance model is learnt using an unsupervised approach. By treating the planar surfaces as structural elements of the scene and reasoning about their visibility in multiple views, we segment the object in each image independently. Finally, these segmentations are refined by probabilistically fusing information across multiple views. We demonstrate that our approach can segment challenging objects with complex shapes and topologies, which may have thin structures and non-Lambertian surfaces. It can also handle scenarios where the object and background color distributions overlap significantly.

POSTER SESSION 7

Thursday, October 11
08:45 - 11:15

[S7-P1A]

Manifold Statistics for Essential Matrices

Gijs Dubbelman, Leo Dorst, and Henk Pijls

Riemannian geometry allows for the generalization of statistics designed for Euclidean vector spaces to Riemannian manifolds. It has recently gained popularity within computer vision as many relevant parameter spaces have such a Riemannian manifold structure. Approaches which exploit this have been shown to exhibit improved efficiency and accuracy. The Riemannian logarithmic and exponential mappings are at the core of these approaches. In this contribution we review recently proposed Riemannian mappings for essential matrices and prove that they lead to sub-optimal manifold statistics. We introduce correct Riemannian mappings by utilizing a multiple-geodesic approach and show experimentally that they provide optimal statistics.

[S7-P2A]

Elastic Shape Matching of Parameterized Surfaces Using Square Root Normal Fields

Ian H. Jermyn, Sebastian Kurtek, Eric Klassen, and Anuj Srivastava

In this paper we define a new methodology for shape analysis of parameterized surfaces, where the main issues are: (1) choice of metric for shape comparisons and (2) invariance to reparameterization. We begin by defining a general elastic metric on the space of parameterized surfaces. The main advantages of this metric are twofold. First, it provides a natural interpretation of elastic shape deformations that are being quantified. Second, this metric is invariant under the action of the reparameterization group. We also introduce a novel representation of surfaces termed square root normal fields or SRNFs. This representation is convenient for shape analysis because, under this representation, a reduced version of the general elastic metric becomes the simple L^2 metric. Thus, this transformation greatly simplifies the implementation of our framework. We validate our approach using multiple shape analysis examples for quadrilateral and spherical surfaces. We also compare the current results with those of Kurtek et al. [1]. We show that the proposed method results in more natural shape matchings, and furthermore, has some theoretical advantages over previous methods.

[S7-P3A]

N-tuple Color Segmentation for Multi-view Silhouette Extraction

Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, François Le Clerc, and Patrick Pérez

We present a new method to extract multiple segmentations of an object viewed by multiple cameras, given only the camera calibration. We introduce the n-tuple color model to express inter-view consistency when inferring in each view the foreground and background color models permitting the final segmentation. A color n-tuple is a set of pixel colors associated to the n projections of a 3D point. The first goal is set as finding the MAP estimate of background/foreground color models based on an arbitrary sample set of such n-tuples, such that samples are consistently classified, in a soft way, as ″empty″ if they project in the background of at least one view, or ″occupied″ if they project to foreground pixels in all views. An Expectation Maximization framework is then used to alternate between color models and soft classifications. In a final step, all views are segmented based on their attached color models. The approach is significantly simpler and faster than previous multi-view segmentation methods, while providing results of equivalent or better quality.

[S7-P4A]

Motion-Aware Structured Light Using Spatio-Temporal Decodable Patterns

Yuichi Taguchi, Amit Agrawal, and Oncel Tuzel

Single-shot structured light methods allow 3D reconstruction of dynamic scenes. However, such methods lose spatial resolution and perform poorly around depth discontinuities. Previous single-shot methods project the same pattern repeatedly; thereby spatial resolution is reduced even if the scene is static or has slowly moving parts. We present a structured light system using a sequence of shifted stripe patterns that is decodable both spatially and temporally. By default, our method allows single-shot 3D reconstruction with any of our projected patterns by using spatial windows. Moreover, the sequence is designed so as to progressively improve the reconstruction quality around depth discontinuities by using temporal windows. Our method enables motion-aware reconstruction for each pixel: The best spatio-temporal window is automatically selected depending on the scene structure, motion, and the number of available images. This significantly reduces the number of pixels around discontinuities where depth cannot be recovered in traditional approaches. Our decoding scheme extends the adaptive window matching commonly used in stereo by incorporating temporal windows with 1D spatial windows. We demonstrate the advantages of our approach for a variety of scenarios including thin structures, dynamic scenes, and scenes containing both static and dynamic regions.

[S7-P5A]

Refractive Calibration of Underwater Cameras

Anne Jordt-Sedlazeck and Reinhard Koch

In underwater computer vision, images are influenced by the water in two different ways. First, while still traveling through the water, light is absorbed and scattered, both of which are wavelength dependent, thus create the typical green or blue hue in underwater images. Secondly, when entering the underwater housing, the rays are refracted, affecting image formation geometrically. When using underwater images in for example Structure-from-Motion applications, both effects need to be taken into account. Therefore, we present a novel method for calibrating the parameters of an underwater camera housing. An evolutionary optimization algorithm is coupled with an analysis-by-synthesis approach, which allows to calibrate the parameters of a light propagation model for the local water body. This leads to a highly accurate calibration method for camera-glass distance and glass normal with respect to the optical axis. In addition, a model for the distance dependent effect of water on light propagation is parametrized and can be used for color correction.

[S7-P6A]

Detection of Independently Moving Objects in Non-planar Scenes via Multi-Frame Monocular Epipolar Constraint

Soumyabrata Dey, Vladimir Reilly, Imran Saleemi, and Mubarak Shah

In this paper we present a novel approach for detection of independently moving foreground objects in non-planar scenes captured by a moving camera. We avoid the traditional assumptions that the stationary background of the scene is planar, or that it can be approximated by dominant single or multiple planes, or that the camera used to capture the video is orthographic. Instead we utilize a multiframe monocular epipolar constraint of camera motion derived for monocular moving cameras defined by an evolving epipolar plane between the moving camera center and 3D scene points. This constraint is parameterized as a polynomial function of time, and unlike repeated computations of inter-frame fundamental matrix, requires the estimation of fewer unknowns, and provides a more consistent separation between moving and static objects for different levels of noise. This constraint allows us to segment out moving objects in a general 3D scene where other approaches fail because their initial assumptions do not hold, and provides a natural way of fusing temporal information across multiple frames. We use a combination of optical flow and particle advection to capture all motion in the video across a number of frames, in the form of particle trajectories. We then apply the derived multi-frame epipolar constraint to these trajectories to determine which trajectories violate it, thus segmenting out the independently moving objects. We show superior results on a number of moving camera sequences observing non-planar scenes, where other methods fail.

[S7-P7A]

Shape from Angle Regularity

Aamer Zaheer, Maheen Rashid, and Sohaib Khan

This paper deals with automatic Single View Reconstruction (SVR) of multi-planar scenes characterized by a profusion of straight lines and mutually orthogonal line-pairs. We provide a new shape-from-X constraint based on this regularity of angles between line-pairs in man-made scenes. First, we show how the presence of such regular angles can be used for 2D rectification of an image of a plane. Further, we propose an automatic SVR method assuming there are enough orthogonal line-pairs available on each plane. This angle regularity is only imposed on physically intersecting line-pairs, making it a local constraint. Unlike earlier literature, our approach does not make restrictive assumptions about the orientation of the planes or the camera and works for both indoor and outdoor scenes. Results are shown on challenging images which would be difficult to reconstruct for existing automatic SVR algorithms.

[S7-P8A]

Pose Invariant Approach for Face Recognition at Distance

Eslam Mostafa, Asem Ali, Naif Alajlan, and Aly Farag

We propose an automatic pose invariant approach for Face Recognition At a Distance (FRAD). Since face alignment is a crucial step in face recognition systems, we propose a novel facial features extraction model, which guides extended ASM to accurately align the face. Our main concern is to recognize human faces under uncontrolled environment at far distances accurately and fast. To achieve this goal, we perform an offline stage where 3D faces are reconstructed from stereo pair images. These 3D shapes are used to synthesize virtual 2D views in novel poses. To obtain good synthesized images from the 3D shape, we propose an accurate 3D reconstruction framework, which carefully handles illumination variance, occlusion, and the disparity discontinuity. The online phase is fast where a 2D image with unknown pose is matched with the closest virtual images in sampled poses. Experiments show that our approach outperforms the-state-of-the-art approaches.

[S7-P9A]

Minimal Correlation Classification

Noga Levy and Lior Wolf

When the description of the visual data is rich and consists of many features, a classification based on a single model can often be enhanced using an ensemble of models. We suggest a new ensemble learning method that encourages the base classifiers to learn different aspects of the data. Initially, a binary classification algorithm such as Support Vector Machine is applied and its confidence values on the training set are considered. Following the idea that ensemble methods work best when the classification errors of the base classifiers are not related, we serially learn additional classifiers whose output confidences on the training examples are minimally correlated. Finally, these uncorrelated classifiers are assembled using the GentleBoost algorithm. Presented experiments in various visual recognition domains demonstrate the effectiveness of the method.

[S7-P10A]

Contextual Object Detection Using Set-Based Classification

Ramazan Gokberk Cinbis and Stan Sclaroff

We propose a new model for object detection that is based on set representations of the contextual elements. In this formulation, relative spatial locations and relative scores between pairs of detections are considered as sets of unordered items. Directly training classification models on sets of unordered items, where each set can have varying cardinality can be difficult. In order to overcome this problem, we propose SetBoost, a discriminative learning algorithm for building set classifiers. The SetBoost classifiers are trained to rescore detected objects based on object-object and object-scene context. Our method is able to discover composite relationships, as well as intra-class and inter-class spatial relationships between objects. The experimental evidence shows that our set-based formulation performs comparable to or better than existing contextual methods on the SUN and the VOC 2007 benchmark datasets.

[S7-P11A]

Age Invariant Face Verification with Relative Craniofacial Growth Model

Tao Wu and Rama Chellappa

Age-separated facial images usually have significant changes in both shape and texture. Although many face recognition algorithms have been proposed in the last two decades, the problem of recognizing facial images across aging remains an open problem. In this paper, we propose a relative craniofacial growth model which is based on the science of craniofacial anthropometry. Compared to the traditional craniofacial growth model, the proposed method introduces a set of linear equations on the relative growth parameters which can be easily applied for facial image verification across aging. We then integrate the relative growth model with the Grassmann manifold and the SVM classifier. We also demonstrate how knowing the age could improve shape-based face recognition algorithms. Experiments show that the proposed method is able to mitigate the variations caused by the aging progress and thus effectively improve the performance of open-set face verification across aging.

[S7-P12A]

Inferring Gene Interaction Networks from ISH Images via Kernelized Graphical Models

Kriti Puniyani and Eric P. Xing

New bio-technologies are being developed that allow high-throughput imaging of gene expressions, where each image captures the spatial gene expression pattern of a single gene in the tissue of interest. This paper addresses the problem of automatically inferring a gene interaction network from such images. We propose a novel kernel-based graphical model learning algorithm, that is both convex and consistent. The algorithm uses multi-instance kernels to compute similarity between the expression patterns of different genes, and then minimizes the L1 regularized Bregman divergence to estimate a sparse gene interaction network. We apply our algorithm on a large, publicly available data set of gene expression images of Drosophila embryos, where our algorithm makes novel and interesting predictions.

[S7-P13A]

Random Forest for Image Annotation

Hao Fu, Qian Zhang, and Guoping Qiu

In this paper, we present a novel method for image annotation and made three contributions. Firstly, we propose to use the tags contained in the training images as the supervising information to guide the generation of random trees, thus enabling the retrieved nearest neighbor images not only visually alike but also semantically related. Secondly, different from conventional decision tree methods, which fuse the information contained at each leaf node individually, our method treats the random forest as a whole, and introduces the new concepts of semantic nearest neighbors (SNN) and semantic similarity measure (SSM). Thirdly, we annotate an image from the tags of its SNN based on SSM and have developed a novel learning to rank algorithm to systematically assign the optimal tags to the image. The new technique is intrinsically scalable and we will present experimental results to demonstrate that it is competitive to state of the art methods.

