ECCV 2012

Errata – Abstracts

Here are collected the errata of the printed booklet.

[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-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.

[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-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-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.

[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-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-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-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.

[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.

[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-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-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-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.

[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.

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-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.

[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-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-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.