****

Schwarps: Locally Projective Image Warps
Based on 2D Schwarzian Derivatives

Rahat Khan, Daniel Pizarro, and Adrien Bartoli
ISIT, UMR 6284 CNRS-UdA, Clermont-Ferrand, France

Abstract. Image warps -or just warps- capture the geometric deformation existing between two images of a deforming surface. The current approach to enforce a warp’s smoothness is to penalize its second order partial derivatives. Because this favors locally a■ne warps, this fails to capture the local projective component of the image deformation. This may have a negative impact on applications such as image registration and deformable 3D reconstruction. We propose a novel penalty designed to smooth the warp while capturing the deformation’s local projective structure. Our penalty is based on equivalents to the Schwarzian derivatives, which are projective di■erential invariants exactly preserved by homographies. We propose a methodology to derive a set of Partial Differential Equations with homographies as solutions. We call this system the Schwarzian equations and we explicitly derive them for 2D functions using di■erential properties of homographies. We name as Schwarp a warp which is estimated by penalizing the residual of Schwarzian equations. Experimental evaluation shows that Schwarps outperform existing warps in modeling and extrapolation power, and lead to far better results in Shape-from-Template and camera calibration from a deformable surface.

Keywords: Schwarzian Penalizer, Bending Energy, Projective Di■eren-
tial Invariants, Image Warps.
1
****
gDLS: A Scalable Solution
to the Generalized Pose and Scale Problem
Chris Sweeney, Victor Fragoso, Tobias H¨ ollerer, and Matthew Turk
University of California, Santa Barbara, USA
{cmsweeney,vfragoso,holl,mturk }@cs.ucsb.edu
Abstract. In this work, we present a scalable least-squares solution for
computing a seven degree-of-freedom similarity transform. Our method
utilizes thegeneralized camera model to computerelative rotation, trans-lation,
and scale from four or more 2D-3D correspondences. In particu-
lar, structure and motion estimations from monocular cameras lack scale
without speci■c calibration. As such, our methods have applications inloop closu
re in visual odometry and registering multiple structure from
motion reconstructions where scale must be recovered. We formulate the
generalized pose and scale problem as a minimization of a least squarescost func
tion and solve this minimization without iterations or initializa-tion. Addition
ally, we obtain all minima of the cost function. The order
of the polynomial system that we solve is independent of the number of
points, allowing our overall approach to scale favorably. We evaluate ourmethod
experimentally on synthetic and real datasets and demonstrate
that our methods produce higher accuracy similarity transform solutions
than existing methods.
1
****
Generalized Connectivity Constraints
for Spatio-temporal 3D Reconstruction
Martin Ralf Oswald, Jan St¨ uhmer, and Daniel Cremers
Department of Computer Science, Technische Universit¨ at M¨unchen■
Boltzmannstr. 3, 85748 Garching, Germany
Abstract. This paper introduces connectivity preserving constraints
intospatio-temporal multi-view reconstruction. Wee■cientlymodel con-
nectivity constraints by precomputing a geodesic shortest path tree on
the occupancy likelihood. Connectivity of the ■nal occupancy labeling is
ensured with a set of linear constraints on the labeling function. In order
to generalize the connectivity constraints from objects with genus 0 to
an arbitrary genus, we detect loops by analyzing the visual hull of the
scene. A modi■cation of the constraints ensures connectivity in the pres-
ence of loops. The proposed e■cient implementation adds little runtime
andmemoryoverheadto thereconstruction method.Several experiments
show signi■cant improvement over state-of-the-art methods and validate
the practical use of this approach in scenes with ■ne structured details.
Keywords: connectivity constraints, spatio-temporal 3D reconstruc-
tion.
1 of 16 input images No Connectivity With a Connectivity Generalized Connec-
Constraint [22] Constraint [25]+[22] tivity Constraint
Fig. 1.Embeddingconnectivityconstraintsintomulti-viewreconstructionclearly helps
to recover ■ne structures like the rope. The tree-shaped connectivity prior [25]
only
works for objects without holes (genus 0), resulting in disconnected parts when
the
rope touches the head. The proposed generalized connectivity constraint works fo
r
objects with arbitrary genus. Dataset: ’jumping rope’ sequence from the INRIA 4D
repository [16].
■This work was supported by the ERC Starting Grant ’Convex Vision’ and the Tech-
nische Universit¨ at M¨unchen - Institute for Advanced Study, funded by the Germ
an