[S7-P14A]

$(MP)^2T$: Multiple People Multiple Parts Tracker

Hamid Izadinia, Imran Saleemi, Wenhui Li, and Mubarak Shah

We present a method for multi-target tracking that exploits the persistence in detection of object parts. While the implicit representation and detection of body parts have recently been leveraged for improved human detection, ours is the first method that attempts to temporally constrain the location of human body parts with the express purpose of improving pedestrian tracking. We pose the problem of simultaneous tracking of multiple targets and their parts in a network flow optimization framework and show that parts of this network need to be optimized separately and iteratively, due to inter-dependencies of node and edge costs. Given potential detections of humans and their parts separately, an initial set of pedestrian tracklets is first obtained, followed by explicit tracking of human parts as constrained by initial human tracking. A merging step is then performed whereby we attempt to include part-only detections for which the entire human is not observable. This step employs a selective appearance model, which allows us to skip occluded parts in description of positive training samples. The result is high confidence, robust trajectories of pedestrians as well as their parts, which essentially constrain each other's locations and associations, thus improving human tracking and parts detection. We test our algorithm on multiple real datasets and show that the proposed algorithm is an improvement over the state-of-the-art.

[S7-P15A]

Mixture Component Identification and Learning for Visual Recognition

Omid Aghazadeh, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson

The non-linear decision boundary between object and background classes - due to large intra-class variations - needs to be modelled by any classifier wishing to achieve good results. While a mixture of linear classifiers is capable of modelling this non-linearity, learning this mixture from weakly annotated data is non-trivial and is the paper's focus. Our approach is to identify the modes in the distribution of our positive examples by clustering, and to utilize this clustering in a latent SVM formulation to learn the mixture model. The clustering relies on a robust measure of visual similarity which suppresses uninformative clutter by using a novel representation based on the exemplar SVM. This subtle clustering of the data leads to learning better mixture models, as is demonstrated via extensive evaluations on Pascal VOC 2007. The final classifier, using a HOG representation of the global image patch, achieves performance comparable to the state-of-the-art while being more efficient at detection time.

[S7-P16A]

Image Retrieval with Structured Object Queries Using Latent Ranking SVM

Tian Lan, Weilong Yang, Yang Wang, and Greg Mori

We consider image retrieval with structured object queries – queries that specify the objects that should be present in the scene, and their spatial relations. An example of such queries is "car on the road". Existing image retrieval systems typically consider queries consisting of object classes (i.e. keywords). They train a separate classifier for each object class and combine the output heuristically. In contrast, we develop a learning framework to jointly consider object classes and their relations. Our method considers not only the objects in the query ("car" and "road" in the above example), but also related object categories can be useful for retrieval. Since we do not have ground-truth labeling of object bounding boxes on the test image, we represent them as latent variables in our model. Our learning method is an extension of the ranking SVM with latent variables, which we call latent ranking SVM. We demonstrate image retrieval and ranking results on a dataset with more than a hundred of object classes.

[S7-P17A]

A Probabilistic Derivative Measure Based on the Distribution of Intensity Difference

Youngbae Hwang and In-So Kweon

In this paper, we propose a novel derivative measure based on the probability of intensity difference that is defined by observed intensities and their true intensities. Because the true intensity cannot be estimated accurately only using two observed intensities, the probability is marginalized to consider an entire set of possible true values. The proposed measure not only considers intensity dependent noise effectively using a distribution of intensity difference, but also computes the relevant difference of two corresponding pixels that have different true intensities by extending the same intensity assumption in previous works. Using the proposed measure, the estimation result of image derivative for synthetic noisy signals is closer to the ground truth than most of previous measures. We apply the proposed measure for block matching and corner detection that compute intensity similarity in the temporal domain and image derivative in the spatial domain, respectively.

[S7-P18A]

Pairwise Rotation Invariant Co-occurrence Local Binary Pattern

Xianbiao Qi, Rong Xiao, Jun Guo, and Lei Zhang

In this work, we introduce a novel pairwise rotation invariant co-occurrence local binary pattern (PRI-CoLBP) feature which incorporates two types of context - spatial co-occurrence and orientation co-occurrence. Different from traditional rotation invariant local features, pairwise rotation invariant co-occurrence features preserve relative angle between the orientations of individual features. The relative angle depicts the local curvature information, which is discriminative and rotation invariant. Experimental results on the CUReT, Brodatz, KTH-TIPS texture dataset, Flickr Material dataset, and Oxford 102 Flower dataset further demonstrate the superior performance of the proposed feature on texture classification, material recognition and flower recognition tasks.

[S7-P19A]

Per-patch Descriptor Selection Using Surface and Scene Properties

Ivo Everts, Jan C. van Gemert, and Theo Gevers

Local image descriptors are generally designed for describing all possible image patches. Such patches may be subject to complex variations in appearance due to incidental object, scene and recording conditions. Because of this, a single-best descriptor for accurate image representation under all conditions does not exist. Therefore, we propose to automatically select from a pool of descriptors the one that is best suitable based on object surface and scene properties. These properties are measured on the fly from a single image patch through a set of attributes. Attributes are input to a classifier which selects the best descriptor. Our experiments on a large dataset of colored object patches show that the proposed selection method outperforms the best single descriptor and a-priori combinations of the descriptor pool.

[S7-P20A]

Mixed-Resolution Patch-Matching

Harshit Sureka and P.J. Narayanan

Matching patches of a source image with patches of itself or a target image is a first step for many operations. Finding the optimum nearest-neighbors of each patch using a global search of the image is expensive. Optimality is often sacrificed for speed as a result. We present the Mixed-Resolution Patch-Matching (MRPM) algorithm that uses a pyramid representation to perform fast global search. We compare mixed-resolution patches at coarser pyramid levels to alleviate the effects of smoothing. We store more matches at coarser resolutions to ensure wider search ranges and better accuracy at finer levels. Our method achieves near optimality in terms of average error compared to exhaustive search. Our approach is simple compared to complex trees or hash tables used by others. This enables fast parallel implementations on the GPU, yielding upto 70× speedup compared to other iterative approaches. Our approach is best suited when multiple, global matches are needed.

[S7-P21A]

Exploiting Sparse Representations for Robust Analysis of Noisy Complex Video Scenes

Gloria Zen, Elisa Ricci, and Nicu Sebe

Recent works have shown that, even with simple low level visual cues, complex behaviors can be extracted automatically from crowded scenes, e.g. those depicting public spaces recorded from video surveillance cameras. However, low level features as optical flow or foreground pixels are inherently noisy. In this paper we propose a novel unsupervised learning approach for the analysis of complex scenes which is specifically tailored to cope directly with features' noise and uncertainty. We formalize the task of extracting activity patterns as a matrix factorization problem, considering as reconstruction function the robust Earth Mover's Distance. A constraint of sparsity on the computed basis matrix is imposed, filtering out noise and leading to the identification of the most relevant elementary activities in a typical high level behavior. We further derive an alternate optimization approach to solve the proposed problem efficiently and we show that it is reduced to a sequence of linear programs. Finally, we propose to use short trajectory snippets to account for object motion information, in alternative to the noisy optical flow vectors used in previous works. Experimental results demonstrate that our method yields similar or superior performance to state-of-the arts approaches.

[S7-P22A]

KAZE Features

Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J. Davison

In this paper, we introduce KAZE features, a novel multiscale 2D feature detection and description algorithm in nonlinear scale spaces. Previous approaches detect and describe features at different scale levels by building or approximating the Gaussian scale space of an image. However, Gaussian blurring does not respect the natural boundaries of objects and smoothes to the same degree both details and noise, reducing localization accuracy and distinctiveness. In contrast, we detect and describe 2D features in a nonlinear scale space by means of nonlinear diffusion filtering. In this way, we can make blurring locally adaptive to the image data, reducing noise but retaining object boundaries, obtaining superior localization accuracy and distinctiviness. The nonlinear scale space is built using efficient Additive Operator Splitting (AOS) techniques and variable conductance diffusion. We present an extensive evaluation on benchmark datasets and a practical matching application on deformable surfaces. Even though our features are somewhat more expensive to compute than SURF due to the construction of the nonlinear scale space, but comparable to SIFT, our results reveal a step forward in performance both in detection and description against previous state-of-the-art methods.

[S7-P1B]

Online Moving Camera Background Subtraction

Ali Elqursh and Ahmed Elgammal

Recently several methods for background subtraction from moving camera were proposed. They use bottom up cues to segment video frames into foreground and background regions. Due to this lack of explicit models, they can easily fail to detect a foreground object when such cues are ambiguous in certain parts of the video. This becomes even more challenging when videos need to be processed online. We present a method which enables learning of pixel based models for foreground and background regions and, in addition, segments each frame in an online framework. The method uses long term trajectories along with a Bayesian filtering framework to estimate motion and appearance models. We compare our method to previous approaches and show results on challenging video sequences.

[S7-P2B]

Coregistration: Simultaneous Alignment and Modeling of Articulated 3D Shape

David A. Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black

Three-dimensional (3D) shape models are powerful because they enable the inference of object shape from incomplete, noisy, or ambiguous 2D or 3D data. For example, realistic parameterized 3D human body models have been used to infer the shape and pose of people from images. To train such models, a corpus of 3D body scans is typically brought into registration by aligning a common 3D human-shaped template to each scan. This is an ill-posed problem that typically involves solving an optimization problem with regularization terms that penalize implausible deformations of the template. When aligning a corpus, however, we can do better than generic regularization. If we have a model of how the template can deform then alignments can be regularized by this model. Constructing a model of deformations, however, requires having a corpus that is already registered. We address this chicken-and-egg problem by approaching modeling and registration together. By minimizing a single objective function, we reliably obtain high quality registration of noisy, incomplete, laser scans, while simultaneously learning a highly realistic articulated body model. The model greatly improves robustness to noise and missing data. Since the model explains a corpus of body scans, it captures how body shape varies across people and poses.

[S7-P3B]

Motion Interchange Patterns for Action Recognition in Unconstrained Videos

Orit Kliper-Gross, Yaron Gurovich, Tal Hassner, and Lior Wolf

Action Recognition in videos is an active research field that is fueled by an acute need, spanning several application domains. Still, existing systems fall short of the applications' needs in real-world scenarios, where the quality of the video is less than optimal and the viewpoint is uncontrolled and often not static. In this paper, we consider the key elements of motion encoding and focus on capturing local changes in motion directions. In addition, we decouple image edges from motion edges using a suppression mechanism, and compensate for global camera motion by using an especially fitted registration scheme. Combined with a standard bag-of-words technique, our methods achieves state-of-the-art performance in the most recent and challenging benchmarks.

[S7-P4B]

A Non-parametric Hierarchical Model to Discover Behavior Dynamics from Tracks

Julian F.P. Kooij, Gwenn Englebienne, and Dariu M. Gavrila

We present a novel non-parametric Bayesian model to jointly discover the dynamics of low-level actions and high-level behaviors of tracked people in open environments. Our model represents behaviors as Markov chains of actions which capture high-level temporal dynamics. Actions may be shared by various behaviors and represent spatially localized occurrences of a person's low-level motion dynamics using Switching Linear Dynamics Systems. Since the model handles real-valued features directly, we do not lose information by quantizing measurements to 'visual words' and can thus discover variations in standing, walking and running without discrete thresholds. We describe inference using Gibbs sampling and validate our approach on several artificial and real-world tracking datasets. We show that our model can distinguish relevant behavior patterns that an existing state-of-the-art method for hierarchical clustering cannot.

