Floor Fields for Tracking in High Density Crowd Scenes

Saad Ali and Mubarak Shah
Computer Vision Lab, University of Central Florida, Orlando, USA
{sali,shah }@eecs.ucf.edu

Abstract. This paper presents an algorithm for tracking individual targets in high density crowd scenes containing hundreds of people. Tracking in such a scene is extremely challenging due to the small number of pixels on the target, appearance ambiguity resulting from the dense packing, and severe inter-object occlusions. The novel tracking algorithm, which is outlined in this paper, will overcome these challenges using a scene structure based force model . In this force model an individual, when moving in a particular scene, is subjected to global and local forces that are functions of the layout of that scene and the locomotive behavior of other individuals in the scene. The key ingredients of the force model are three floor fields, which are inspired by the research in the field of evacuation dynamics, namely Static Floor Field (SFF), Dynamic Floor Field (DFF), and Boundary Floor Field (BFF). These fields determine the probability of move from one location to another by converting the long-range forces into local ones. The SFF specifies regions of the scene which are attractive in nature (e.g. an exit location). The DFF specifies the immediate behavior of the crowd in the vicinity of the individual being tracked. The BFF specifies influences exhibited by the barriers in the scene (e.g. walls, no-go areas). By combining cues from all three fields with the available appearance information, we track individual targets in high density crowds.

1

****

The Naked Truth: Estimating Body Shape Under Clothing

Alexandru O. Bălan and Michael J. Black
Department of Computer Science, Brown University, Providence, RI 02912, USA
{alb,black }@cs.brown.edu

Abstract. We propose a method to estimate the detailed 3D shape of a person from images of that person wearing clothing. The approach exploits a model of human body shapes that is learned from a database of over 2000 range scans. We show that the parameters of this shape model can be recovered independently of body pose. We further propose a generalization of the visual hull to account for the fact that observed silhouettes of clothed people do not provide a tight bound on the true 3D shape. With clothed subjects, different poses provide different constraints on the possible underlying 3D body shape. We consequently combine constraints across pose to more accurately estimate 3D body shape in the presence of occluding clothing. Finally we use the recovered 3D shape to estimate the gender of subjects and then employ gender-specific body models to refine our shape estimates. Results on a novel database of thousands of images of clothed and “naked” subjects, as well as sequences from the HumanEva dataset, suggest the method may be accurate enough for biometric shape analysis in video.

1

****

Temporal Surface Tracking Using Mesh Evolution

Kiran Varanasi, Andrei Zaharescu, Edmond Boyer, and Radu Horaud
L J K-I N R I AR h ˆ one-Alpes, France

Abstract. In this paper, we address the problem of surface tracking in multiple camera environments and over time sequences. In order to fully track a surface undergoing significant deformations, we cast the problem as a mesh evolution over time. Such an evolution is driven by 3D displacement fields estimated between meshes recovered independently at different time frames. Geometric and photometric information is used to identify a robust set of matching vertices. T his provides a sparse displacement field that is densified over the mesh by Laplacian diffusion. In contrast to existing approaches that evolve meshes, we do not assume a known model or a fixed topology. The contribution is a novel mesh evolution based framework that allows to fully track, over long sequences, an unknown surface encountering deformations, including topological changes. Results on very challenging and publicly available image based 3D mesh sequences demonstrate the ability of our framework to efficiently recover surface motions .

1

****

Grassmann Registration Manifolds for Face Recognition■

Yui Man Lui and J. Ross Beveridge
Department of Computer Science, Colorado State University,
Fort Collins, CO 80523, USA
{lui,ross }@cs.colostate.edu

Abstract. Motivated by image perturbation and the geometry of manifolds, we present a novel method combining these two elements. First, we form a tangent space from a set of perturbed images and observe that the tangent space admits a vector space structure. Second, we embed the approximated tangent spaces on a Grassmann manifold and employ a chordal distance as the means for comparing subspaces. The matching process is accelerated using a coarse to fine strategy. Experiments on the FERET database suggest that the proposed method yields excellent results using both holistic and local features. Specifically, on the FERET Dup2 data set, our proposed method achieves 83 .8% rank 1 recognition: to our knowledge the currently the best result among all non-trained methods. Evidence is also presented that peak recognition performance is achieved using roughly 100 distinct perturbed images.

1

****

Facial Expression Recognition Based on 3D Dynamic Range Model Sequences

Yi Sun and Lijun Yin
Department of Computer Science, State University of New York at Binghamton
Binghamton, New York, 13902 USA

Abstract. Traditionally, facial expression recognition (FER) issues have been studied mostly based on modalities of 2D images, 2D videos, and 3D static models. In this paper, we propose a spatio-temporal expression analysis approach based on a new modality, 3D dynamic geometric facial model sequences, to tackle the FER problems. Our approach integrates a 3D facial surface descriptor and Hidden Markov Models (HMM) to recognize facial expressions. To study the dynamics of 3D dynamic models for FER, we investigated three types of HMMs: temporal 1D-HMM, pseudo 2D-HMM (a combination of a spatial HMM and a temporal HMM), and real 2D-HMM. We also created a new dynamic 3D facial expression database for the research community. The results show that our approach achieves a 90.44% person-independent recognition rate for distinguishing six prototypic facial expressions. The advantage of our method is demonstrated as compared to methods based on 2D texture images, 2D/3D Motion Units, and 3D static range models. Further experimental evaluations also verify the benefits of our approach with respect to partial facial surface occlusion, expression intensity changes, and 3D model resolution variations.

1

****

Face Alignment Via Component-Based Discriminative Search

Lin Liang, Rong Xiao, Fang Wen, and Jian Sun
Microsoft Research Asia
Beijing, China
{lliang,rxiao,fangwen,jiansun }@microsoft.com

Abstract. In this paper, we propose a component-based discriminative approach for face alignment without requiring initialization1. Unlike many approaches which locally optimize in a small range, our approach searches the face shape in a large range at the component level by a discriminative search algorithm. Specifically, a set of direction classifiers guide the search of the configurations of facial components among multiple detected modes of facial components. The direction classifiers are learned using a large number of aligned local patches and misaligned local patches from the training data. The discriminative search is extremely effective and able to find very good alignment results only in a few (2 ~ 3) search iterations. As the new approach gives excellent alignment results on the commonly used datasets (e.g., AR [18], FERET [21]) created under-controlled conditions, we evaluate our approach on a more challenging dataset containing over 1,700 well-labeled facial images with a large range of variations in pose, lighting, expression, and background. The experimental results show the superiority of our approach on both accuracy and efficiency.

1

****

Improving People Search Using Query Expansions

How Friends Help to Find People

Thomas Mensink and Jakob Verbeek
LEAR - INRIA Rhône Alpes - Grenoble, France
{thomas.mensink,jakob.verbeek }@inria.fr

Abstract. In this paper we are interested in finding images of people on the web, and more specifically within large databases of captioned news images. It has recently been shown that visual analysis of the faces in images returned on a text-based query over captions can significantly improve search results. The underlying idea to improve the text-based results is that although this initial result is imperfect, it will render the queried person to be relatively frequent as compared to other people, so we can search for a large group of highly similar faces.

The performance of such methods depends strongly on this assumption: for people whose face appears in less than about 40% of the initial text-based result, the performance may be very poor. The contribution of this paper is to improve search results by exploiting faces of other people that co-occur frequently with the queried person. We refer to this process as ‘query expansion’. In the face analysis we use the query expansion to provide a query-specific relevant set of ‘negative’ examples which should be separated from the potentially positive examples in the text-based result set. We apply this idea to a recently-proposed method which filters the initial result set using a Gaussian mixture model, and apply the same idea using a logistic discriminant model. We experimentally evaluate the methods using a set of 23 queries on a database of 15.000 captioned news stories from Yahoo! News . The results show that (i) query expansion improves both methods, (ii) that our discriminative models outperform the generative ones, and (iii) our best results surpass the state-of-the-art results by 10% precision on average.

1

****

Fast Automatic Single-View 3-d Reconstruction of Urban Scenes

Olga Barinova[1], Vadim Konushin[2], Anton Yakubenko[1],
KeeChang Lee[3], Hwasup Lim[3], and Anton Konushin[1]
[1]Moscow State University, Department of Computational Mathematics and Cybernetics, Graphics & Media Lab
{obarinova,toh,ktosh }@graphics.cs.msu.ru
[2]The Keldysh Institute of Applied Mathematic Russian Academy of Sciences
vadim@graphics.cs.msu.ru
[3]Samsung Advanced Institute of Technology
{lkc.lee,hwasup.lim }@samsung.com

Abstract. We consider the problem of estimating 3-d structure from a single still image of an outdoor urban scene. Our goal is to efficiently create 3-d models which are visually pleasant. We chose an appropriate 3-d model structure and formulate the task of 3-d reconstruction as model fitting problem. Our 3-d models are composed of a number of vertical walls and a ground plane, where ground-vertical boundary is a continuous polyline. We achieve computational efficiency by special preprocessing together with stepwise search of 3-d model parameters dividing the problem into two smaller sub-problems on chain graphs. The use of Conditional Random Field models for both problems allows to various cues. We infer orientation of vertical walls of 3-d model vanishing points.