Excellence Initiative.
D. Fleet et al. (Eds.): ECCV 2014, Part IV, LNCS 8692, pp. 32–46, 2014.
c/circlecopyrtSpringer International Publishing Switzerland 201
****

Passive Tomography of Turbulence Strength
Marina Alterman1,Y o a vY .S c h e c h n e r1,
Minh Vo2, and Srinivasa G. Narasimhan2
1Dept. Electrical Eng., Technion - Israel Institute of Technology, Haifa, Israel
2Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
Abstract. Turbulence is studied extensively in remote sensing, astron-
omy, meteorology, aerodynamics and ■uid dynamics. The strength of
turbulence is a statistical measure of local variations in the turbulent
medium. It in■uences engineering decisions made in these domains. Tur-bulence st
rength (TS) also a■ects safety of aircraft and tethered bal-
loons, and reliability of free-space electromagnetic relays. We show that
it is possible to estimate TS, without having to reconstruct instantaneous■uid ■
ow ■elds. Instead, the TS ■eld can be directly recovered, passively,
using videos captured from di■erent viewpoints. We formulate this as a
linear tomography problem with a structure unique to turbulence ■elds.No tight s
ynchronization between cameras is needed. Thus, realization is
very simple to deploy using consumer-grade cameras. We experimentally
demonstrate this both in a lab and in a large-scale uncontrolled complexoutdoor
environment, which includes industrial, rural and urban areas.

1 The Need to Recover Turbulence Strength

Turbulence creates refra ctive perturbations to lig ht passing through a scene.
This causes random distortions when imaging background objects. Hence, mod-
eling and trying to compensate for rando m refractive distortions has long been
studied in remote sensing [40], astronomy [34] and increasingly in computer vi-s
ion [2,4,10,14,18,35,38,41,52,55]. Nevertheless,these distortionarenot necessar-
ily a problem: they o■er information about the medium and the scene itself [44].
This insight is analogous to imaging in scattering media (fog [29], haze [19,37]
,water [11,30]), where visibility reduction yields ranging and characterizingof
the
medium. Similar e■orts are made to reconstruct refracting (transparent) solids
or water surfaces [3,16,28,43,46] from images of a distorted background or light
■eld [50,51]. In turbulence, refraction occurs continuously throughout a volume.
Weexploitrandom image distortions as a means to estimate the spatial (vol-
umetric) distribution of turbulence strength (TS). The strength of turbulence
is a statistical measure of local variations in the medium [20,21]. Often, it is
not necessary to estimate an instantaneous snapshot of air density or refraction
■eld [32,42]. Rather local statistics is relied upon heavily in many application
s.

Meteorologists rely on TS to understand convection (which forms clouds), wind,
and atmospheric stability. This is measured using special Doppler lidars [9,31],
which are very expensive. Turbulence si gni■cantly a■ects the e■ciency of wind
D. Fleet et al. (Eds.): ECCV 2014, Part IV, LNCS 8692, pp. 47–60, 2014.
c/circlecopyrtSpringer International Publishing Switzerland 201
****

A Non-local Method
for Robust Noisy Image Completion
Wei Li, Lei Zhao, Duanqing Xu, and Dongming Lu
Zhejiang University, Hangzhou, China
Abstract. The problem of noisy image completion refers to recovering
an image from a random subset of its noisy intensities. In this paper,
we propose a non-local patch-based algorithm to settle the noisy image
completion problem following the methodology “grouping and collabo-ratively ■lte
ring”. The target of “grouping” is to form patch matrices by
matching and stacking similar image patches. And the “collaboratively
■ltering” is achieved by transforming the tasks of simultaneously esti-mating mi
ssing values and removing noises for the stacked patch matri-

ces into low-rank matrix completion problems, which can be e■ciently solved by minimizing the nuclear norm of the matrix with linear con-straints. The ■nal output is produced by synthesizing all the restored patches. To improve the robustness of our algorithm, we employ an ef-■cient and accurate patch matching method with adaptations includingpre-completion and outliers removal, etc. Experiments demonstrate that our approach achieves state-of-the-art performance for the noisy image completion problem in terms of both PSNR andsubjective visual quality.