[S7-P5B]

Scene Semantics from Long-Term Observation of People

Vincent Delaitre, David F. Fouhey, Ivan Laptev, Josef Sivic, Abhinav Gupta, and Alexei A. Efros

Our everyday objects support various tasks and can be used by people for different purposes. While object classification is a widely studied topic in computer vision, recognition of object function, i.e., what people can do with an object and how they do it, is rarely addressed. In this paper we construct a functional object description with the aim to recognize objects by the way people interact with them. We describe scene objects (sofas, tables, chairs) by associated human poses and object appearance. Our model is learned discriminatively from automatically estimated body poses in many realistic scenes. In particular, we make use of time-lapse videos from YouTube providing a rich source of common human-object interactions and minimizing the effort of manual object annotation. We show how the models learned from human observations significantly improve object recognition and enable prediction of characteristic human poses in new scenes. Results are shown on a dataset of more than 400,000 frames obtained from 146 time-lapse videos of challenging and realistic indoor scenes.

[S7-P6B]

Efficient Exact Inference for 3D Indoor Scene Understanding

Alexander G. Schwing and Raquel Urtasun

In this paper we propose the first exact solution to the problem of estimating the 3D room layout from a single image. This problem is typically formulated as inference in a Markov random field, where potentials count image features (e.g., geometric context, orientation maps, lines in accordance with vanishing points) in each face of the layout. We present a novel branch and bound approach which splits the label space in terms of candidate sets of 3D layouts, and efficiently bounds the potentials in these sets by restricting the contribution of each individual face. We employ integral geometry in order to evaluate these bounds in constant time, and as a consequence, we not only obtain the exact solution, but also in less time than approximate inference tools such as message-passing. We demonstrate the effectiveness of our approach in two benchmarks and show that our bounds are tight, and only a few evaluations are necessary.

[S7-P7B]

Seam Segment Carving: Retargeting Images to Irregularly-Shaped Image Domains

Shaoyu Qi and Jeffrey Ho

Image retargeting algorithms aim to adapt the image to the display screen with the goal of preserving the image content as much as possible. However, existing methods and research efforts have mostly been directed towards retargeting algorithms that retarget images to rectangular domains. This significantly hampers its application to broader classes of display devices and platforms for which the display area can be of any origins and shapes. For example, seam carving-based methods retarget images by carving out seams that run from the top to the bottom of the images, and this results in changing the width and therefore aspect ratio of the image without changing the shape of the image boundary in any essential way. However, by carving out appropriately-chosen seam segments, seams that are not required to cut across the entire image, it is then possible to retarget the images to a broader array of image domains with non-rectangular boundaries. Based on this simple idea of carving out the seam segments, the main contribution of this paper is a novel image retargeting algorithm that is capable of retargeting images to non-rectangular domains. We evaluate the effectiveness of the proposed method on a number of challenging indoor and outdoor scene images, and the results demonstrate that the proposed algorithm is both efficient and effective, and it is capable of providing good-quality retargeted images for a variety of interesting boundary shapes.

[S7-P8B]

Estimation of Intrinsic Image Sequences from Image+Depth Video

Kyong Joon Lee, Qi Zhao, Xin Tong, Minmin Gong, Shahram Izadi, Sang Uk Lee, Ping Tan, and Stephen Lin

We present a technique for estimating intrinsic images from image+depth video, such as that acquired from a Kinect camera. Intrinsic image decomposition in this context has importance in applications like object modeling, in which surface colors need to be recovered without illumination effects. The proposed method is based on two new types of decomposition constraints derived from the multiple viewpoints and reconstructed 3D scene geometry of the video data. The first type provides shading constraints that enforce relationships among the shading components of different surface points according to their similarity in surface orientation. The second type imposes temporal constraints that favor consistency in the intrinsic color of a surface point seen in different video frames, which improves decomposition in cases of view-dependent non-Lambertian reflections. Local and non-local variants of the two constraints are employed in a manner complementary to local and non-local reflectance constraints used in previous works. Together they are formulated within a linear system that allows for efficient optimization. Experimental results demonstrate that each of the new constraints appreciably elevates the quality of intrinsic image estimation, and that they jointly yield decompositions that compare favorably to current techniques.

[S7-P9B]

Bayesian Blind Deconvolution with General Sparse Image Priors

S. Derin Babacan, Rafael Molina, Minh N. Do, and Aggelos K. Katsaggelos

We present a general method for blind image deconvolution using Bayesian inference with super-Gaussian sparse image priors. We consider a large family of priors suitable for modeling natural images, and develop the general procedure for estimating the unknown image and the blur. Our formulation includes a number of existing modeling and inference methods as special cases while providing additional flexibility in image modeling and algorithm design. We also present an analysis of the proposed inference compared to other methods and discuss its advantages. Theoretical and experimental results demonstrate that the proposed formulation is very effective, efficient, and flexible.

[S7-P10B]

3D^{2}PM – 3D Deformable Part Models

Bojan Pepik, Peter Gehler, Michael Stark, and Bernt Schiele

As objects are inherently 3-dimensional, they have been modeled in 3D in the early days of computer vision. Due to the ambiguities arising from mapping 2D features to 3D models, 2D feature-based models are the predominant paradigm in object recognition today. While such models have shown competitive bounding box (BB) detection performance, they are clearly limited in their capability of fine-grained reasoning in 3D or continuous viewpoint estimation as required for advanced tasks such as 3D scene understanding. This work extends the deformable part model [1] to a 3D object model. It consists of multiple parts modeled in 3D and a continuous appearance model. As a result, the model generalizes beyond BB oriented object detection and can be jointly optimized in a discriminative fashion for object detection and viewpoint estimation. Our 3D Deformable Part Model (3D^{2}PM) leverages on CAD data of the object class, as a 3D geometry proxy.

[S7-P11B]

Efficient Similarity Derived from Kernel-Based Transition Probability

Takumi Kobayashi and Nobuyuki Otsu

Semi-supervised learning effectively integrates labeled and unlabeled samples for classification, and most of the methods are founded on the pair-wise similarities between the samples. In this paper, we propose methods to construct similarities from the probabilistic viewpoint, whilst the similarities have so far been formulated in a heuristic manner such as by k-NN. We first propose the kernel-based formulation of transition probabilities via considering kernel least squares in the probabilistic framework. The similarities are consequently derived from the kernel-based transition probabilities which are efficiently computed, and the similarities are inherently sparse without applying k-NN. In the case of multiple types of kernel functions, the multiple transition probabilities are also obtained correspondingly. From the probabilistic viewpoint, they can be integrated with prior probabilities, i.e., linear weights, and we propose a computationally efficient method to optimize the weights in a discriminative manner, as in multiple kernel learning. The novel similarity is thereby constructed by the composite transition probability and it benefits the semi-supervised learning methods as well. In the various experiments on semi-supervised learning problems, the proposed methods demonstrate favorable performances, compared to the other methods, in terms of classification performances and computation time.

[S7-P12B]

A Convex Discrete-Continuous Approach for Markov Random Fields

Christopher Zach and Pushmeet Kohli

We propose an extension of the well-known LP relaxation for Markov random fields to explicitly allow continuous label spaces. Unlike conventional continuous formulations of labelling problems which assume that the unary and pairwise potentials are convex, our formulation allows them to be general piecewise convex functions with continuous domains. Furthermore, we present the extension of the widely used efficient scheme for handling L^1 smoothness priors over discrete ordered label sets to continuous label spaces. We provide a theoretical analysis of the proposed model, and empirically demonstrate that labelling problems with huge or continuous label spaces can benefit from our discrete-continuous representation.

[S7-P13B]

Generalized Roof Duality for Multi-Label Optimization: Optimal Lower Bounds and Persistency

Thomas Windheuser, Hiroshi Ishikawa, and Daniel Cremers

We extend the concept of generalized roof duality from pseudo-boolean functions to real-valued functions over multi-label variables. In particular, we prove that an analogue of the persistency property holds for energies of any order with any number of linearly ordered labels. Moreover, we show how the optimal submodular relaxation can be constructed in the first-order case.

[S7-P14B]

Sparse Embedding: A Framework for Sparsity Promoting Dimensionality Reduction

Hien V. Nguyen, Vishal M. Patel, Nasser M. Nasrabadi, and Rama Chellappa

We introduce a novel framework, called sparse embedding (SE), for simultaneous dimensionality reduction and dictionary learning. We formulate an optimization problem for learning a transformation from the original signal domain to a lower-dimensional one in a way that preserves the sparse structure of data. We propose an efficient optimization algorithm and present its non-linear extension based on the kernel methods. One of the key features of our method is that it is computationally efficient as the learning is done in the lower-dimensional space and it discards the irrelevant part of the signal that derails the dictionary learning process. Various experiments show that our method is able to capture the meaningful structure of data and can perform significantly better than many competitive algorithms on signal recovery and object classification tasks.

[S7-P15B]

Automatic Localization of Balloon Markers and Guidewire in Rotational Fluoroscopy with Application to 3D Stent Reconstruction

Yu Wang, Terrence Chen, Peng Wang, Christopher Rohkohl, and Dorin Comaniciu

A fully automatic framework is proposed to identify consistent landmarks and wire structures in a rotational X-ray scan. In our application, we localize the balloon marker pair and the guidewire in between the marker pair on each projection angle from a rotational fluoroscopic sequence. We present an effective offline balloon marker tracking algorithm that leverages learning based detectors and employs the Viterbi algorithm to track the balloon markers in a globally optimal manner. Localizing the guidewire in between the tracked markers is formulated as tracking the middle control point of the spline fitting the guidewire. The experimental studies demonstrate that our methods achieve a marker tracking accuracy of 96.33% and a mean guidewire localization error of 0.46 mm, suggesting a great potential of our methods for clinical applications. The proposed offline marker tracking method is also successfully applied to the problem of automatic self-initialization of generic online marker trackers for 2D live fluoroscopy stream, demonstrating a success rate of 95.9% on 318 sequences. Its potential applications also include localization of landmarks in a generic rotational scan.

[S7-P16B]

Improving NCC-Based Direct Visual Tracking

Glauco Garcia Scandaroli, Maxime Meilland, and Rogério Richa

Direct visual tracking can be impaired by changes in illumination if the right choice of similarity function and photometric model is not made. Tracking using the sum of squared differences, for instance, often needs to be coupled with a photometric model to mitigate illumination changes. More sophisticated similarities, e.g. mutual information and cross cumulative residual entropy, however, can cope with complex illumination variations at the cost of a reduction of the convergence radius, and an increase of the computational effort. In this context, the normalized cross correlation (NCC) represents an interesting alternative. The NCC is intrinsically invariant to affine illumination changes, and also presents low computational cost. This article proposes a new direct visual tracking method based on the NCC. Two techniques have been developed to improve the robustness to complex illumination variations and partial occlusions. These techniques are based on subregion clusterization, and weighting by a residue invariant to affine illumination changes. The last contribution is an efficient Newton-style optimization procedure that does not require the explicit computation of the Hessian. The proposed method is compared against the state of the art using a benchmark database with ground-truth, as well as real-world sequences.

[S7-P17B]

Simultaneous Compaction and Factorization of Sparse Image Motion Matrices

Susanna Ricco and Carlo Tomasi

Matrices that collect the image coordinates of point features tracked through video – one column per feature – have often low rank, either exactly or approximately. This observation has led to many matrix factorization methods for 3D reconstruction, motion segmentation, or regularization of feature trajectories. However, temporary occlusions, image noise, and variations in lighting, pose, or object geometry often confound trackers. A feature that reappears after a temporary tracking failure – whatever the cause – is regarded as a new feature by typical tracking systems, resulting in very sparse matrices with many columns and rendering factorization problematic. We propose a method to simultaneously factor and compact such a matrix by merging groups of columns that correspond to the same feature into single columns. This combination of compaction and factorization makes trackers more resilient to changes in appearance and short occlusions. Preliminary experiments show that imputation of missing matrix entries – and therefore matrix factorization – becomes significantly more reliable as a result. Clean column merging also required us to develop a history-sensitive feature reinitialization method we call feature snapping that aligns merged feature trajectory segments precisely to each other.