1

****

Fourier Analysis of the 2D Screened Poisson Equation for Gradient Domain Problems

Pravin Bhat[1], Brian Curless[1], Michael Cohen[1,2], and C. Lawrence Zitnick[1]
[1]University of Washington
[2]Microsoft Research

Abstract. We analyze the problem of reconstructing a 2D function that approximates a set of desired gradients and a data term. The combined data and gradient terms enable operations like modifying the gradients of an image while staying close to the original image. Starting with a variational formulation, we arrive at the “screened Poisson equation” known in physics. Analysis of this equation in the Fourier domain leads to a direct, exact, and efficient solution to the problem. Further analysis reveals the structure of the spatial filters that solve the 2D screened Poisson equation and shows gradient scaling to be a well-defined sharpen filter that generalizes Laplacian sharpening, which itself can be mapped to gradient domain filtering. Results using a DCT-based screened Poisson solver are demonstrated on several applications including image blending for panoramas, image sharpening, and de-blocking of compressed images.

1

****

Anisotropic Geodesics for Perceptual Grouping and Domain Meshing■

Sébastien Bougleux, Gabriel Peyré, and Laurent Cohen
Université Paris-Dauphine, CEREMADE, 75775 Paris Cedex 16, France
{bougleux,peyre,cohen }@ceremade.dauphine.fr

Abstract. This paper shows how Voronoi diagrams and their dual Delaunay complexes, defined with geodesic distances over 2D Reimannian manifolds, can be used to solve two important problems encountered in computer vision and graphics. The first problem studied is perceptual grouping which is a curve reconstruction problem where one should complete in a meaningful way a sparse set of noisy curves. From this latter curves, our grouping algorithm first designs an anisotropic tensor field that corresponds to a Reimannian metric. Then, according to this metric, the Delaunay graph is constructed and pruned in order to correctly link together salient features. The second problem studied is planar domain meshing, where one should build a good quality triangulation of a given domain. Our meshing algorithm is a geodesic Delaunay refinement method that exploits an anisotropic tensor field in order to locally impose the orientation and aspect ratio of the created triangles.

1

****

Regularized Partial Matching of Rigid Shapes

Alexander M. Bronstein and Michael M. Bronstein
Technion – Israel Institute of Technology, Haifa 32000, Israel
{bron,mbron }@cs.technion.ac.il

Abstract. Matching of rigid shapes is an important problem in numerous applications across the boundary of comput er vision, pattern recognition and computer graphics communities. A particularly challenging setting of this problem is partial matching, where the two shapes are dissimilar in general, but have significant similar parts. In this paper, we show a rigorous approach allowing to find matching parts of rigid shapes with controllable size and regularity. The regularity term we use is similar to the spirit of the Mumford-Shah functional, extended to non-Euclidean spaces. Numerical experiments show that the regularized partial matching produces better results compared to the non-regularized one.

1

****

Compressive Sensing for Background Subtraction

V olkan Cevher1, Aswin Sankaranarayanan2, Marco F. Duarte1, Dikpal Reddy2,
Richard G. Baraniuk1, and Rama Chellappa2
1Rice University, ECE, Houston TX 77005
2University of Maryland, UMIACS, College Park, MD 20947

Abstract. Compressive sensing (CS) is an emerging field that provides a framework for image recovery using sub-Nyquist sampling rates. The CS theory shows that a signal can be reconstructed from a small set of random projections, provided that the signal is sparse in some basis, e.g., wavelets. In this paper, we describe a method to directly recover background subtracted images using CS and discuss its applications in some communication constrained multi-camera computer vision problems. We show how to apply the CS theory to recover object silhouettes (binary background subtracted images) when the objects of interest occupy a small portion of the camera view, i.e., when they are sparse in the spatial domain. We cast the background subtraction as a sparse approximation problem and provide different solutions based on convex optimization and total variation. In our method, as opposed to learning the background, we learn and adapt a low dimensional compressed representation of it, which is sufficient to determine spatial innovations; object silhouettes are then estimated directly using the compressive samples without any auxiliary image reconstruction. We also discuss simultaneous appearance recovery of the objects using compressive measurements. In this case, we show that it may be necessary to reconstruct one auxiliary image. To demonstrate the pe rformance of the proposed algorithm, we provide results on data captured using a compressive single-pixel camera. We also illustrate that our approach is suitable for image coding in communication constrained problems by using data captured by multiple conventional cameras to provide 2D tracking and 3D shape reconstruction results with compressive measurements.

1

****

Robust 3D Pose Estimation and Efficient 2D
Region-Based Segmentation from a 3D Shape Prior

Samuel Dambreville, Romeil Sandhu, Anthony Yezzi, and Allen Tannenbaum
Georgia Institute of Technology

Abstract. In this work, we present an approach to jointly segment a rigid object in a 2D image and estimate its 3D pose, using the knowledge of a 3D model. We naturally couple the two processes together into a unique energy functional that is minimized through a variational approach. Our methodology differs from the standard monocular 3D pose estimation algorithms since it does not rely on local image features. Instead, we use global image statistics to drive the pose estimation process. This confers a satisfying level of robustness to noise and initialization for our algorithm, and bypasses the need to establish correspondences between image and object features. Moreover, our methodology possesses the typical qualities of region-based active contour techniques with shape priors, such as robustness to occlusions or missing information, without the need to evolve an infinite dimensional curve. Another novelty of the proposed contribution is to use a unique 3D model surface of the object, instead of learning a large collection of 2D shapes to accommodate for the diverse aspects that a 3D object can take when imaged by a camera. Experimental results on both synthetic and real images are provided, which highlight the robust performance of the technique on challenging tracking and segmentation applications.

1 Motivation and Related Work

2D image segmentation and 2D-3D pose estimation are ubiquitous tasks in computer vision applications and have received a great deal of attention in the past few years.

These two fundamental techniques are usually studied separately in the literature. In this work, we combine both approaches in a variational framework. To appreciate the contribution of this work, we recall some of the results and specifics of both fields.

2D-3D pose estimation aims at determining the pose of a 3D object relative to a calibrated camera from one unique or a collection of 2D images. By knowing the mapping between the world coordinates and image coordinates from the camera calibration matrix, and after establishing correspondences between 2D features in the image and their 3D counterparts on the model, it is then possible to solve the pose transformation (from a set of equations that express these correspondences). The literature concerned with 3D pose estimation is very large and a complete survey is beyond the scope of this paper. However, most methods can be distinguished by the type of local image features used to establish correspondences, such as points [1], lines or segments [2,3], multi-part curve segments [4], or complete contours [5,6].

Segmentation consists of separating an object from the background in an image. The geometric active contour (GAC) framework, in which a curve is evolved continuously to

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part II, LNCS 5303, pp. 169–182, 2008.

****

Linear Time Maximally Stable Extremal Regions

David Nist´ er and Henrik Stew´ enius

[1]Microsoft Live Labs

[2]Google Switzerland

Abstract. In this paper we present a new algorithm for computing Maximally Stable Extremal Regions (MSER), as invented by Matas et al. The standard algorithm makes use of a union-find data structure and takes quasi-linear time in the number of pixels. The new algorithm provides exactly identical results in true worst-case linear time. Moreover, the new algorithm uses significantly less memory and has better cache-locality, resulting in faster execution. Our CPU implementation performs twice as fast as a state-of-the-art FPGA implementation based on the standard algorithm.

The new algorithm is based on a different computational ordering of the pixels, which is suggested by another immersion analogy than the one corresponding to the standard connected-component algorithm. With the new computational ordering, the pixels considered or visited at any point during computation consist of a single connected component of pixels in the image, resembling a flood-fill that adapts to the grey-level landscape. The computation only needs a priority queue of candidate pixels (the boundary of the single connected component), a single bit image masking visited pixels, and information for as many components as there are grey-levels in the image. This is substantially more compact in practice than the standard algorithm, where a large number of connected components must be considered in parallel. The new algorithm can also generate the component tree of the image in true linear time. The result shows that MSER detection is not tied to the union-find data structure, which may open more possibilities for parallelization.