1

****

Improved Motion Invariant Deblurring through Motion Estimation

Scott McCloskey
Honeywell Labs, USA

Abstract. We address the capture of sharp images of fast-moving objects, and build on the Motion Invariant photographic technique. The key advantage of motion invariance is that, unlike other computationalphotographic techniques, it does not require pre-exposure velocity estimation in order to ensure numerically stable deblurring. Its disadvantage is that the invariance is only approximate - objects moving with non-zero velocity will exhibit artifacts in the deblurred image related to tail clipping in the motion Point Spread Function (PSF). We model these artifacts as a convolution of the desired latent image with an error PSF,and demonstrate that the spatial scale of these artifacts corresponds to the object velocity. Surprisingly, despite the use of parabolic motion to capture an image in which blur is invariant to motion, we demonstratethat the motion invariant image can be used to estimate object motion post-capture . With real camera images, we demonstratesigni■cant reductions in the artifacts by using the estimated motion for deblurring. Wealso quantify a 96% reduction in recons truction error, relative to a ■oor established by exact PSF deconvolution, via simulation with a large test set of photographic images.

1

****

Consistent Matting for Light Field Images

Donghyeon Cho, Sunyeong Kim, and Yu-Wing Tai
Korea Advanced Institute of Science and Technology (KAIST)

Abstract. We present a new image matting algorithm to extract consistent alpha mattes across sub-images of a light ■eld image. Instead ofmatting each sub-image individually, our approach utilizes the epipolar plane image (EPI) to construct comprehensive foreground and background sample sets across the sub-images without missing a true sample.The sample sets represent all color variation of foreground and background in a light ■eld image, and the optimal alpha matte is obtained by choosing the best combination of foreground and background samplesthat minimizes the linear composite error subject to the EPI correspondence constraint. To further preserve consistency of the estimated alpha mattes across di■erent sub-images, we impose a smoothness constraintalong the EPI of alpha mattes. In experimental evaluations, we have created a dataset where the ground truth alpha mattes of light ■eld images were obtained by using the blue screen technique. A variety of exper-iments show that our proposed algorithm produces both visually and quantitatively high-quality matting results for light ■eld images.

Keywords: Image Matting, Light ■eld image, EPI.

1

****

Consensus of Regression for Occlusion-Robust Facial Feature Localization

Xiang Yu1,Z h eL i n2, Jonathan Brandt2, and Dimitris N. Metaxas1
1Rutgers University, Piscataway, NJ 08854, USA

2Adobe Research, San Jose, CA 95110, USA
Abstract. We address the problem of robust facial feature localization in the presence of occlusions, which remains a lingering problem in facial analysis despite intensive long-term studies. Recently, regression-basedapproaches to localization have produced accurate results in many cases, yet are still subject to signi■cant error when portions of the face are occluded. To overcome this weakness, we propose an occlusion-robustregression method by forming a consensus from estimates arising from a set of occlusion-speci■c regressors. That is, each regressor is trained to estimate facial feature locations under the precondition that a particularpre-de■ned region of the face is occluded. The predictions from each regressor are robustly merged using a Bayesian model that models each regressor's prediction correctness likelihood based on local appearance andconsistency with other regressors with overlapping occlusion regions. After localization, the occlusion state for each landmark point is estimated using a Gaussian MRF semi-supervised learning method. Experimentson both non-occluded and occluded face databases demonstrate that our approach achieves consistently better results over state-of-the-art methods for facial landmark localization and occlusion detection.
Keywords: Facial feature localization, Consensus of Regression, Occlusion detection, Face alignment.

****

Learning the Face Prior
for Bayesian Face Recognition
Chaochao Lu and Xiaoou Tang
Department of Information Engineering,
The Chinese University of Hong Kong, China
Abstract. For the traditional Bayesian face recognition methods, a simple prior on face representation cannot cover large variations in facialposes, illuminations, expressions, aging, and occlusions in the wild. In thispaper,wepropose anewapproachtolearn theface priorfor Bayesian face recognition. First, we extend Manifold Relevance Determination tolearn the identity subspace for each individual automatically. Based on the structure of the learned identity subspaces, we then propose to estimate Gaussian mixture densities in the observation space with Gaussianprocess regression. During the training of our approach, the leave-set-out algorithm is also developed for over■tting avoidance. On extensive experimental evaluations, the learned face prior can improve the per-formance of the traditional Bayesian face and other related methods signi■cantly. It is also proved that the simple Bayesian face method with the learned face prior can handle the complex intra-personal variationssuch as large poses and large occlusions. Experiments on the challenging LFW benchmark shows that our algorithm outperforms most of the state-of-art methods.