[S7-P18B]

Low-Rank Sparse Learning for Robust Visual Tracking

Tianzhu Zhang, Bernard Ghanem, Si Liu, and Narendra Ahuja

In this paper, we propose a new particle-filter based tracking algorithm that exploits the relationship between particles (candidate targets). By representing particles as sparse linear combinations of dictionary templates, this algorithm capitalizes on the inherent low-rank structure of particle representations that are learned jointly. As such, it casts the tracking problem as a low-rank matrix learning problem. This low-rank sparse tracker (LRST) has a number of attractive properties. (1) Since LRST adaptively updates dictionary templates, it can handle significant changes in appearance due to variations in illumination, pose, scale, etc. (2) The linear representation in LRST explicitly incorporates background templates in the dictionary and a sparse error term, which enables LRST to address the tracking drift problem and to be robust against occlusion respectively. (3) LRST is computationally attractive, since the low-rank learning problem can be efficiently solved as a sequence of closed form update operations, which yield a time complexity that is linear in the number of particles and the template size. We evaluate the performance of LRST by applying it to a set of challenging video sequences and comparing it to 6 popular tracking methods. Our experiments show that by representing particles jointly, LRST not only outperforms the state-of-the-art in tracking accuracy but also significantly improves the time complexity of methods that use a similar sparse linear representation model for particles [1].

[S7-P19B]

Towards Optimal Design of Time and Color Multiplexing Codes

Tsung-Han Chan, Kui Jia, Eliot Wycoff, Chong-Yung Chi, and Yi Ma

Multiplexed illumination has been proved to be valuable and beneficial, in terms of noise reduction, in wide applications of computer vision and graphics, provided that the limitations of photon noise and saturation are properly tackled. Existing optimal multiplexing codes, in the sense of maximum signal-to-noise ratio (SNR), are primarily designed for time multiplexing, but they only apply to a multiplexing system requiring the number of measurements (M) equal to the number of illumination sources (N). In this paper, we formulate a general code design problem, where $M \geq N$, for time and color multiplexing, and develop a sequential semi-definite programming to deal with the formulated optimization problem. The proposed formulation and method can be readily specialized to time multiplexing, thereby making such optimized codes have a much broader application. Computer simulations will discover the main merit of the method— a significant boost of SNR as M increases. Experiments will also be presented to demonstrate the effectiveness and superiority of the method in object illumination.

[S7-P20B]

Dating Historical Color Images

Frank Palermo, James Hays, and Alexei A. Efros

We introduce the task of automatically estimating the age of historical color photographs. We suggest features which attempt to capture temporally discriminative information based on the evolution of color imaging processes over time and evaluate the performance of both these novel features and existing features commonly utilized in other problem domains on a novel historical image data set. For the challenging classification task of sorting historical color images into the decade during which they were photographed, we demonstrate significantly greater accuracy than that shown by untrained humans on the same data set. Additionally, we apply the concept of data-driven camera response function estimation to historical color imagery, demonstrating its relevance to both the age estimation task and the popular application of imitating the appearance of vintage color photography.

[S7-P21B]

Rainbow Flash Camera: Depth Edge Extraction Using Complementary Colors

Yuichi Taguchi

We present a novel color multiplexing method for extracting depth edges in a scene. It has been shown that casting shadows from different light positions provides a simple yet robust cue for extracting depth edges. Instead of flashing a single light source at a time as in conventional methods, our method flashes all light sources simultaneously to reduce the number of captured images. We use a ring light source around a camera and arrange colors on the ring such that the colors form a hue circle. Because complementary colors are arranged at any position and its antipole on the ring, shadow regions where a half of the hue circle is occluded are colorized according to the orientations of depth edges, while non-shadow regions where all the hues are mixed have a neutral color in the captured image. In an ideal situation, the colored shadows in a single image directly provide depth edges and their orientations. In practice, we present a robust depth edge extraction algorithm using an additional image captured by rotating the hue circle with 180°. We demonstrate the advantages of our approach using a camera prototype consisting of a standard camera and 8 color LEDs.

[S7-P22B]

Stixels Motion Estimation without Optical Flow Computation

Bertan Günyel, Rodrigo Benenson, Radu Timofte, and Luc Van Gool

This paper presents a new approach to estimate the motion of objects seen from a stereo rig mounted on a ground mobile robot. We exploit the prior knowledge on ground plane presence and rough shape of objects, to extract a simplified world model, named stixel world. The contribution of this paper is to show that stixels motion can be estimated directly solving a single dynamic programming problem instead of an image wide optical flow computation. We compare this new method with baseline methods, show competitive results quality-wise, and a significant gain speed-wise.

[S7-P23B]

Video Matting Using Multi-frame Nonlocal Matting Laplacian

Inchang Choi, Minhaeng Lee, and Yu-Wing Tai

We present an algorithm for extracting high quality temporally coherent alpha mattes of objects from a video. Our approach extends the conventional image matting approach, i.e. closed-form matting, to video by using multi-frame nonlocal matting Laplacian. Our multi-frame nonlocal matting Laplacian is defined over a nonlocal neighborhood in spatial temporal domain, and it solves the alpha mattes of several video frames all together simultaneously. To speed up computation and to reduce memory requirement for solving the multi-frame nonlocal matting Laplacian, we use the approximate nearest neighbor(ANN) to find the nonlocal neighborhood and the k-d tree implementation to divide the nonlocal matting Laplacian into several smaller linear systems. Finally, we adopt the nonlocal mean regularization to enhance temporal coherence of the estimated alpha mattes and to correct alpha matte errors at low contrast regions. We demonstrate the effectiveness of our approach on various examples with qualitative comparisons to the results from previous matting algorithms.

[S7-P24B]

Super-Resolution-Based Inpainting

Olivier Le Meur and Christine Guillemot

This paper introduces a new examplar-based inpainting framework. A coarse version of the input image is first inpainted by a non-parametric patch sampling. Compared to existing approaches, some improvements have been done (e.g. filling order computation, combination of K nearest neighbours). The inpainted of a coarse version of the input image allows to reduce the computational complexity, to be less sensitive to noise and to work with the dominant orientations of image structures. From the low-resolution inpainted image, a single-image super-resolution is applied to recover the details of missing areas. Experimental results on natural images and texture synthesis demonstrate the effectiveness of the proposed method.

[S7-P25B]

Fast Planar Correlation Clustering for Image Segmentation

Julian Yarkony, Alexander Ihler, and Charless C. Fowlkes

We describe a new optimization scheme for finding high-quality clusterings in planar graphs that uses weighted perfect matching as a subroutine. Our method provides lower-bounds on the energy of the optimal correlation clustering that are typically fast to compute and tight in practice. We demonstrate our algorithm on the problem of image segmentation where this approach outperforms existing global optimization techniques in minimizing the objective and is competitive with the state of the art in producing high-quality segmentations.

ORAL SESSION 7
LIGHTS, ACTION!

Thursday, October 11
11:20 - 13:00

[S7-O1]

Reflectance and Natural Illumination from a Single Image

Stephen Lombardi and Ko Nishino

Estimating reflectance and natural illumination from a single image of an object of known shape is a challenging task due to the ambiguities between reflectance and illumination. Although there is an inherent limitation in what can be recovered as the reflectance band-limits the illumination, explicitly estimating both is desirable for many computer vision applications. Achieving this estimation requires that we derive and impose strong constraints on both variables. We introduce a probabilistic formulation that seamlessly incorporates such constraints as priors to arrive at the maximum a posteriori estimates of reflectance and natural illumination. We begin by showing that reflectance modulates the natural illumination in a way that increases its entropy. Based on this observation, we impose a prior on the illumination that favors lower entropy while conforming to natural image statistics. We also impose a prior on the reflectance based on the directional statistics BRDF model that constrains the estimate to lie within the bounds and variability of real-world materials. Experimental results on a number of synthetic and real images show that the method is able to achieve accurate joint estimation for different combinations of materials and lighting.

[S7-O2]

Frequency-Space Decomposition and Acquisition of Light Transport under Spatially Varying Illumination

Dikpal Reddy, Ravi Ramamoorthi, and Brian Curless

We show that, under spatially varying illumination, the light transport of diffuse scenes can be decomposed into direct, near-range (subsurface scattering and local inter-reflections) and far-range transports (diffuse inter-reflections). We show that these three component transports are redundant either in the spatial or the frequency domain and can be separated using appropriate illumination patterns. We propose a novel, efficient method to sequentially separate and acquire the component transports. First, we acquire the direct transport by extending the direct-global separation technique from floodlit images to full transport matrices. Next, we separate and acquire the near-range transport by illuminating patterns sampled uniformly in the frequency domain. Finally, we acquire the far-range transport by illuminating low-frequency patterns. We show that theoretically, our acquisition method achieves the lower bound our model places on the required number of patterns. We quantify the savings in number of patterns over the brute force approach. We validate our observations and acquisition method with rendered and real examples throughout.

[S7-O3]

A Naturalistic Open Source Movie for Optical Flow Evaluation

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black

Ground truth optical flow is difficult to measure in real scenes with natural motion. As a result, optical flow data sets are restricted in terms of size, complexity, and diversity, making optical flow algorithms difficult to train and test on realistic data. We introduce a new optical flow data set derived from the open source 3D animated short film Sintel. This data set has important features not present in the popular Middlebury flow evaluation: long sequences, large motions, specular reflections, motion blur, defocus blur, and atmospheric effects. Because the graphics data that generated the movie is open source, we are able to render scenes under conditions of varying complexity to evaluate where existing flow algorithms fail. We evaluate several recent optical flow algorithms and find that current highly-ranked methods on the Middlebury evaluation have difficulty with this more complex data set suggesting further research on optical flow estimation is needed. To validate the use of synthetic data, we compare the image- and flow-statistics of Sintel to those of real films and videos and show that they are similar. The data set, metrics, and evaluation website are publicly available.

[S7-O4]

Streaming Hierarchical Video Segmentation

Chenliang Xu, Caiming Xiong, and Jason J. Corso

The use of video segmentation as an early processing step in video analysis lags behind the use of image segmentation for image analysis, despite many available video segmentation methods. A major reason for this lag is simply that videos are an order of magnitude bigger than images; yet most methods require all voxels in the video to be loaded into memory, which is clearly prohibitive for even medium length videos. We address this limitation by proposing an approximation framework for streaming hierarchical video segmentation motivated by data stream algorithms: each video frame is processed only once and does not change the segmentation of previous frames. We implement the graph-based hierarchical segmentation method within our streaming framework; our method is the first streaming hierarchical video segmentation method proposed. We perform thorough experimental analysis on a benchmark video data set and longer videos. Our results indicate the graph-based streaming hierarchical method outperforms other streaming video segmentation methods and performs nearly as well as the full-video hierarchical graph-based method.

[S7-O5]

Motion Capture of Hands in Action Using Discriminative Salient Points

Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys

Capturing the motion of two hands interacting with an object is a very challenging task due to the large number of degrees of freedom, self-occlusions, and similarity between the fingers, even in the case of multiple cameras observing the scene. In this paper we propose to use discriminatively learned salient points on the fingers and to estimate the finger-salient point associations simultaneously with the estimation of the hand pose. We introduce a differentiable objective function that also takes edges, optical flow and collisions into account. Our qualitative and quantitative evaluations show that the proposed approach achieves very accurate results for several challenging sequences containing hands and objects in action.