1

****

Efficient Edge-Based Methods for Estimating Manhattan Frames in Urban Imagery

Patrick Denis[1], James H. Elder[1], and Francisco J. Estrada[2]

[1]York University
{pdenis,jelder }@yorku.ca

[2]University of Toronto
strider@cs.utoronto.ca

Abstract. We address the problem of efficiently estimating the rotation of a camera relative to the canonical 3D Cartesian frame of an urban scene, under the so-called “Manhattan World” assumption [1,2]. While the problem has received considerable attention in recent years, it is unclear how current methods stack up in terms of accuracy and efficiency, and how they might best be improved. It is often argued that it is best to base estimation on all pixels in the image [2]. However, in this paper, we argue that in a sense, less can be more: that basing estimation on sparse, accurately localized edges, rather than dense gradient maps, permits the derivation of more accurate statistical models and leads to more efficient estimation. We also introduce and compare several different search techniques that have advantages over prior approaches. A cornerstone of the paper is the establishment of a new public groundtruth database which we use to derive required statistics and to evaluate and compare algorithms.

1

****

Multiple Component Learning for Object Detection

Piotr Dollár[1,2], Boris Babenko[2], Serge Belongie[1,2], Pietro Perona[1], and Zhuowen Tu[3]

[1]Electrical Engineering California Inst. of Tech.
{pdollar,perona }@caltech.edu

[2]Comp. Science & Eng. Univ. of CA, San Diego
{bbabenko,sjb }@cs.ucsd.edu

3Lab of Neuro Imaging Univ. of CA, Los Angeles
zhuowen.tu@loni.ucla.edu

Abstract. Object detection is one of the key problems in computer vision. In the last decade, discriminative learning approaches have proven effective in detecting rigid objects, achieving very low false positives rates. The field has also seen a resurgence of part-based recognition methods, with impressive results on highly articulated, diverse object categories. In this paper we propose a discriminative learning approach for detection that is inspired by part-based recognition approaches. Our method, Multiple Component Learning (mcl), automatically learns individual component classifiers and combines these into an overall classifier. Unlike previous methods, which rely on either fairly restricted part models or labeled part data, mcllearns powerful component classifiers in a weakly supervised manner, where object labels are provided but part labels are not. The basis of mcllies in learning a set classifier; we achieve this by combining boosting with weakly supervised learning, specifically the Multiple Instance Learning framework (mil).mclis general, and we demonstrate results on a range of data from computer audition and computer vision. In particular, mcloutperforms all existing methods on the challenging INRIA pedestrian detection dataset, and unlike methods that are not part-based, mclis quite robust to occlusions.

1

****

A Probabilistic Approach to Integrating Multiple Cues in Visual Tracking

Wei Du and Justus Piater
University of Li` ege, Department of Electrical Engineering and Computer Science
Montefiore Institute, B28, B-4000 Liege, Belgium
{wei.du,justus.piater }@ulg.ac.be

Abstract. This paper presents a novel probabilistic approach to integrating multiple cues in visual tracking. We perform tracking in different cues by interacting processes. Each process is represented by a Hidden Markov Model, and these parallel pro cesses are arranged in a chain topology. The resulting Linked Hidden Markov Models naturally allow the use of particle filters and Belief Propagation in a unified framework. In particular, a target is tracked in each cue by a particle filter, and the particle filters in different cues interact via a message passing scheme. The general framework of our approach allows a customized combination of different cues in different situations, which is desirable from the implementation point of view. Our examples selectively integrate four visual cues including color, edges, motion and contours. We demonstrate empirically that the ordering of the cues is nearly inconsequential, and that our approach is superior to other approaches such as Independent Integration and Hierarchical Integration in terms of flexibility and robustness.

1

****

Fast and Accurate Rotation Estimation on the 2-Sphere without Correspondences

Janis Fehr, Marco Reisert, and Hans Burkhardt
Chair of Pattern Recognition and Image Processing
Institute for Computer Science
Albert-Ludwigs-University, Freiburg, Germany
fehr@informatik.uni-freiburg.de

Abstract. We present a refined method for rotation estimation of signals on the 2-Sphere. Our approach utilizes a fast correlation in the harmonic domain to estimate rotation angles of arbitrary size and resolution. The method is able to achieve great accuracy even for very low spherical harmonic expansions of the input signals without using correspondences or any other kind of a priori information. The rotation parameters are computed analytically without additional iterative post-

processing or “fine tuning”.
The theoretical advances presented in this paper can be applied to a wide range of practical problems such as: shape description and shape retrieval, 3D rigid registration, robot positioning with omni-directional cameras or 3D invariant feature design.

1

****

A Lattice-Preserving Multig rid Method for Solving the Inhomogeneous Poisson Equations Used in Image Analysis

Leo Grady
Siemens Corporate Research
Department of Imaging and Visualization
755 College Road East
Princeton, NJ 08540

Abstract. The inhomogeneous Poisson (Laplace) equation with internal Dirichlet boundary conditions has recently appear ed in several applications ranging from image segmentation [1, 2, 3] to image colorization [4], digital photo matting [5, 6] and image filtering [7, 8]. In addition, the problem we address may also be considered as the generalized eigenvector problem associated with Normalized Cuts [9], the linearized anisotropic diffusion problem [10, 11, 8] solved with a backward Euler method, visual surface reconstruction with discontinuities [12, 13] or opti cal flow [14]. Although these appr oaches have demonstrated quality results, the computational burden of finding a solution requires an efficient solver. Design of an efficient multigrid solver is difficult for these problems due to unpredictable inhomogeneity in the equation coefficients and internal Dirichlet boundary conditions with unpredictable loca tion and value. Previous approaches to multigrid solvers have typically employed either a data-driven operator (with fast convergence) or the maintenance of a lattice structure at coarse levels (with low memory overhead). In addition to memory efficiency, a lattice structure at coarse levels is also essential to taking advantage of the power of a GPU implementation [15,16,5,3]. In this work, we present a multigrid method that maintains the low memory overhead (and GPU suitability) associated with a regular lattice while benefiting from the fast convergence of a data-driven coarse operator.

1

****

SMD: A Locally Stable Monotonic Change Invariant Feature Descriptor

Raj Gupta and Anurag Mittal
Indian Institute of Tec hnology, Madras, India
gupta.raj@gmail.com, amittal@cse.iitm.ernet.in

Abstract. Extraction and matching of discriminative feature points in images is an important problem in computer vision with applications in image classification, object recognition, mosaicing, automatic 3D reconstruction and stereo. Features are represented and matched via descriptors that must be invariant to small errors in the localization and scale of the extracted feature point, viewpoint changes, and other kinds of changes such as illumination, image compression and blur. While currently used feature descr iptors are able to deal with many of such changes, they are not invariant to a ge neric monotonic change in the intensities, which occurs in many cases. Furthermore, their performance degrades rapidly with many image degradations such as blur and compression where the intensity transformation is non-linear. In this paper, we present a new feature descriptor that obtains invariance to a monotonic change in the intensity of the patch by looking at orders between certain pixels in the patch. An order change between pixels indicates a difference between the patches which is penalized. Summation of such penalties over carefully chosen pixel pairs that are stable to small errors in their localization and are independent of each other leads to a robust measure of change between two features. Promising results were obtained using this approach that show significant improvement over existing methods, especially in the case of illumination change, blur and JPEG compression where the intensity of the points changes from one image to the next.

1

****

Finding Actions Using Shape Flows

Hao Jiang and David R. Martin
Computer Science Department, Bost on College, Chestnut Hill, MA 02467, USA
{hjiang,dmartin }@cs.bc.edu

Abstract. We propose a novel method for action detection based on a new action descriptor called a shape flow that represents both the shape and movement of an object in a holistic and parsimonious manner. We find actions by finding shape flows in a target video that are similar to a template shape flow. Shape flows are largely independent of appearance, and the match cost function that we propose is invariant to scale changes and smooth nonlinear deformation in space and time. The problem of matching shape flows is difficult, however, yielding a large, non-convex, integer program. We propose a novel relaxation method based onsuccessive convexification that converts this hard program into a vastly smaller linear program: By using only those variables that appear on the 4D lower convex hull of the matching cost volume, most of the variables in the linear program may be eliminated. Experiments confirm that the proposed shape flow method can successfully detect complex actions in cluttered video, even with self-occlusion, camera motion, and intra-class variation.