Photo Sequencing

Tali Basha, Yael Moses, and Shai Avidan

Dynamic events such as family gatherings, concerts or sports events are often captured by a group of people. The set of still images obtained this way is rich in dynamic content but lacks accurate temporal information. We propose a method for photo-sequencing – temporally ordering a set of still images taken asynchronously by a set of uncalibrated cameras. Photo-sequencing is an essential tool in analyzing (or visualizing) a dynamic scene captured by still images. The first step of the method detects sets of corresponding static and dynamic feature points across images. The static features are used to determine the epipolar geometry between pairs of images, and each dynamic feature votes for the temporal order of the images in which it appears. The partial orders provided by the dynamic features are not necessarily consistent, and we use rank aggregation to combine them into a globally consistent temporal order of images. We demonstrate successful photo sequencing on several challenging collections of images taken using a number of mobile phones.

POSTER SESSION 8

Thursday, October 11
14:30 - 17:00

[S8-P1A]

Co-inference for Multi-modal Scene Analysis

Daniel Munoz, James Andrew Bagnell, and Martial Hebert

We address the problem of understanding scenes from multiple sources of sensor data (e.g., a camera and a laser scanner) in the case where there is no one-to-one correspondence across modalities (e.g., pixels and 3-D points). This is an important scenario that frequently arises in practice not only when two different types of sensors are used, but also when the sensors are not co-located and have different sampling rates. Previous work has addressed this problem by restricting interpretation to a single representation in one of the domains, with augmented features that attempt to encode the information from the other modalities. Instead, we propose to analyze all modalities simultaneously while propagating information across domains during the inference procedure. In addition to the immediate benefit of generating a complete interpretation in all of the modalities, we demonstrate that this co-inference approach also improves performance over the canonical approach.

[S8-P2A]

A Unified View on Deformable Shape Factorizations

Roland Angst and Marc Pollefeys

Multiple-view geometry and structure-from-motion are well established techniques to compute the structure of a moving rigid object. These techniques are all based on strong algebraic constraints imposed by the rigidity of the object. Unfortunately, many scenes of interest, e.g. faces or cloths, are dynamic and the rigidity constraint no longer holds. Hence, there is a need for non-rigid structure-from-motion (NRSfM) methods which can deal with dynamic scenes. A prominent framework to model deforming and moving non-rigid objects is the factorization technique where the measurements are assumed to lie in a low-dimensional subspace. Many different formulations and variations for factorization-based NRSfM have been proposed in recent years. However, due to the complex interactions between several subspaces, the distinguishing properties between two seemingly related approaches are often unclear. For example, do two approaches just vary in the optimization method used or is really a different model beneath? In this paper, we show that these NRSfM factorization approaches are most naturally modeled with tensor algebra. This results in a clear presentation which subsumes many previous techniques. In this regard, this paper brings several strings of research together and provides a unified point of view. Moreover, the tensor formulation can be extended to the case of a camera network where multiple static affine cameras observe the same deforming and moving non-rigid object. Thanks to the insights gained through this tensor notation, a closed-form and an efficient iterative algorithm can be derived which provide a reconstruction even if there are no feature point correspondences at all between different cameras. An evaluation of the theory and algorithms on motion capture data show promising results.

[S8-P3A]

Finding the Exact Rotation between Two Images Independently of the Translation

Laurent Kneip, Roland Siegwart, and Marc Pollefeys

In this paper, we present a new epipolar constraint for computing the rotation between two images independently of the translation. Against the common belief in the field of geometric vision that it is not possible to find one independently of the other, we show how this can be achieved by relatively simple two-view constraints. We use the fact that translation and rotation cause fundamentally different flow fields on the unit sphere centered around the camera. This allows to establish independent constraints on translation and rotation, and the latter is solved using the Gröbner basis method. The rotation computation is completed by a solution to the cheiriality problem that depends neither on translation, nor on feature triangulations. Notably, we show for the first time how the constraint on the rotation has the advantage of remaining exact even in the case of translations converging to zero. We use this fact in order to remove the error caused by model selection via a non-linear optimization of rotation hypotheses. We show that our method operates in real-time and compare it to a standard existing approach in terms of both speed and accuracy.

[S8-P4A]

A New Set of Quartic Trivariate Polynomial Equations for Stratified Camera Self-calibration under Zero-Skew and Constant Parameters Assumptions

Adlane Habed, Kassem Al Ismaeil, and David Fofi

This paper deals with the problem of self-calibrating a moving camera with constant parameters. We propose a new set of quartic trivariate polynomial equations in the unknown coordinates of the plane at infinity derived under the no-skew assumption. Our new equations allow to further enforce the constancy of the principal point across all images while retrieving the plane at infinity. Six such polynomials, four of which are independent, are obtained for each triplet of images. The proposed equations can be solved along with the so-called modulus constraints and allow to improve the performance of existing methods.

[S8-P5A]

A Minimal Solution for Camera Calibration Using Independent Pairwise Correspondences

Francisco Vasconcelos, João Pedro Barreto, and Edmond Boyer

We propose a minimal algorithm for fully calibrating a camera from 11 independent pairwise point correspondences with two other calibrated cameras. Unlike previous approaches, our method neither requires triple correspondences, nor prior knowledge about the viewed scene. This algorithm can be used to insert or re-calibrate a new camera into an existing network, without having to interrupt operation. Its main strength comes from the fact that it is often difficult to find triple correspondences in a camera network. This makes our algorithm, for the specified use cases, probably the most suited calibration solution that does not require a calibration target, and hence can be performed without human interaction.

[S8-P6A]

Real-Time Human Pose Tracking from Range Data

Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun

Tracking human pose in real-time is a difficult problem with many interesting applications. Existing solutions suffer from a variety of problems, especially when confronted with unusual human poses. In this paper, we derive an algorithm for tracking human pose in real-time from depth sequences based on MAP inference in a probabilistic temporal model. The key idea is to extend the iterative closest points (ICP) objective by modeling the constraint that the observed subject cannot enter free space, the area of space in front of the true range measurements. Our primary contribution is an extension to the articulated ICP algorithm that can efficiently enforce this constraint. The resulting filter runs at 125 frames per second using a single desktop CPU core. We provide extensive experimental results on challenging real-world data, which show that the algorithm outperforms the previous state-of-the-art trackers both in computational efficiency and accuracy.

[S8-P7A]

Large-Lexicon Attribute-Consistent Text Recognition in Natural Images

Tatiana Novikova, Olga Barinova, Pushmeet Kohli, and Victor Lempitsky

This paper proposes a new model for the task of word recognition in natural images that simultaneously models visual and lexicon consistency of words in a single probabilistic model. Our approach combines local likelihood and pairwise positional consistency priors with higher order priors that enforce consistency of characters (lexicon) and their attributes (font and colour). Unlike traditional stage-based methods, word recognition in our framework is performed by estimating the maximum a posteriori (MAP) solution under the joint posterior distribution of the model. MAP inference in our model is performed through the use of weighted finite-state transducers (WFSTs). We show how the efficiency of certain operations on WFSTs can be utilized to find the most likely word under the model in an efficient manner. We evaluate our method on a range of challenging datasets (ICDAR'03, SVT, ICDAR'11). Experimental results demonstrate that our method outperforms state-of-the-art methods for cropped word recognition.

[S8-P8A]

Dictionary-Based Face Recognition from Video

Yi-Chen Chen, Vishal M. Patel, P. Jonathon Phillips, and Rama Chellappa

The main challenge in recognizing faces in video is effectively exploiting the multiple frames of a face and the accompanying dynamic signature. One prominent method is based on extracting joint appearance and behavioral features. A second method models a person by temporal correlations of features in a video. Our approach introduces the concept of video-dictionaries for face recognition, which generalizes the work in sparse representation and dictionaries for faces in still images. Video-dictionaries are designed to implicitly encode temporal, pose, and illumination information. We demonstrate our method on the Face and Ocular Challenge Series (FOCS) Video Challenge, which consists of unconstrained video sequences. We show that our method is efficient and performs significantly better than many competitive video-based face recognition algorithms.

[S8-P9A]

Relaxed Pairwise Learned Metric for Person Re-identification

Martin Hirzer, Peter M. Roth, Martin Köstinger, and Horst Bischof

Matching persons across non-overlapping cameras is a rather challenging task. Thus, successful methods often build on complex feature representations or sophisticated learners. A recent trend to tackle this problem is to use metric learning to find a suitable space for matching samples from different cameras. However, most of these approaches ignore the transition from one camera to the other. In this paper, we propose to learn a metric from pairs of samples from different cameras. In this way, even less sophisticated features describing color and texture information are sufficient for finally getting state-of-the-art classification results. Moreover, once the metric has been learned, only linear projections are necessary at search time, where a simple nearest neighbor classification is performed. The approach is demonstrated on three publicly available datasets of different complexity, where it can be seen that state-of-the-art results can be obtained at much lower computational costs.

[S8-P10A]

Connecting Missing Links: Object Discovery from Sparse Observations Using 5 Million Product Images

Hongwen Kang, Martial Hebert, Alexei A. Efros, and Takeo Kanade

Object discovery algorithms group together image regions that originate from the same object. This process is effective when the input collection of images contains a large number of densely sampled views of each object, thereby creating strong connections between nearby views. However, existing approaches are less effective when the input data only provide sparse coverage of object views. We propose an approach for object discovery that addresses this problem. We collect a database of about 5 million product images that capture 1.2 million objects from multiple views. We represent each region in the input image by a ''bag'' of database object regions. We group input regions together if they share similar ''bags of regions''. Our approach can correctly discover links between regions of the same object even if they are captured from dramatically different viewpoints. With the help from these added links, our proposed approach can robustly discover object instances even with sparse coverage of the viewpoints.

[S8-P11A]

Disentangling Factors of Variation for Facial Expression Recognition

Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza

We propose a semi-supervised approach to solve the task of emotion recognition in 2D face images using recent ideas in deep learning for handling the factors of variation present in data. An emotion classification algorithm should be both robust to (1) remaining variations due to the pose of the face in the image after centering and alignment, (2) the identity or morphology of the face. In order to achieve this invariance, we propose to learn a hierarchy of features in which we gradually filter the factors of variation arising from both (1) and (2). We address (1) by using a multi-scale contractive convolutional network (CCNET) in order to obtain invariance to translations of the facial traits in the image. Using the feature representation produced by the CCNET, we train a Contractive Discriminative Analysis (CDA) feature extractor, a novel variant of the Contractive Auto-Encoder (CAE), designed to learn a representation separating out the emotion-related factors from the others (which mostly capture the subject identity, and what is left of pose after the CCNET). This system beats the state-of-the-art on a recently proposed dataset for facial expression recognition, the Toronto Face Database, moving the state-of-art accuracy from 82.4% to 85.0%, while the CCNET and CDA improve accuracy of a standard CAE by 8%.

[S8-P12A]

Simultaneous Image Classification and Annotation via Biased Random Walk on Tri-relational Graph

Xiao Cai, Hua Wang, Heng Huang, and Chris Ding

Image annotation as well as classification are both critical and challenging work in computer vision research. Due to the rapid increasing number of images and inevitable biased annotation or classification by the human curator, it is desired to have an automatic way. Recently, there are lots of methods proposed regarding image classification or image annotation. However, people usually treat the above two tasks independently and tackle them separately. Actually, there is a relationship between the image class label and image annotation terms. As we know, an image with the sport class label rowing is more likely to be annotated with the terms water, boat and oar than the terms wall, net and floor, which are the descriptions of indoor sports. In this paper, we propose a new method for jointly class recognition and terms annotation. We present a novel Tri-Relational Graph (TG) model that comprises the data graph, annotation terms graph, class label graph, and connect them by two additional graphs induced from class label as well as annotation assignments. Upon the TG model, we introduce a Biased Random Walk (BRW) method to jointly recognize class and annotate terms by utilizing the interrelations between two tasks. We conduct the proposed method on two benchmark data sets and the experimental results demonstrate our joint learning method can achieve superior prediction results on both tasks than the state-of-the-art methods.