1

****

Cross-View Action Recognition from Temporal Self-similarities

Imran N. Junejo, Emilie Dexter, Ivan Laptev, and Patrick P´ erez
INRIA Rennes - Bretagne Atlantique
35042 Rennes Cedex - France

Abstract. This paper concerns recognition of human actions under view changes. We explore self-similarities of action sequences over time and observe the striking stability of such measures across views. Building upon this key observation we develop an action descriptor that captures the structure of temporal similarities and dissimilarities within an action sequence. Despite this descriptor not being strictly view-invariant, we provide intuition and experimental validation demonstrating the high stability of self-similariti es under view changes. Self- similarity descriptors are also shown stable under action variations within a class as well as discriminative for action recognition. Interestingly, self-similarities computed from different image features possess similar properties and can be used in a complementary fashion. Our method is simple and requires neither structure recovery nor multi-view correspondence estimation. Instead, it relies on weak geometric properties and combines them with machine learning for efficient cross-view action recognition. The method is validated on three public datasets, it has similar or superior performance compared to related methods and it performs well even in extreme conditions such as when recognizing actions from top views while using side views for training only.

1

****

Window Annealing over Square Lattice Markov Random Field

Ho Yub Jung, Kyoung Mu Lee, and Sang Uk Lee
Department of EECS, ASRI, Seoul National University, 151-742, Seoul, Korea
hoyub@diehard.snu.ac.kr, kyoungmu@snu.ac.kr, sanguk@ipl.snu.ac.kr

Abstract. Monte Carlo methods and their subsequent simulated annealing are able to minimize general energy functions. However, the slow convergence of simulated annealin g compared with more recent deterministic algorithms such as graph cuts and belief propagation hinders its popularity over the large dimensional Markov Random Field (MRF). In this paper, we propose a new efficient sampling-based optimization algorithm called WA (Window Annealing) over squared lattice MRF, in which cluster sampling and annealing concepts are combined together. Unlike the conventional annealing process in which only the temperature variable is scheduled, we design a series of artificial ”guiding” (auxiliary) probability distributions based on the general sequential Monte Carlo framework. These auxiliary distributions lead to the maximum a posteriori (MAP) state by scheduling both the temperature and the proposed maximum size of the windows (rectangular cluster) variable. This new annealing scheme greatly enhances the mixing rate and consequently reduces convergence time. Moreover, by adopting the integral image technique for computation of the proposal probability of a sampled window, we can achieve a dramatic reduction in overall computations. The proposed WA is compared with several existing Monte Carlo based optimization techniques as well as state-of-the-art deterministic methods including Graph Cut (GC) and sequential tree re-weighted belief propagation (TRW-S) in the pairwise MRF stereo problem. The experimental results demonstrate that the propo sed WA method is comparable with GC in both speed and obtained energy level.

1

****

Unsupervised Classification and Part Localization by Consistency Amplification

Leonid Karlinsky, Michael Dinerstein, Dan Levi, and Shimon Ullman
Weizmann Institute of Science, Rehovot 76100, Israel
{leonid.karlinsky,michael.dinerstein,dan.levi,
shimon.ullman }@weizmann.ac.il

Abstract. We present a novel method for unsupervised classification, including the discovery of a new category and precise object and part localization. Given a set of unlabelled images, some of which contain an object of an unknown category, with unknown location and unknown size relative to the background, the method automatically identifies the images that contain the objects, localizes them and their parts, and reliably learns their appearance and geometry for subsequent classification. Current unsupervised methods construct classifiers based on a fixed set of initial features. Instead, we propose a new approach which iteratively extracts new features and re-learns the induced classifier, improving class vs. non-class separation at each iteration. We develop two main tools that allow this iterative combined search. The first is a novel star-like model capable of learning a geometric class representation in the unsupervised setting. The second is learning of ”part specific features” that are optimized for parts detection, and which optimally combine different part appearances discovered in the training examples. These novel aspects lead to precise part localization and to improvement in overall classification performance compared with previous methods. We applied our method to multiple object classes from Caltech-101, UIUC and a

sub-classification problem from PASCAL. The obtained results are comparable to state-of-the-art supervised classification techniques and superior to state-of-the-art unsupervised approaches previously applied to the same image sets.

1

****

Simultaneous Visual Recognition of Manipulation Actions and Manipulated Objects

Hedvig Kjellstr¨ om, Javier Romero, David Mart´ ■nez, and Danica Kragi´ c
Computational Vision and Active Perception Lab
School of Computer Science and Communication
KTH, SE-100 44 Stockholm, Sweden
{hedvig,jrgn,davidmm,dani }@kth.se

Abstract. The visual analysis of human manipulation actions is of interest for e.g. human-robot interaction applications where a robot learns how to perform a task by watching a human. In this paper, a method forclassifying manipulation actions in the context of the objects manipulated ,a n dclassifying objects in the context of the actions used to manipulate them is presented. Hand and object features are extracted from the video sequence using a segmentation based approach. A shape based representation is used for both the hand and the object. Experiments show this representation suitable for representing generic shape classes. The action-object correlation over time is then modeled using conditional random fields. Experimental comparison show great improvement in classification rate when the action-object correlation is taken into account, compared to separate classification of manipulation actions and manipulated objects.

1

****

Active Contour Based Segmentation of 3D Surfaces

Matthias Krueger, Patrice Delmas, and Georgy Gimel’farb
Dept. of Computer Science, Tamaki Campus
The University of Auckland, Auckland, New Zealand
mkru007@aucklanduni.ac.nz

Abstract. Algorithms incorporating 3D information have proven to be superior to purely 2D approaches in many areas of computer vision including face biometrics and recognition. Still, the range of methods for feature extraction from 3D surfaces is limited. Very popular in 2D image analysis, active contours have been generalized to curved surfaces only recently. Current implementations require a global surface parametrisation. We show that a balloon force cannot be included properly in existing methods, making them unsuitable for applications with noisy data. To overcome this drawback we propose a new algorithm for evolving geodesic active contours on implicit surfaces. We also introduce a new narrowband scheme which results in linear computational complexity. The performance of our model is illustrated on various real and synthetic 3D surfaces.

1

****

What Is a Good Nearest Neighbors Algorithm for Finding Similar Patches in Images?

Neeraj Kumar1,■, Li Zhang2, and Shree Nayar1
1Columbia University
2University of Wisconsin-Madison

Abstract. Many computer vision algorithms require searching a set of images for similar patches, which is a very expensive operation. In this work, we compare and evaluate a number of nearest neighbors algorithms for speeding up this task. Since image patches follow very different distributions from the uniform and Gaussian distributions that are typically used to evaluate nearest neighbors methods, we determine the method with the best performance via extensive experimentation on real images. Furthermore, we take advantage of the inherent structure and properties of images to achieve highly efficient implementations of these algorithms. Our results indicate that vantage point trees, which are not well known in the vision community, generally offer the best performance.

1

****

Learning for Optical Flow Using Stochastic Optimization

Yunpeng Li and Daniel P. Huttenlocher
Department of Computer Science, Cornell University, Ithaca, NY 14853
{yuli,dph }@cs.cornell.edu

Abstract. We present a technique for learning the parameters of a continuous-state Markov random field (MRF) model of optical flow, by minimizing the training loss for a set of ground-truth images using simultaneous perturbation stochastic approximation (SPSA). The use of SPSA to directly minimize the training loss offers several advantages over most previous work on learning MRF models for low-level vision, which instead seek to maximize the likelihood of the data given the model parameters. In particular, our approach explicitly optimizes the error criterion used to evaluate the quality of the flow field, naturally handles missing data values in the ground truth, and does not require the kinds of approximations that current methods use to address the intractable nature of maximum-likelihood estimation for such problems. We show that our method achieves state-of-the-art results and requires only a very small number of training images. We also find that our method generalizes well to unseen data, including data with quite different characteristics than the training set.

1

****

Region-Based 2D Deformable Generalized Cylinder for Narrow Structures Segmentation

Julien Mille[1], Romuald Boné[1], and Laurent D. Cohen[2]
[1]Laboratoire d'Informatique, Université François Rabelais de Tours
64 avenue Jean Portalis, 37200 Tours, France
[2]CEREMADE, CNRS UMR 7534, Université Paris Dauphine
Place du Maréchal de Lattre de Tassigny, 75775 Paris, France
julien.mille@univ-tours.fr

Abstract. In this paper, we present a region-based deformable cylinder model, extending the work on classical region-based active contours and gradient-based ribbon snakes. Defined by a central curve playing the role of the medial axis and a variable thickness, the model is endowed with a region-dependent term.This energy follows the narrow band principle, in order to handle local region properties while overcoming limitations of classical edge-based models. The energy is subsequently transformed and derived in order to allow implementation on a polygonal line deformed with gradient descent. The model is used to extract path-like objects in medical and aerial images.