[S8-P13A]

Spring Lattice Counting Grids: Scene Recognition Using Deformable Positional Constraints

Alessandro Perina and Nebojsa Jojic

Adopting the Counting Grid (CG) representation [1], the Spring Lattice Counting Grid (SLCG) model uses a grid of feature counts to capture the spatial layout that a variety of images tend to follow. The images are mapped to the counting grid with their features rearranged so as to strike a balance between the mapping quality and the extent of the necessary rearrangement. In particular, the feature sets originating from different image sectors are mapped to different sub-windows in the counting grid in a configuration that is close, but not exactly the same as the configuration of the source sectors. The distribution over deformations of the sector configuration is learnable using a new spring lattice model, while the rearrangement of features within a sector is unconstrained. As a result, the CG model gains a more appropriate level of invariance to realistic image transformations like view point changes, rotations or scales. We tested SLCG on standard scene recognition datasets and on a dataset collected with a wearable camera which recorded the wearer's visual input over three weeks. Our algorithm is capable of correctly classifying the visited locations more than 80% of the time, outperforming previous approaches to visual location recognition. At this level of performance, a variety of real-world applications of wearable cameras become feasible.

[S8-P14A]

Hand Pose Estimation and Hand Shape Classification Using Multi-layered Randomized Decision Forests

Cem Keskin, Furkan Kiraç, Yunus Emre Kara, and Lale Akarun

Vision based articulated hand pose estimation and hand shape classification are challenging problems. This paper proposes novel algorithms to perform these tasks using depth sensors. In particular, we introduce a novel randomized decision forest (RDF) based hand shape classifier, and use it in a novel multi–layered RDF framework for articulated hand pose estimation. This classifier assigns the input depth pixels to hand shape classes, and directs them to the corresponding hand pose estimators trained specifically for that hand shape. We introduce two novel types of multi–layered RDFs: Global Expert Network (GEN) and Local Expert Network (LEN), which achieve significantly better hand pose estimates than a single–layered skeleton estimator and generalize better to previously unseen hand poses. The novel hand shape classifier is also shown to be accurate and fast. The methods run in real–time on the CPU, and can be ported to the GPU for further increase in speed.

[S8-P15A]

Information Theoretic Learning for Pixel-Based Visual Agents

Marco Gori, Stefano Melacci, Marco Lippi, and Marco Maggini

In this paper we promote the idea of using pixel-based models not only for low level vision, but also to extract high level symbolic representations. We use a deep architecture which has the distinctive property of relying on computational units that incorporate classic computer vision invariances and, especially, the scale invariance. The learning algorithm that is proposed, which is based on information theory principles, develops the parameters of the computational units and, at the same time, makes it possible to detect the optimal scale for each pixel. We give experimental evidence of the mechanism of feature extraction at the first level of the hierarchy, which is very much related to SIFT-like features. The comparison shows clearly that, whenever we can rely on the massive availability of training data, the proposed model leads to better performances with respect to SIFT.

[S8-P16A]

Attribute Discovery via Predictable Discriminative Binary Codes

Mohammad Rastegari, Ali Farhadi, and David Forsyth

We present images with binary codes in a way that balances discrimination and learnability of the codes. In our method, each image claims its own code in a way that maintains discrimination while being predictable from visual data. Category memberships are usually good proxies for visual similarity but should not be enforced as a hard constraint. Our method learns codes that maximize separability of categories unless there is strong visual evidence against it. Simple linear SVMs can achieve state-of-the-art results with our short codes. In fact, our method produces state-of-the-art results on Caltech256 with only 128-dimensional bit vectors and outperforms state of the art by using longer codes. We also evaluate our method on ImageNet and show that our method outperforms state-of-the-art binary code methods on this large scale dataset. Lastly, our codes can discover a discriminative set of attributes.

[S8-P17A]

Local Higher-Order Statistics (LHS) for Texture Categorization and Facial Analysis

Gaurav Sharma, Sibt ul Hussain, and Frédéric Jurie

This paper proposes a new image representation for texture categorization and facial analysis, relying on the use of higher-order local differential statistics as features. In contrast with models based on the global structure of textures and faces, it has been shown recently that small local pixel pattern distributions can be highly discriminative. Motivated by such works, the proposed model employs higher-order statistics of local non-binarized pixel patterns for the image description. Hence, in addition to being remarkably simple, it requires neither any user specified quantization of the space (of pixel patterns) nor any heuristics for discarding low occupancy volumes of the space. This leads to a more expressive representation which, when combined with discriminative SVM classifier, consistently achieves state-of-the-art performance on challenging texture and facial analysis datasets outperforming contemporary methods (with similar powerful classifiers).

[S8-P18A]

SEEDS: Superpixels Extracted via Energy-Driven Sampling

Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin de Capitani, and Luc Van Gool

Superpixel algorithms aim to over-segment the image by grouping pixels that belong to the same object. Many state-of-the-art superpixel algorithms rely on minimizing objective functions to enforce color homogeneity. The optimization is accomplished by sophisticated methods that progressively build the superpixels, typically by adding cuts or growing superpixels. As a result, they are computationally too expensive for real-time applications. We introduce a new approach based on a simple hill-climbing optimization. Starting from an initial superpixel partitioning, it continuously refines the superpixels by modifying the boundaries. We define a robust and fast to evaluate energy function, based on enforcing color similarity between the boundaries and the superpixel color histogram. In a series of experiments, we show that we achieve an excellent compromise between accuracy and efficiency. We are able to achieve a performance comparable to the state-of-the-art, but in real-time on a single Intel i7 CPU at 2.8GHz.

[S8-P19A]

Recording and Playback of Camera Shake: Benchmarking Blind Deconvolution with a Real-World Database

Rolf Köhler, Michael Hirsch, Betty Mohler, Bernhard Schölkopf, and Stefan Harmeling

Motion blur due to camera shake is one of the predominant sources of degradation in handheld photography. Single image blind deconvolution (BD) or motion deblurring aims at restoring a sharp latent image from the blurred recorded picture without knowing the camera motion that took place during the exposure. BD is a long-standing problem, but has attracted much attention recently, cumulating in several algorithms able to restore photos degraded by real camera motion in high quality. In this paper, we present a benchmark dataset for motion deblurring that allows quantitative performance evaluation and comparison of recent approaches featuring non-uniform blur models. To this end, we record and analyse real camera motion, which is played back on a robot platform such that we can record a sequence of sharp images sampling the six dimensional camera motion trajectory. The goal of deblurring is to recover one of these sharp images, and our dataset contains all information to assess how closely various algorithms approximate that goal. In a comprehensive comparison, we evaluate state-of-the-art single image BD algorithms incorporating uniform and non-uniform blur models.

[S8-P20A]

Learning-Based Symmetry Detection in Natural Images

Stavros Tsogkas and Iasonas Kokkinos

In this work we propose a learning-based approach to symmetry detection in natural images. We focus on ribbon-like structures, i.e. contours marking local and approximate reflection symmetry and make three contributions to improve their detection. First, we create and make publicly available a ground-truth dataset for this task by building on the Berkeley Segmentation Dataset. Second, we extract features representing multiple complementary cues, such as grayscale structure, color, texture, and spectral clustering information. Third, we use supervised learning to learn how to combine these cues, and employ MIL to accommodate the unknown scale and orientation of the symmetric structures. We systematically evaluate the performance contribution of each individual component in our pipeline, and demonstrate that overall we consistently improve upon results obtained using existing alternatives.

[S8-P21A]

Similarity Constrained Latent Support Vector Machine: An Application to Weakly Supervised Action Classification

Nataliya Shapovalova, Arash Vahdat, Kevin Cannons, Tian Lan, and Greg Mori

We present a novel algorithm for weakly supervised action classification in videos. We assume we are given training videos annotated only with action class labels. We learn a model that can classify unseen test videos, as well as localize a region of interest in the video that captures the discriminative essence of the action class. A novel Similarity Constrained Latent Support Vector Machine model is developed to operationalize this goal. This model specifies that videos should be classified correctly, and that the latent regions of interest chosen should be coherent over videos of an action class. The resulting learning problem is challenging, and we show how dual decomposition can be employed to render it tractable. Experimental results demonstrate the efficacy of the method.

[S8-P22A]

Team Activity Recognition in Sports

Cem Direkoglu and Noel E. O'Connor

We introduce a novel approach for team activity recognition in sports. Given the positions of team players from a plan view of the playing field at any given time, we solve a particular Poisson equation to generate a smooth distribution defined on whole playground, termed the position distribution of the team. Computing the position distribution for each frame provides a sequence of distributions, which we process to extract motion features for team activity recognition. The motion features are obtained at each frame using frame differencing and optical flow. We investigate the use of the proposed motion descriptors with Support Vector Machines (SVM) classification, and evaluate on a publicly available European handball dataset. Results show that our approach can classify six different team activities and performs better than a method that extracts features from the explicitly defined positions. Our method is new and different from other trajectory-based methods. These methods extract activity features using the explicitly defined trajectories, where the players have specific positions at any given time, and ignore the rest of the playground. In our work, on the other hand, given the specific positions of the team players at a frame, we construct a position distribution for the team on the whole playground and process the sequence of position distribution images to extract motion features for activity recognition. Results show that our approach is effective.

[S8-P1B]

Space-Variant Descriptor Sampling for Action Recognition Based on Saliency and Eye Movements

Eleonora Vig, Michael Dorr, and David Cox

Algorithms using "bag of features"-style video representations currently achieve state-of-the-art performance on action recognition tasks, such as the challenging Hollywood2 benchmark [1,2,3]. These algorithms are based on local spatiotemporal descriptors that can be extracted either sparsely (at interest points) or densely (on regular grids), with dense sampling typically leading to the best performance [1]. Here, we investigate the benefit of space-variant processing of inputs, inspired by attentional mechanisms in the human visual system. We employ saliency-mapping algorithms to find informative regions and descriptors corresponding to these regions are either used exclusively, or are given greater representational weight (additional codebook vectors). This approach is evaluated with three state-of-the-art action recognition algorithms [1,2,3], and using several saliency algorithms. We also use saliency maps derived from human eye movements to probe the limits of the approach. Saliency-based pruning allows up to 70% of descriptors to be discarded, while maintaining high performance on Hollywood2. Meanwhile, pruning of 20-50% (depending on model) can even improve recognition. Further improvements can be obtained by combining representations learned separately on salience-pruned and unpruned descriptor sets. Not surprisingly, using the human eye movement data gives the best mean Average Precision (mAP; 61.9%), providing an upper bound on what is possible with a high-quality saliency map. Even without such external data, the Dense Trajectories model [1] enhanced by automated saliency-based descriptor sampling achieves the best mAP (60.0%) reported on Hollywood2 to date.

[S8-P2B]

Dynamic Probabilistic CCA for Analysis of Affective Behaviour

Mihalis A. Nicolaou, Vladimir Pavlovic, and Maja Pantic

Fusing multiple continuous expert annotations is a crucial problem in machine learning and computer vision, particularly when dealing with uncertain and subjective tasks related to affective behaviour. Inspired by the concept of inferring shared and individual latent spaces in probabilistic CCA (PCCA), we firstly propose a novel, generative model which discovers temporal dependencies on the shared/individual spaces (DPCCA). In order to accommodate for temporal lags which are prominent amongst continuous annotations, we further introduce a latent warping process. We show that the resulting model (DPCTW) (i) can be used as a unifying framework for solving the problems of temporal alignment and fusion of multiple annotations in time, and (ii) that by incorporating dynamics, modelling annotation/sequence specific biases, noise estimation and time warping, DPCTW outperforms state-of-the-art methods for both the aggregation of multiple, yet imperfect expert annotations as well as the alignment of affective behavior.

[S8-P3B]

Loss-Specific Training of Non-Parametric Image Restoration Models: A New State of the Art

Jeremy Jancsary, Sebastian Nowozin, and Carsten Rother

After a decade of rapid progress in image denoising, recent methods seem to have reached a performance limit. Nonetheless, we find that state-of-the-art denoising methods are visually clearly distinguishable and possess complementary strengths and failure modes. Motivated by this observation, we introduce a powerful non-parametric image restoration framework based on Regression Tree Fields (RTF). Our restoration model is a densely-connected tractable conditional random field that leverages existing methods to produce an image-dependent, globally consistent prediction. We estimate the conditional structure and parameters of our model from training data so as to directly optimize for popular performance measures. In terms of peak signal-to-noise-ratio (PSNR), our model improves on the best published denoising method by at least 0.26dB across a range of noise levels. Our most practical variant still yields statistically significant improvements, yet is over 20× faster than the strongest competitor. Our approach is well-suited for many more image restoration and low-level vision problems, as evidenced by substantial gains in tasks such as removal of JPEG blocking artefacts.

[S8-P4B]

A Probabilistic Approach to Robust Matrix Factorization

Naiyan Wang, Tiansheng Yao, Jingdong Wang, and Dit-Yan Yeung

Matrix factorization underlies a large variety of computer vision applications. It is a particularly challenging problem for large-scale applications and when there exist outliers and missing data. In this paper, we propose a novel probabilistic model called Probabilistic Robust Matrix Factorization (PRMF) to solve this problem. In particular, PRMF is formulated with a Laplace error and a Gaussian prior which correspond to an l_1 loss and an l_2 regularizer, respectively. For model learning, we devise a parallelizable expectation-maximization (EM) algorithm which can potentially be applied to large-scale applications. We also propose an online extension of the algorithm for sequential data to offer further scalability. Experiments conducted on both synthetic data and some practical computer vision applications show that PRMF is comparable to other state-of-the-art robust matrix factorization methods in terms of accuracy and outperforms them particularly for large data matrices.

[S8-P5B]

Fast Parameter Sensitivity Analysis of PDE-Based Image Processing Methods

Torben Pätz and Tobias Preusser

We present a fast parameter sensitivity analysis by combining recent developments from uncertainty quantification with image processing operators. The approach is not based on a sampling strategy, instead we combine the polynomial chaos expansion and stochastic finite elements with PDE-based image processing operators. With our approach and a moderate number of parameters in the models the full sensitivity analysis is obtained at the cost of a few Monte Carlo runs. To demonstrate the efficiency and simplicity of the approach we show a parameter sensitivity analysis for Perona-Malik diffusion, random walker and Ambrosio-Tortorelli segmentation, and discontinuity-preserving optical flow computation.

[S8-P6B]

The Lazy Flipper: Efficient Depth-Limited Exhaustive Search in Discrete Graphical Models

Bjoern Andres, Jörg H. Kappes, Thorsten Beier, Ullrich Köthe, and Fred A. Hamprecht

We propose a new exhaustive search algorithm for optimization in discrete graphical models. When pursued to the full search depth (typically intractable), it is guaranteed to converge to a global optimum, passing through a series of monotonously improving local optima that are guaranteed to be optimal within a given and increasing Hamming distance. For a search depth of 1, it specializes to ICM. Between these extremes, a tradeoff between approximation quality and runtime is established. We show this experimentally by improving approximations for the non-submodular models in the MRF benchmark [1] and Decision Tree Fields [2].

[S8-P7B]

Face Association across Unconstrained Video Frames Using Conditional Random Fields

Ming Du and Rama Chellappa

Automatic face association across unconstrained video frames has many practical applications. Recent advances in the area of object detection have made it possible to replace the traditional tracking-based association approaches with the more robust detection-based ones. However, it is still a very challenging task for real-world unconstrained videos, especially if the subjects are in a moving platform and at distances exceeding several tens of meters. In this paper, we present a novel solution based on a Conditional Random Field (CRF) framework. The CRF approach not only gives a probabilistic and systematic treatment of the problem, but also elegantly combines global and local features. When ambiguities in labels cannot be solved by using the face appearance alone, our method relies on multiple contextual features to provide further evidence for association. Our algorithm works in an on-line mode and is able to reliably handle real-world videos. Results of experiments using challenging video data and comparisons with other methods are provided to demonstrate the effectiveness of our method.

[S8-P8B]

Contraction Moves for Geometric Model Fitting

Oliver J. Woodford, Minh-Tri Pham, Atsuto Maki, Riccardo Gherardi, Frank Perbet, and Björn Stenger

This paper presents a new class of moves, called α-expansion-contraction, which generalizes α-expansion graph cuts for multi-label energy minimization problems. The new moves are particularly useful for optimizing the assignments in model fitting frameworks whose energies include Label Cost (LC), as well as Markov Random Field (MRF) terms. These problems benefit from the contraction moves' greater scope for removing instances from the model, reducing label costs. We demonstrate this effect on the problem of fitting sets of geometric primitives to point cloud data, including real-world point clouds containing millions of points, obtained by multi-view reconstruction.

[S8-P9B]

General and Nested Wiberg Minimization: L_2 and Maximum Likelihood

Dennis Strelow

Wiberg matrix factorization breaks a matrix Y into low-rank factors U and V by solving for V in closed form given U, linearizing V(U) about U, and iteratively minimizing ||Y – UV(U)||2 with respect to U only. This approach factors the matrix while effectively removing V from the minimization. We generalize the Wiberg approach beyond factorization to minimize an arbitrary function that is nonlinear in each of two sets of variables. In this paper we focus on the case of L2 minimization and maximum likelihood estimation (MLE), presenting an L2 Wiberg bundle adjustment algorithm and a Wiberg MLE algorithm for Poisson matrix factorization. We also show that one Wiberg minimization can be nested inside another, effectively removing two of three sets of variables from a minimization. We demonstrate this idea with a nested Wiberg algorithm for L2 projective bundle adjustment, solving for camera matrices, points, and projective depths.

[S8-P10B]

Nonmetric Priors for Continuous Multilabel Optimization

Evgeny Strekalovskiy, Claudia Nieuwenhuis, and Daniel Cremers

We propose a novel convex prior for multilabel optimization which allows to impose arbitrary distances between labels. Only symmetry, $d(i,j) \geq 0$ and $d(i,i) = 0$ are required. In contrast to previous grid based approaches for the nonmetric case, the proposed prior is formulated in the continuous setting avoiding grid artifacts. In particular, the model is easy to implement, provides a convex relaxation for the Mumford-Shah functional and yields comparable or superior results on the MSRC segmentation database comparing to metric or grid based approaches.

[S8-P11B]

Real-Time Camera Tracking: When is High Frame-Rate Best?

Ankur Handa, Richard A. Newcombe, Adrien Angeli, and Andrew J. Davison

Higher frame-rates promise better tracking of rapid motion, but advanced real-time vision systems rarely exceed the standard 10–60Hz range, arguing that the computation required would be too great. Actually, increasing frame-rate is mitigated by reduced computational cost per frame in trackers which take advantage of prediction. Additionally, when we consider the physics of image formation, high frame-rate implies that the upper bound on shutter time is reduced, leading to less motion blur but more noise. So, putting these factors together, how are application-dependent performance requirements of accuracy, robustness and computational cost optimised as frame-rate varies? Using 3D camera tracking as our test problem, and analysing a fundamental dense whole image alignment approach, we open up a route to a systematic investigation via the careful synthesis of photorealistic video using ray-tracing of a detailed 3D scene, experimentally obtained photometric response and noise models, and rapid camera motions. Our multi-frame-rate, multi-resolution, multi-light-level dataset is based on tens of thousands of hours of CPU rendering time. Our experiments lead to quantitative conclusions about frame-rate selection and highlight the crucial role of full consideration of physical image formation in pushing tracking performance.

[S8-P12B]

A Bayesian Approach to Alignment-Based Image Hallucination

Marshall F. Tappen and Ce Liu

In most image hallucination work, a strong assumption is held that images can be aligned to a template on which the prior of high-res images is formulated and learned. Realizing that one template can hardly generalize to all images of an object such as faces due to pose and viewpoint variation as well as occlusion, we propose an example-based prior distribution via dense image correspondences. We introduce a Bayesian formulation based on an image prior that can implement different effective behaviors based on the value of a single parameter. Using faces as examples, we show that our system outperforms the prior state of art.

[S8-P13B]

Continuous Regression for Non-rigid Image Alignment

Enrique Sánchez-Lozano, Fernando De la Torre, and Daniel González-Jiménez

Parameterized Appearance Models (PAMs) such as Active Appearance Models (AAMs), Morphable Models and Boosted Appearance Models have been extensively used for face alignment. Broadly speaking, PAMs methods can be classified into generative and discriminative. Discriminative methods learn a mapping between appearance features and motion parameters (rigid and non-rigid). While discriminative approaches have some advantages (e.g., feature weighting, improved generalization), they suffer from two major drawbacks: (1) they need large amounts of perturbed samples to train a regressor or classifier, making the training process computationally expensive in space and time. (2) It is not practical to uniformly sample the space of motion parameters. In practice, there are regions of the motion space that are more densely sampled than others, resulting in biased models and lack of generalization. To solve these problems, this paper proposes a computationally efficient continuous regressor that does not require the sampling stage. Experiments on real data show the improvement in memory and time requirements to train a discriminative appearance model, as well as improved generalization.

[S8-P14B]

Non-rigid Shape Registration: A Single Linear Least Squares Framework

Mohammad Rouhani and Angel D. Sappa

This paper proposes a non-rigid registration formulation capturing both global and local deformations in a single framework. This formulation is based on a quadratic estimation of the registration distance together with a quadratic regularization term. Hence, the optimal transformation parameters are easily obtained by solving a liner system of equations, which guarantee a fast convergence. Experimental results with challenging 2D and 3D shapes are presented to show the validity of the proposed framework. Furthermore, comparisons with the most relevant approaches are provided.

[S8-P15B]

Robust and Accurate Shape Model Fitting Using Random Forest Regression Voting

Tim F. Cootes, Mircea C. Ionita, Claudia Lindner, and Patrick Sauer

A widely used approach for locating points on deformable objects is to generate feature response images for each point, then to fit a shape model to the response images. We demonstrate that Random Forest regression can be used to generate high quality response images quickly. Rather than using a generative or a discriminative model to evaluate each pixel, a regressor is used to cast votes for the optimal position. We show this leads to fast and accurate matching when combined with a statistical shape model. We evaluate the technique in detail, and compare with a range of commonly used alternatives on several different datasets. We show that the random forest regression method is significantly faster and more accurate than equivalent discriminative, or boosted regression based methods trained on the same data.

[S8-P16B]

Shape from Fluorescence

Tali Treibitz, Zak Murez, B. Greg Mitchell, and David Kriegman

Beyond day glow highlighters and psychedelic black light posters, it has been estimated that fluorescence is a property exhibited by 20% of objects. When a fluorescent material is illuminated with a short wavelength light, it re-emits light at a longer wavelength isotropically in a similar manner as a Lambertian surface reflects light. This hitherto neglected property opens the doors to using fluorescence to reconstruct 3D shape with some of the same techniques as for Lambertian surfaces – even when the surface's reflectance is highly non-Lambertian. Thus, performing reconstruction using fluorescence has advantages over purely Lambertian surfaces. Single image shape-from-shading and calibrated Lambertian photometric stereo can be applied to fluorescence images to reveal 3D shape. When performing uncalibrated photometric stereo, both fluorescence and reflectance can be used to recover Euclidean shape and resolve the generalized bas relief ambiguity. Finally for objects that fluoresce in wavelengths distinct from their reflectance (such as plants and vegetables), reconstructions do not suffer from problems due to inter-reflections. We validate these claims through experiments.

[S8-P17B]

Separability Oriented Preprocessing for Illumination-Insensitive Face Recognition

Hu Han, Shiguang Shan, Xilin Chen, Shihong Lao, and Wen Gao

In the last decade, some illumination preprocessing approaches were proposed to eliminate the lighting variation in face images for lighting-invariant face recognition. However, we find surprisingly that existing preprocessing methods were seldom modeled to directly enhance the separability of different faces, which should have been the essential goal. To address the issue, we propose to explicitly exploit maximizing separability of different subjects' faces as the preprocessing objective. With this in mind, a novel approach, named by us Separability Oriented Preprocessing (SOP), is proposed to enhance face images by maximizing the Fisher separability criterion in scale-space. Extensive experiments on both laboratory-controlled and real-world face databases using different recognition methods show the effectiveness of the proposed approach.