1

****

Pose Priors for Simultaneously Solving Alignment and Correspondence■

Francesc Moreno-Noguer, Vincent Lepetit, and Pascal Fua
Computer Vision Laboratory
École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland

Abstract. Estimating a camera pose given a set of 3D-object and 2D-image feature points is a well understood problem when correspondences are given. However, when such correspondences cannot be established a priori, one must simultaneously compute them along with the pose. Most current approaches to solving this problem are too computationally intensive to be practical. An interesting exception is the SoftPosit algorithm, that looks for the solution as the minimum of a suitable objective function. It is arguably one of the best algorithms but its iterative nature means it can fail in the presence of clutter, occlusions, or repetitive patterns. In this paper, we propose an approach that overcomes this limitation by taking advantage of the fact that, in practice, some prior on the camera pose is often available. We model it as a Gaussian Mixture Model that we progressively refine by hypothesizing new correspondences. This rapidly reduces the number of potential matches for each 3D point and lets us explore the pose space more thoroughly than SoftPosit at a similar computational cost. We will demonstrate the superior performance of our approach on both synthetic and real data.

****

Latent Pose Estimator for Continuous Action Recognition

Huazhong Ning[1], Wei Xu[2], Yihong Gong[2], and Thomas Huang[1]
[1]ECE, U. of Illinois at Urbana-Champaign, USA
{hning2,huang }@ifp.uiuc.edu
[2]NEC Laboratories America, Inc., USA
{xw,ygong }@sv.nec-labs.com

Abstract. Recently, models based on conditional random fields (CRF) have produced promising results on labeling sequential data in several scientific fields. However, in the vision task of continuous action recognition, the observations of visual features have dimensions as high as hundreds or even thousands. This might pose severe difficulties on parameter estimation and even degrade the performance. To bridge the gap between the high dimensional observations and the random fields, we propose a novel model that replace the observation layer of a traditional random fields model with a latent pose estimator. In training stage, the human pose is not observed in the action data, and the latent pose estimator is learned under the supervision of the labeled action data, instead of image-to-pose data. The advantage of this model is twofold. First, it learns to convert the high dimensional observations into more compact and informative representations. Second, it enables transfer learning to fully utilize the existing knowledge and data on image-to-pose relationship. The parameters of the latent pose estimator and the random fields are jointly optimized through a gradient ascent algorithm. Our approach is tested on HumanEva [1] – a publicly available dataset. The experiments show that our approach can improve recognition accuracy over standard CRF model and its variations. The performance can be further significantly improved by using additional image-to-pose data for training. Our experiments also show that the model trained on HumanEva can generalize to different environment and human subjects.

****

Relevant Feature Selection for Human Pose Estimation and Localization in Cluttered Images

Ryuzo Okada and Stefano Soatto
Computer Science Department, University of California, Los Angeles
{okada,soatto }@cs.ucla.edu

Abstract. We address the problem of estimating human body pose from a single image with cluttered background. We train multiple local linear regressors for estimating the 3D pose from a feature vector of gradient orientation histograms.

Each linear regressor is capable of selecting relevant components of the feature vector depending on pose by training it on a pose cluster which is a subset of the training samples with similar pose. For discriminating the pose clusters, we use kernel Support Vector Machines (SVM) with pose-dependent feature selection. We achieve feature selection for kernel SVMs by estimating scale parameters of RBF kernel through minimization of the radius/margin bound, which is an upper bound of the expected generalization error, with efficient gradient descent. Human detection is also possible with these SVMs. Quantitative experiments show the effectiveness of pose-dependent feature selection to both human detection and pose estimation.

****

Determining Patch Saliency Using Low-Level Context

Devi Parikh[1], C. Lawrence Zitnick[2], and Tsuhan Chen[1]
[1]Carnegie Mellon University, Pittsburgh, PA, USA
[2]Microsoft Research, Redmond, WA, USA

Abstract. The increased use of context for high level reasoning has been popular in recent works to increase recognition accuracy. In this paper, we consider an orthogonal application of context. We explore the use of context to determine which low-level appearance cues in an image are salient or representative of an image’s contents. Existing classes of low-level saliency measures for image patches include those based on interest points, as well as supervised discriminative measures. We propose a new class of unsupervised contextual saliency measures based on co-occurrence and spatial information between image patches. For recognition, image patches are sampled using a weighted random sampling based on saliency, or using a sequential approach based on maximizing the likelihoods of the image patches. We compare the different classes of saliency measures, along with a baseline uniform measure, for the task of scene and object recognition using the bag-of-features paradigm. In our results, the contextual saliency measures achieve improved accuracies over the previous methods. Moreover, our highest accuracy is achieved using a sparse sampling of the image, unlike previous approaches who’s performance increases with the sampling density.

****

Edge-Preserving Smoothing and Mean-Shift Segmentation of Video Streams

Sylvain Paris
Adobe Systems, Inc.

Abstract. Video streams are ubiquitous in applications such as surveillance, games, and live broadcast. Processing and analyzing these data is challenging because algorithms have to be efficient in order to process the data on the fly. From a theoretical standpoint, vid eo streams have their own specificities – they mix spatial and temporal dimensions, and compared to standard video sequences, half of the information is missing, i.e.the future is unknown. The theoretical part of our work is motivated by the ubiquitous use of the Gaussian kernel in tools such as bilateral filtering and mean-shift segmentation. We formally derive its equivalent for video streams as well as a dedicated expression of isotropic diffusion. Building upon this theoretical ground, we adapt a number of classical algorithms to video streams: bilateral filtering, mean-shift segmentation, and anisotropic diffusion.

****

Deformed Lattice Discovery Via Efficient
Mean-Shift Belief Propagation

Minwoo Park1, Robert T. Collins1, and Yanxi Liu1,2
1Department of Computer Science and Engineering
2Department of Electrical Engineering
The Pennsylvania State University, University Park, PA 16802
{mipark,rcollins,yanxi }@cse.psu.edu

Abstract. We introduce a novel framework for automatic detection of repeated patterns in real images. The novelty of our work is to formulate the extraction of an underlying deformed lattice as a spatial, multi-target tracking problem using a new and efficient Mean-Shift Belief Propagation (MSBP) method. Compared to existing work, our approach has multiple advantages, including: 1) incorporating higher order constraints early-on to propose highly plausible lattice points; 2) growing a lattice in multiple directions simultaneously instead of one at a time sequentially; and 3) achieving more efficient and more accurate performance than state-of-the-art algorithms. These advantages are demonstrated by quantitative experimental results on a diverse set of real world photos.

1

****

Local Statistic Based Region Segmentation with
Automatic Scale Selection

Jérome Piovano and Théodore Papadopoulo
Odyssée Project Team, INRIA Sophia Antipolis - Méditerranée, France
{Jerome.Piovano,Theodore.Papadopoulo }@sophia.inria.fr

Abstract. Recently, new segmentation models based on local information have emerged. They combine local statistics of the regions along the contour (inside and outside) to drive the segmentation procedure. Since they are based on local decisions, these models are more robust to local variations of the regions of interest(contrast, noise, blur,...). They nonetheless also introduce some new difficulties which are inherent to the fact of basing a global property (the segmentation) on pure local decisions. This papers explores some of those difficulties and proposes some possible corrections. Results on both 2D and 3D data are compared to those obtained without these corrections.

1

****

A Comparative Analysis of RANSAC
Techniques Leading to Adaptive Real-Time
Random Sample Consensus

Rahul Raguram1, Jan-Michael Frahm1, and Marc Pollefeys1,2
1Department of Computer Science, The University of North Carolina at Chapel Hill
{rraguram,jmf,marc }@cs.unc.edu
2Department of Computer Science, ETH Zürich
marc.pollefeys@inf.ethz.ch

Abstract. The Random Sample Consensus (RANSAC) algorithm is a popular tool for robust estimation problems in computer vision, primarily due to its ability to tolerate a tremendous fraction of outliers. There have been a number of recent efforts that aim to increase the efficiency of the standard RANSAC algorithm. Relatively fewer efforts, however, have been directed towards formulating RANSAC in a manner that is suitable for real-time implementation. The contributions of this work are two-fold: First, we provide a comparative analysis of the state-of-the-art RANSAC algorithms and categorize the various approaches. Second, we develop a powerful new framework for real-time robust estimation. The technique we develop is capable of efficiently adapting to the constraints presented by a fixed time budget, while at the same time providing accurate estimation over a wide range of inlier ratios. The method shows significant improvements in accuracy and speed over existing techniques.