[S8-P18B]

Saliency Modeling from Image Histograms

Shijian Lu and Joo-Hwee Lim

We proposed a computational visual saliency modeling technique. The proposed technique makes use of a color co-occurrence histogram (CCH) that captures not only "how many" but also "where and how" image pixels are composed into a visually perceivable image. Hence the CCH encodes image saliency information that is usually perceived as the discontinuity between an image region or object and its surrounding. The proposed technique has a number of distinctive characteristics: It is fast, discriminative, tolerant to image scale variation, and involves minimal parameter tuning. Experiments over benchmarking datasets show that it predicts fixational eye tracking points accurately and a superior AUC of 71.25 is obtained.

[S8-P19B]

A Theoretical Analysis of Camera Response Functions in Image Deblurring

Xiaogang Chen, Feng Li, Jie Yang, and Jingyi Yu

Motion deblurring is a long standing problem in computer vision and image processing. In most previous approaches, the blurred image is modeled as the convolution of a latent intensity image with a blur kernel. However, for images captured by a real camera, the blur convolution should be applied to scene irradiance instead of image intensity and the blurred results need to be mapped back to image intensity via the camera's response function (CRF). In this paper, we present a comprehensive study to analyze the effects of CRFs on motion deblurring. We prove that the intensity-based model closely approximates the irradiance model at low frequency regions. However, at high frequency regions such as edges, the intensity-based approximation introduces large errors and directly applying deconvolution on the intensity image will produce strong ringing artifacts even if the blur kernel is invertible. Based on the approximation error analysis, we further develop a dual-image based solution that captures a pair of sharp/blurred images for both CRF estimation and motion deblurring. Experiments on synthetic and real images validate our theories and demonstrate the robustness and accuracy of our approach.

[S8-P20B]

Robust and Efficient Subspace Segmentation via Least Squares Regression

Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan

This paper studies the subspace segmentation problem which aims to segment data drawn from a union of multiple linear subspaces. Recent works by using sparse representation, low rank representation and their extensions attract much attention. If the subspaces from which the data drawn are independent or orthogonal, they are able to obtain a block diagonal affinity matrix, which usually leads to a correct segmentation. The main differences among them are their objective functions. We theoretically show that if the objective function satisfies some conditions, and the data are sufficiently drawn from independent subspaces, the obtained affinity matrix is always block diagonal. Furthermore, the data sampling can be insufficient if the subspaces are orthogonal. Some existing methods are all special cases. Then we present the Least Squares Regression (LSR) method for subspace segmentation. It takes advantage of data correlation, which is common in real data. LSR encourages a grouping effect which tends to group highly correlated data together. Experimental results on the Hopkins 155 database and Extended Yale Database B show that our method significantly outperforms state-of-the-art methods. Beyond segmentation accuracy, all experiments demonstrate that LSR is much more efficient.

[S8-P21B]

Local Label Descriptor for Example Based Semantic Image Labeling

Yiqing Yang, Zhouyuan Li, Li Zhang, Christopher Murphy, Jim Ver Hoeve, and Hongrui Jiang

In this paper we introduce the concept of local label descriptor, which is a concatenation of label histograms for each cell in a patch. Local label descriptors alleviate the label patch misalignment issue in combining structured label predictions for semantic image labeling. Given an input image, we solve for a label map whose local label descriptors can be approximated as a sparse convex combination of exemplar label descriptors in the training data, where the sparsity is regularized by the similarity measure between the local feature descriptor of the input image and that of the exemplars in the training data set. Low-level image over-segmentation can be incorporated into our formulation to improve efficiency. Our formulation and algorithm compare favorably with the baseline method on the CamVid and Barcelona datasets.

[S8-P22B]

Road Scene Segmentation from a Single Image

Jose M. Alvarez, Theo Gevers, Yann LeCun, and Antonio M. Lopez

Road scene segmentation is important in computer vision for different applications such as autonomous driving and pedestrian detection. Recovering the 3D structure of road scenes provides relevant contextual information to improve their understanding. In this paper, we use a convolutional neural network based algorithm to learn features from noisy labels to recover the 3D scene layout of a road image. The novelty of the algorithm relies on generating training labels by applying an algorithm trained on a general image dataset to classify on--board images. Further, we propose a novel texture descriptor based on a learned color plane fusion to obtain maximal uniformity in road areas. Finally, acquired (off--line) and current (on--line) information are combined to detect road areas in single images. From quantitative and qualitative experiments, conducted on publicly available datasets, it is concluded that convolutional neural networks are suitable for learning 3D scene layout from noisy labels and provides a relative improvement of 7% compared to the baseline. Furthermore, combining color planes provides a statistical description of road areas that exhibits maximal uniformity and provides a relative improvement of 8% compared to the baseline. Finally, the improvement is even bigger when acquired and current information from a single image are combined.

[S8-P23B]

Efficient Recursive Algorithms for Computing the Mean Diffusion Tensor and Applications to DTI Segmentation

Guang Cheng, Hesamoddin Salehian, and Baba C. Vemuri

Computation of the mean of a collection of symmetric positive definite (SPD) matrices is a fundamental ingredient of many algorithms in diffusion tensor image (DTI) processing. For instance, in DTI segmentation, clustering, etc. In this paper, we present novel recursive algorithms for computing the mean of a set of diffusion tensors using several distance/divergence measures commonly used in DTI segmentation and clustering such as the Riemannian distance and symmetrized Kullback-Leibler divergence. To the best of our knowledge, to date, there are no recursive algorithms for computing the mean using these measures in literature. Recursive algorithms lead to a gain in computation time of several orders in magnitude over existing non-recursive algorithms. The key contributions of this paper are: (i) we present novel theoretical results on a recursive estimator for Karcher expectation in the space of SPD matrices, which in effect is a proof of the law of large numbers (with some restrictions) for the manifold of SPD matrices. (ii) We also present a recursive version of the symmetrized KL-divergence for computing the mean of a collection of SPD matrices. (iii) We present comparative timing results for computing the mean of a group of SPD matrices (diffusion tensors) depicting the gains in compute time using the proposed recursive algorithms over existing non-recursive counter parts. Finally, we also show results on gains in compute times obtained by applying these recursive algorithms to the task of DTI segmentation.

[S8-P24B]

Semi-Nonnegative Matrix Factorization for Motion Segmentation with Missing Data

Quanyi Mo and Bruce A. Draper

Motion segmentation is an old problem that is receiving renewed interest because of its role in video analysis. In this paper, we present a Semi-Nonnegative Matrix Factorization (SNMF)method that models dense point tracks in terms of their optical flow, and decomposes sets of point tracks into semantically meaningful motion components. We show that this formulation of SNMF with missing values outperforms the state-of-the-art algorithm of Brox and Malik in terms of accuracy on 10-frame video segments from the Berkeley test set, while being over 100 times faster. We then show how SNMF can be applied to longer videos using sliding windows. The result is competitive in terms of accuracy with Brox and Malik's algorithm, while still being two orders of magnitude faster.

ORAL SESSION 8
SEMANTIC SEGMENTATION AND OBJECT DISCOVERY

Thursday, October 11
17:05 - 18:30

[S8-O1]

A Three-Layered Approach to Facade Parsing

Andelo Martinovi, Markus Mathias, Julien Weissenberg, and Luc Van Gool

We propose a novel three-layered approach for semantic segmentation of building facades. In the first layer, starting from an oversegmentation of a facade, we employ the recently introduced machine learning technique Recursive Neural Networks (RNN) to obtain a probabilistic interpretation of each segment. In the second layer, initial labeling is augmented with the information coming from specialized facade component detectors. The information is merged using a Markov Random Field. In the third layer, we introduce weak architectural knowledge, which enforces the final reconstruction to be architecturally plausible and consistent. Rigorous tests performed on two existing datasets of building facades demonstrate that we significantly outperform the current-state of the art, even when using outputs from earlier layers of the pipeline. Also, we show how the final output of the third layer can be used to create a procedural reconstruction.

[S8-O2]

Semantic Segmentation with Second-Order Pooling

João Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu

Feature extraction, coding and pooling, are important components on many contemporary object recognition paradigms. In this paper we explore novel pooling techniques that encode the second-order statistics of local descriptors inside a region. To achieve this effect, we introduce multiplicative second-order analogues of average and max-pooling that together with appropriate non-linearities lead to state-of-the-art performance on free-form region recognition, without any type of feature coding. Instead of coding, we found that enriching local descriptors with additional image information leads to large performance gains, especially in conjunction with the proposed pooling methodology. We show that second-order pooling over free-form regions produces results superior to those of the winning systems in the Pascal VOC 2011 semantic segmentation challenge, with models that are 20,000 times faster.

[S8-O3]

Shape Sharing for Object Segmentation

Jaechul Kim and Kristen Grauman

We introduce a category-independent shape prior for object segmentation. Existing shape priors assume class-specific knowledge, and thus are restricted to cases where the object class is known in advance. The main insight of our approach is that shapes are often shared between objects of different categories. To exploit this "shape sharing" phenomenon, we develop a non-parametric prior that transfers object shapes from an exemplar database to a test image based on local shape matching. The transferred shape priors are then enforced in a graph-cut formulation to produce a pool of object segment hypotheses. Unlike previous multiple segmentation methods, our approach benefits from global shape cues; unlike previous top-down methods, it assumes no class-specific training and thus enhances segmentation even for unfamiliar categories. On the challenging PASCAL 2010 and Berkeley Segmentation datasets, we show it outperforms the state-of-the-art in bottom-up or category-independent segmentation.

[S8-O4]

Segmentation Propagation in ImageNet

Daniel Kuettel, Matthieu Guillaumin, and Vittorio Ferrari

ImageNet is a large-scale hierarchical database of object classes. We propose to automatically populate it with pixelwise segmentations, by leveraging existing manual annotations in the form of class labels and bounding-boxes. The key idea is to recursively exploit images segmented so far to guide the segmentation of new images. At each stage this propagation process expands into the images which are easiest to segment at that point in time, e.g. by moving to the semantically most related classes to those segmented so far. The propagation of segmentation occurs both (a) at the image level, by transferring existing segmentations to estimate the probability of a pixel to be foreground, and (b) at the class level, by jointly segmenting images of the same class and by importing the appearance models of classes that are already segmented. Through an experiment on 577 classes and 500k images we show that our technique (i) annotates a wide range of classes with accurate segmentations; (ii) effectively exploits the hierarchical structure of ImageNet; (iii) scales efficiently; (iv) outperforms a baseline GrabCut [1] initialized on the image center, as well as our recent segmentation transfer technique [2] on which this paper is based. Moreover, our method also delivers state-of-the-art results on the recent iCoseg dataset for co-segmentation.

[S8-O5]

"Clustering by Composition" for Unsupervised Discovery of Image Categories

Alon Faktor and Michal Irani

We define a "good image cluster" as one in which images can be easily composed (like a puzzle) using pieces from each other, while are difficult to compose from images outside the cluster. The larger and more statistically significant the pieces are, the stronger the affinity between the images. This gives rise to unsupervised discovery of very challenging image categories. We further show how multiple images can be composed from each other simultaneously and efficiently using a collaborative randomized search algorithm. This collaborative process exploits the "wisdom of crowds of images", to obtain a sparse yet meaningful set of image affinities, and in time which is almost linear in the size of the image collection. "Clustering-by-Composition" can be applied to very few images (where a 'cluster model' cannot be 'learned'), as well as on benchmark evaluation datasets, and yields state-of-the-art results.

NOTES