1

****

Video Registration Using Dynamic Textures

Avinash Ravichandran and Ren´ e Vidal
Center for Imaging Science, Johns Hopkins University, Baltimore, MD 21218, USA

Abstract. We propose a dynamic texture feature-based algorithm for registering two video sequences of a rigid or nonrigid scene taken from two synchronous or asynchronous cameras. We model each video sequence as the output of a linear dynamical system, and transform the task of registering frames of the two sequences to that of registering the parameters of the corresponding models. This allows us to perform registration using the more classical image-based features as opposed to space-time features, such as space-time volumes or feature trajectories. As the model parameters are not uniquely defined, we propose a generic method to resolve these ambiguities by jointly identifying the parameters from multiple video sequences. We finally test our algorithm on a wide variety of challenging video sequences and show that it matches the performance of significantly more computationally expensive existing methods.

1

****

Hierarchical Support Vector Random Fields: Joint Training to Combine Local and Global Features

Paul Schnitzspan, Mario Fritz, and Bernt Schiele
Computer Science Department, TU Darmstadt, Germany
{schnitzspan,fritz,schiele }@cs.tu-darmstadt.de

Abstract. Recently, impressive results have been reported for the detection of objects in challenging real-world scenes. Interestingly however, the underlying models vary greatly even between the most successful approaches. Methods using a global feature descriptor (e.g. [1]) paired with discriminative classifiers such as SVMs enable high levels of performance, but require large amounts of training data and typically degrade in the presence of partial occlusions. Local feature-based approaches (e.g. [2–4]) are more robust in the presence of partial occlusions but often produce a significant number of false positives. This paper proposes a novel approach called hierarchical support vector random field that allows 1) to combine the power of global feature-based approaches with the flexibility of local feature-based methods in one consistent multi-layer framework and 2) to automatically learn the tradeoff and the optimal interplay between local, semi-local and global feature contributions. Experiments show that both the combination of local and global features as well as the joint training result in improved detection performance on challenging datasets.

1

****

Scene Segmentation Using the Wisdom of Crowds

Ian Simon and Steven M. Seitz
University of Washington
{iansimon,seitz }@cs.washington.edu

Abstract. Given a collection of images of a static scene taken by many different people, we identify and segment interesting objects. To solve this problem, we use the distribution of images in the collection along with a new field-of-view cue, which leverages the observation that people tend to take photos that frame an object of interest within the field of view. Hence, image features that appear together in many images are likely to be part of the same object. We evaluate the effectiveness of this cue by comparing the segmentations computed by our method against hand-labeled ones for several different models. We also show how the results of our segmentations can be used to highlight important objects in the scene and label them using noisy user-specified textual tag data. These methods are demonstrated on photos of several popular tourist sites downloaded from the Internet.

1

****

Optimization of Symmetric Transfer Error for
Sub-frame Video Synchronization

Meghna Singh1, Irene Cheng2, Mrinal Mandal1, and Anup Basu2
1Department of Electrical and Computer Engineering
2Department of Computing Science
University of Alberta, Edmonton, Alberta, Canada

Abstract. In this work we present a method to synchronize video sequences of events that are acquired via uncalibrated cameras at unknown and dynamically varying temporal offsets. Unlike existing methods that synchronize videos of similar events (i.e., videos related to each other through the motion in the scene) up to an integer alignment, we establish sub-frame video synchronization. While contemporary synchronization algorithms implement a unidirectional alignment which biases the results towards a single reference sequence, we adopt a bi-directional or symmetrical alignment approach that results in a more optimal synchronization. To this end, we propose a novel symmetric transfer error which is dynamically minimized, and reduces the propagation of error from feature extraction and spatial mapping into temporal synchronization. The advantages of our approach are validated by tests conducted on (publicly available) real and synthetic sequences. We present qualitative and quantitative comparisons with another state-of-the-art algorithm. A unique application of this work in generating high-resolution 4D MRI data from multiple low-resolution MRI scans is described.

1

****

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part II, LNCS 5303, pp. 568 – 581, 2008.

© Springer-Verlag Berlin Heidelberg 2008 Shape-Based Retrieval of Heart Sounds for Disease
Similarity Detection

Tanveer Syeda-Mahmood and Fei Wang
IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120
{stf,wangfe}@almaden.ibm.com

Abstract. Retrieval of similar heart sounds from a sound database has applications in physician training, diagnostic screening, and decision support. In this paper, we exploit a visual rendering of heart sounds and model the morphological variations of audio envelopes through a constrained non-rigid translation transform. Similar heart sounds are then retrieved by recovering the corresponding alignment transform using a variant of shape-based dynamic time warping. Results of similar heart sound retrieval are demonstrated for various diseases on a large database of heart sounds.

Keywords: Sound pattern analysis, audio retrieval, curve analysis, healthcare application.

1

****

Learning CRFs Using Graph Cuts

Martin Szummer1, Pushmeet Kohli1, and Derek Hoiem2
1Microsoft Research, Cambridge CB3 0FB, United Kingdom
2Beckman Institute, University of Illinois at Urbana-Champaign, USA

Abstract. Many computer vision problems are naturally formulated as random fields, specifically MRFs or CRFs. The introduction of graph cuts has enabled efficient and optimal inference in associative random fields, greatly advancing applications such as segmentation, stereo reconstruction and many others. However, while fast inference is now widespread, parameter learning in random fields has remained an intractable problem. This paper shows how to apply fast inference algorithms, in particular graph cuts, to learn parameters of random fields with similar efficiency. We find optimal parameter values under standard regularized objective functions that ensure good generalization. Our algorithm enables learning of many parameters in reasonable time, and we explore further speedup techniques. We also discuss extensions to non-associative and multi-class problems. We evaluate the method on image segmentation and geometry recognition.

1

****

Feature Correspondence Via Graph Matching: Models and Global Optimization

Lorenzo Torresani1, Vladimir Kolmogorov2, and Carsten Rother1
1Microsoft Research Lt d., Cambridge, UK
{ltorre,carrot }@microsoft.com
2University College London, UK
vnk@adastral.ucl.ac.uk

Abstract. In this paper we present a new approach for establishing correspondences between sparse image features related by an unknown non-rigid mapping and corrupted by clutter and occlusion, such as points extracted from a pair of images containing a human figure in distinct poses. We formulate this matching task as an energy minimization problem by defining a complex objective function of the appearance and the spatial arrangement of the features. Optimization of this energy is an instance of graph matching, which is in general a NP-hard problem. We describe a novel graph matching optimization technique, which we refer to as dual decomposition (DD), and demonstrate on a variety of examples that this method outperforms existing graph matching algorithms. In the majority of our examples DD is able to find the global minimum within a minute. The ability to globally optimize the objective allows us to accurately learn the parameters of our matching model from training examples. We show on several matching tasks that our learned model yields results superior to those of state-of-the-art methods.

1

****

Event Modeling and Recognition Using Markov Logic Networks■

Son D. Tran and Larry S. Davis
Department of Computer Science
University of Maryland, College Park, MD 20742 USA
{sontran,lsd }@cs.umd.edu

Abstract. We address the problem of visual event recognition in surveillance where noise and missing observations are serious problems. Common sense domain knowledge is exploited to overcome them. The knowledge is represented as first-order logic production rules with associated weights to indicate their confidence. These rules are used in combination with a relaxed deduction algorithm to construct a network of grounded atoms, the Markov Logic Network. The network is used to perform probabilistic inference for input queries about events of interest. The system's performance is demonstrated on a number of videos from a parking lot domain that contains complex interactions of people and vehicles.

1

****

Illumination and Person-Insensitive Head Pose Estimation Using Distance Metric Learning

Xianwang Wang, Xinyu Huang, Jizhou Gao, and Ruigang Yang
Center for Visualization & Virtual Environments,
University of Kentucky,

Lexington KY 40507, USA
{xwang,xhuan4,jgao5,ryang }@cs.uky.edu

Abstract. Head pose estimation is an important task for many face analysis applications, such as face recognition systems and human computer interactions. In this paper we aim to address the pose estimation problem under some challenging conditions, e.g., from a single image, large pose variation, and un-even illumination conditions. The approach we developed combines non-linear dimension reduction techniques with a learned distance metric transformation. The learned distance metric provides better intra-class clustering, therefore preserving a smooth low-dimensional manifold in the presence of large variation in the input images due to illumination changes. Experiments show that our method improves the performance, achieving accuracy within 2-3 degrees for face images with varying poses and within 3-4 degrees error for face images with varying pose and illumination changes.

1

****

2D Image Analysis by Generalized Hilbert Transforms in Conformal Space■

Lennart Wietzke, Oliver Flei schmann, and Gerald Sommer
Kiel University, Department of Computer Science
Christian-Albrechts-Platz 4, 24118 Kiel, Germany
lw@ks.informatik.uni-kiel.de

Abstract. This work presents a novel rotational invariant quadrature filter approach - called the conformal monogenic signal - for analyzing i(ntrinsic)1D and i2D local features of any curved 2D signal such as lines, edges, corners and junctions without the use of steering. The conformal monogenic signal contains the monogenic signal as a special case for i1D signals and combines monogenic scale space, phase, direction/orientation, energy and curvature in one unified algebraic framework. The conformal monogenic signal will be theoretically illustrated and motivated in detail by the relation of the 3D Radon transform and the generalized Hilbert transform on the sphere. The main idea is to lift up 2D signals to the higher dimensional conformal space where the signal features can be analyzed with more degrees of freedom. Results of this work are the low computational time complexity, the easy implementation into existing Computer Vision applications and the numerical robustness of determining curvature without the need of any derivatives.

1

****

An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector

Geert Willems1, Tinne Tuytelaars1, and Luc Van Gool1,2
1ESAT-PSI, K.U. Leuven, Belgium
{gwillems,tuytelaa,vangool }@esat.kuleuven.be
2ETH, Z¨ urich, Switzerland

Abstract. Over the years, several spatio-temporal interest point detectors have been proposed. While some detectors can only extract a sparse set of scale-invariant features, others allow for the detection of a larger amount of features at user-defined scales. This paper presents for the first time spatio-temporal interest points that are at the same time scale-invariant (both spatially and temporally) and densely cover the video content. Moreover, as opposed to earlier work, the features can be computed efficiently. Applying scale-space theory, we show that this can be achieved by using the determinant of the Hessian as the saliency measure. Computations are speeded-up further through the use of approximative box-filter operations on an integral video structure. A quantitative evaluation and experimental results on action recognition show t he strengths of the p roposed detect

or
in terms of repeatability, accuracy and speed, in comparison with previously proposed detectors.

1

****

A Graph Based Subspace Semi-supervised Learning Framework for Dimensionality Reduction

Wuyi Yang, Shuwu Zhang, and Wei Liang
Digital Content Technology Research Center, Institute of Automation,
Chinese Academy of Sciences, Beijing, China, 100190

Abstract. The key to the graph based semi-supervised learning algorithms for classification problems is how to construct the weight matrix of the p-nearest neighbor graph. A new method to construct the weight matrix is proposed and a graph based Subspace Semi-supervised Learning Framework (SSLF) is developed. The Framework aims to find an embedding transformation which respects the discriminant structure inferred from the labeled data, as well as the intrinsic geometrical structure inferred from both the labeled and unlabeled data. By utilizing this framework as a tool, we drive three semi-supervised dimensionality reduction algorithms: Subspace Semi-supervised Linear Discriminant Analysis (SSLDA), Subspace Semi-supervised Locality Preserving Projection (SSLPP), and Subspace Semi-supervised Marginal Fisher Analysis (SSMFA). The experimental results on face recognition demonstrate our subspace semi-supervised algorithms are able to use unlabeled samples effectively.

1

****

Online Tracking and Reacquisition Using Co-trained Generative and Discriminative Trackers

Qian Yu, Thang Ba Dinh, and G´ erard Medioni
University of Southern California, Los Angeles, CA 90089-0273, USA
{qianyu,thangdin,medioni }@usc.edu

Abstract. Visual tracking is a challenging problem, as an object may change its appearance due to viewpoint variations, illumination changes, and occlusion. Also, an object may leave the field of view and then reappear. In order to track and reacquire an unknown object with limited labeling data, we propose to learn these changes online and build a model that describes all seen appearance while tracking. To address this semi-supervised learning problem, we propose a co-training based approach to continuously label incoming data and online update a hybrid discriminative generative model. The generative model uses a number of low dimension linear subspaces to describe the appearance of the object. In order to reacquire an object, the generative model encodes all the appearance variations that have been seen. A discriminative classifier is implemented as an online support vector machine, which is trained to focus on recent appearance variations. The online co-training of this hybrid approach accounts for appearance changes and allows reacquisition of an object after total occlusion. We demonstrate that under challenging situations , this method has strong r eacquisition ability and robustness to distracters in background.

1

****

Statistical Analysis of Global Motion Chains

Jenny Y uen1,■andYasuyukiMatsush ita2
1CSAIL, Massachusetts Institute of Technology, Cambridge MA 02139
2Visual Computing Group, Microsoft Research Asia, Beijing 100080, China
jenny@csail.mit.edu, yasumat@microsoft.com

Abstract. Multiple elements such as lighting, colors, dialogue, and camera motion contribute to the style of a movie. Among them, camera motion is commonly overlooked yet a crucial point. For instance, documentaries tend to use long smooth pans whereas action movies usually have short and dynamic movements. This information, also referred to as global motion, could be leveraged by various applications in video clustering, stabilization, and editing. We perform analyses to study the in-class characteristics of these motions as well as their relationship with motions of other movie types. In particular, we model global motion as a multi-scale distribution of transformation matrices from frame to frame. Secondly, we quantify the difference between pairs of videos using the KL-divergence of these distributions. Finally, we demonstrate an application modeling and clustering commercial and amateur videos. Experiments performed show advantage compared to the usage of some local motion-based approaches.

1

****

Active Image Labeling and Its Application to Facial Action Labeling

Lei Zhang[1], Yan Tong[2], and Qiang Ji[1]

[1]Electrical, Computer, and Systems Engineering Department, Rensselaer Polytechnic Institute

[2]Visualization and Computer Vision Lab, GE Global Research Center

zhangl2@rpi.edu,tongyan@research.ge.com,qji@ecse.rpi.edu

Abstract. For many tasks in computer vision, it is very important to produce the groundtruth data. At present, this is mostly done manually. Manual data labeling is labor-intensive and prone to the human errors. The training data it produces often lacks in both quantity and quality. Fully automatic data labeling, on the other hand, is not feasible and reliable. In this paper, we propose an interactive image labeling technique for efficient and accurate data labeling. The proposed technique includes two parts: an automatic labeling part and a human intervention part. Constructed on a Bayesian Network, the automatic image labeler produces an initial labeling of the image. A person then examines the initial labeling and makes some minor corrections. The selected human corrections and the image measurements are then integrated by the Bayesian Network framework to produce a refined labeling. To minimize the human involvement, an active user feedback strategy is developed, through which the optimal user feedback is determined, so that the labeling errors in the subsequent re-labeling process can be maximally reduced. The proposed framework combines the advantages of the human input with those of the machine so that the reliable, accurate, and efficient data labeling can be achieved. We demonstrate the validity of the proposed framework for interactive labeling of facial action units. The proposed methodology, however, is not limited to labeling of facial action units. It can be easily extended to other areas such as interactive image segmentation.

1

****

Real Time Feature Based 3-D Deformable Face Tracking

Wei Zhang[1],■, Qiang Wang[2], and Xiaoou Tang[1]

[1]Dept. of Information Engineering, The Chinese University of Hong Kong,
Hong Kong, China

{dylan,xtang }@ie.cuhk.edu.hk

[2]Microsoft Research Asia, Beijing, China

qiangwa@microsoft.com

Abstract. In this paper, we develop a novel framework for 3D tracking

of the non-rigid face deformation from a single camera. The difficulty of the problem lies in the fact that 3D deformation parameter estimation becomes unstable when there are few reliable facial features correspondences. Unfortunately, this often occurs in real tracking scenario when there is significant illumination change, motion blur or large pose variation. In order to extract more information of feature correspondences, the proposed framework integrates three types of features which discriminate face deformation across different views: 1) the semantic features which provide constant correspondences between 3D model points and major facial features; 2) the silhouette features which provide dynamic correspondences between 3D model points and facial silhouette under varying views; 3) the online tracking features that provide redundant correspondences between 3D model points and salient image features.
The integration of these complementary features is important for robust estimation of the 3D parameters. In order to estimate the high dimensional 3D deformation parameters, we develop a hierarchical parameter estimation algorithm to robustly estimate both rigid and non-rigid 3D parameters. We show the importance of both features fusion and hierarchical parameter estimation for reliable tracking 3D face deformation. Experiments demonstrate the robustness and accuracy of the proposed algorithm especially in the cases of agile head motion, drastic illumination change, and large pose change up to profile view.

1

****

Rank Classification of Linear Line Structure in Determining Trifocal Tensor

Ming Zhao and Ronald Chung■
Department of Mechanical & Automation Engineering
The Chinese University of Hong Kong
{mzhao,rchung }@mae.cuhk.edu.hk

Abstract. The problem we address is: given line correspondences over three views, what is the condition of the line correspondences for the spatial relation of the three associated camera positions to be uniquely recoverable? We tackle the problem from the perspective of trifocal tensor, a quantity that captures the relative positions of the cameras in relation to the three views. We show that the rank of the matrix that leads to the estimation of the tensor reduces to 7, 11, 15 respectively for line pencil, point star, and ruled plane, which are structures that belong to linear line space; and 12, 19, 23 for general ruled surface, general linear congruence, and general linear line complex. These critical structures are quite typical in reality, and thus the findings are important to the validity and stability of practically all algorithms related to structure from motion and projective reconstruction using line correspondences.

1

****

Learning Visual Shape Lexicon for Document Image Content Recognition

Guangyu Zhu, Xiaodong Yu, Yi Li, and David Doermann
University of Maryland, College Park, MD 20742, USA

Abstract. Developing effective content recognition methods for diverse imagery continues to challenge computer vision researchers. We present a new approach for document image content categorization using a lexicon of shape features. Each lexical word corresponds to a scale and rotation invariant shape feature that is generic enough to be detected repeatably and segmentation free. We learn a concise, structurally indexed shape lexicon from training by clustering and partitioning feature types through graph cuts. We demonstrate our approach on two challenging document image content recognition problems: 1) The classification of 4,500 Web images crawled from Google Image Search into three content categories — pure image, image with text, and document image, and 2)

Language identification of 8 languages (Arabic, Chinese, English, Hindi, Japanese, Korean, Russian, and Thai) on a 1 ,512 complex document image database composed of mixed machine printed text and handwriting. Our approach is capable to handle high intra-class variability and shows results that exceed other state-of-the-art approaches, allowing it to be used as a content recognizer in image indexing and retrieval systems.

1

****

Unsupervised Structure Learning: Hierarchical
Recursive Composition, Suspicious Coincidence
and Competitive Exclusion

Long (Leo) Zhu[1], Chenxi Lin[2], Haoda Huang[2], Yuanhao Chen[3],
and Alan Yuille[1,4]
[1] Department of Statistics, University of California, Los Angeles
{lzhu,yuille }@stat.ucla.edu
[2] Microsoft Research Asia
{chenxi.lin,hahuang }@microsoft.com
[3] University of Science and Technology of China
yhchen4@ustc.edu
[4] Department of Psychology and Computer Science, UCLA

Abstract. We describe a new method for unsupervised structure learning of a hierarchical compositional model (HCM) for deformable objects. The learning is unsupervised in the sense that we are given a training dataset of images containing the object in cluttered backgrounds but we do not know the position or boundary of the object. The structure learning is performed by a bottom-up and top-down process. The bottom-up process is a novel form of hierarchical clustering which recursively composes proposals for simple structures to generate proposals for more complex structures. We combine standard clustering with the suspicious coincidence principle and the competitive exclusion principle to prune the number of proposals to a practical number and avoid an exponential explosion of possible structures. The hierarchical clustering stops automatically, when it fails to generate new proposals, and outputs a proposal for the object model. The top-down process validates the proposals and fills in missing elements. We tested our approach by using it to learn a hierarchical compositional model for parsing and segmenting horses on Weizmann dataset. We show that the resulting model is comparable with (or better than) alternative methods. The versatility of our approach is demonstrated by learning models for other objects (e.g., faces, pianos, butterflies, monitors, etc.). It is worth noting that the low-levels of the object hierarchies automatically learn generic image features while the higher levels learn object specific features.

1

****

Contour Context Selection for Object Detection:
A Set-to-Set Contour Matching Approach

Qihui Zhu[1], Liming Wang[2], Yang Wu[3], and Jianbo Shi[1]
[1] Department of Computer and Information Science, University of Pennsylvania
qihuizhu@seas.upenn.edu, jshi@cis.upenn.edu
[2] Department of Computer Science and Engineering, Fudan University
wanglm@fudan.edu.cn
[3] Institue of Artificial Intelligence and Robotics, Xi ’an Jiaotong University
ywu@aiar.xjtu.edu.cn

Abstract. We introduce a shape detection framework called Contour Context Selection for detecting objects in cluttered images using only one exemplar. Shape based detection is invariant to changes of object appearance, and can reason with geometrical abstraction of the object. Our approach uses salient contours as integral tokens for shape matching. We seek a maximal, holistic matching of shapes, which checks shape features from a large spatial extent, as well as long-range contextual relationships among object parts. This amounts to finding the correct figure/ground contour labeling, and optimal correspondences between control points on/around contours. This removes accidental alignments and does not hallucinate objects in background clutter, without negative training examples. We formulate this task as a set-to-set contour matching problem. Naive methods would require searching over ’exponentially’ many figure/ground contour labelings. We simplify this task by encoding the shape descriptor algebraically in a linear form of contour figure/ground variables. This allows us to use the reliable optimization technique of Linear Programming. We demonstrate our approach on the challenging task of detecting bottles, swans and other objects in cluttered images.

1

****

Robust Object Tracking by Hierarchical Association of
Detection Responses

Chang Huang, Bo Wu, and Ramakant Nevatia
University of Southern California, Los Angeles, CA 90089-0273, USA
{huangcha,bowu,nevatia }@usc.edu

Abstract. We present a detection-based three-level hierarchical association approach to robustly track multiple objects in crowded environments from a single camera. At the low level, reliable tracklets (i.e. short tracks for further analysis) are generated by linking detection responses based on conservative affinity constraints. At the middle level, these tracklets are further associated to form longer tracklets based on more complex affinity measures. The association is formulated as a MAP problem and solved by the Hungarian algorithm. At the high level, entries, exits and scene occluders are estimated using the already computed tracklets, which are used to refine the final trajectories. This approach is applied to the pedestrian class and evaluated on two challenging datasets. The experimental results show a great improvement in performance compared to previous methods.

1

****

Improving the Agility of Keyframe-Based SLAM

Georg Klein and David Murray
Active Vision Laboratory, University of Oxford, UK
{gk,dwm }@robots.ox.ac.uk

Abstract. The ability to localise a camera mov ing in a previously unknown environment is desirable for a wide range of applications. In computer vision this problem is studied as monocular SLAM. Recent years have seen improvements to the usability and scalab ility of m onocular SLAM systems to the point that they may soon find uses outside of laboratory conditions. However, the robustness of these systems to rapid camera motions (we refer to this quality as agility) still lags behind that of tracking systems which use known object models. In this paper we attempt to remedy this. We present two approaches to improving the agility of a keyframe-based SLAM system: Firstly, we add edge features to the map and exploit their resilience to motion blur t o improve tracking under fast motion. Secondly, we implement a very simple inter-fra me rotation estimat or to aid tracking when the camera is rapidly panning – and demonstrate that this method also enables a trivially simple yet effective relocalisation method. Results show that a SLAM system combining points, edge features and motion initialisation allows highly agile tracking at a moderate increase in processing time.

1

****

Articulated Multi-body Tracking under Egomotion

Stephan Gammeter1, Andreas Ess1, Tobias J¨ aggli1, Konrad Schindler1,
Bastian Leibe1,2, and Luc Van Gool1,3
1ETH Z¨ urich
2RWTH Aachen
3KU Leuven, IBBT
{stephaga,aess,jaeggli,schindler,leibe }@vision.ee.ethz.ch

Abstract. In this paper, we address the problem of 3D articulated multi-person tracking in busy street scenes from a moving, human-level observer. In order to handle the complexity of multi-person interactions, we propose to pursue a two-stage strategy. A multi-body detection-based tracker first analyzes the scene and recovers individual pedestrian trajectories, bridging sensor gaps and resolving temporary occlusions. A specialized articulated tracker is then applied to each recovered pedestrian trajectory in parallel to estimate the tracked person's precise body pose over time. This articulated tracker is implemented in a Gaussian Process framework and operates on global pedestrian silhouettes using a learned statistical representation of human body dynamics. We interface the two tracking levels through a guided segmentation stage, which combines traditional bottom-up cues with top-down information from a human detector and the articulated tracker's shape prediction. We show the proposed approach's viability and demonstrate its performance for articulated multi-person tracking on several challenging video sequences of a busy inner-city scenario.

1

****

Robust Real-Time Visual Tracking Using
Pixel-Wise Posteriors■

Charles Bibby and Ian Reid
Active Vision Lab
Department of Engineering Science
University of Oxford
cbibby@robots.ox.ac.uk,ian@robots.ox.ac.uk

Abstract. We derive a probabilistic framework for robust, real-time, visual tracking of previously unseen objects from a moving camera. The tracking problem is handled using a bag-of-pixels representation and comprises a rigid registration between frames, a segmentation and online appearance learning. The registration compensates for rigid motion, segmentation models any residual shape deformation and the online appearance learning provides continual refinement of both the object and background appearance models. The key to the success of our method is the use of pixel-wise posteriors, as opposed to likelihoods. We demonstrate the superior performance of our tracker by comparing cost function statistics against those commonly used in the visual tracking literature. Our comparison method provides a way of summarising tracking performance using lots of data from a variety of different sequences.

1

****