****

Spatio-temporal Event Classi■cation Using
Time-Series Kernel Based Structured Sparsity■
L´aszl´oA .J e n i1, Andr´as L■orincz2,Z o l t ´an Szab´o3,
Je■rey F. Cohn1,4, and Takeo Kanade1
1Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
2Faculty of Informatics, E¨ otv¨os Lor´and University, Budapest, Hungary
3Gatsby Computational Neuroscience Unit, University College London, London, UK
4Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA
laszlo.jeni@ieee.org, andras.lorincz@elte.hu,
zoltan.szabo@gatsby.ucl.ac.uk, {jeffcohn,tk }@cs.cmu.edu
Abstract. In many behavioral domains, such as facial expression and gesture, sparse structure is prevalent. This sparsity would be well suited foreventdetectionbutforoneproblem.Featurestypicallyareconfounded

byalignmenterrorinspaceandtime.Asaconsequence,high-dimensional representations such as SIFT and Gabor features have been favored despite their much greater computational cost and potential loss of information.WeproposeaKernelStructuredSparsity(KSS)methodthatcan handle both the temporal alignment problem and the structured sparse reconstructionwithinacommonframework,anditcanrelyonsimplefeatures. We characterize spatio-temporal events as time-series of motion patternsandbyutilizingtime-serieskernelsweapplystandardstructured-sparse coding techniques to tackle this important problem. We evaluated the KSS method using both gesture and facial expression datasets that include spontaneous behavior and di■er in degree of di■culty and type of ground truth coding. KSS outperformed both sparse and non-sparse methods that utilize complex image features and their temporal extensions. In the case of early facial event classi■cation KSS had 10% higher accuracy as measured by F1score over kernel SVM methods1.

Keywords: structured sparsity, time-series kernels, facial expression classi■cation, gesture recognition.

1

****

Feature Disentangling Machine - A Novel Approach of Feature Selection and Disentangling in Facial Expression Analysis

Ping Liu1, Joey Tianyi Zhou2, Ivor Wai-Hung Tsang3, Zibo Meng1, Shizhong Han1,a n dY a nT o n g1
1Department of Computer Science, University of South Carolina, USA
2Center for Computational Intelligence, Nanyang Technology University, Singapore
3Center for Quantum Computation and Intelligent Systems, University of Technology, Australia

Abstract. Studies in psychology show that not all facial regions are of importance in recognizing facial expressions and di■erent facial regions make di■erent contributions in various facial expressions. Motivated by this, a novel framework, named Feature Disentangling Machine(FDM), is proposed to e■ectively select active features characterizing facial expressions. More importantly, the FDM aims to disentangle these selected features into non-overlapped groups, in particular, common featuresthat are shared across di■erent expressions and expression-speci■c features that are discriminative only for a target expression. Speci■cally, the FDM integrates sparse support vector machine and multi-task learn-ing in a uni■ed framework, where a novel loss function and a set of constraints are formulated to precisely control the sparsity and naturally disentangle active features. Extensive experiments on two well-knownfacial expression databases have demonstrated that the FDM outperforms the state-of-the-art methods for facial expression analysis. More importantly, the FDM achieves an impressive performance in a cross-database validation, which demonstrates the generalization capability of the selected features.

1

****

Joint Unsupervised Face Alignment and Behaviour Analysis■

Lazaros Zafeiriou, Epameinondas Antonakos, Stefanos Zafeiriou, and Maja Pantic
Computing Department, Imperial College London, UK
{l.zafeiriou12,e.antonakos,s.zafeiriou,m.pantic }@imperial.ac.uk

Abstract. The predominant strategy for facial expressions analysis and temporal analysis of facial events is the following: a generic facial landmarkstracker,usually trained on thousandsofcarefully annotated examples,isappliedtotrackthelandmarkpoints,andthenanalysisisperformed usingmostlytheshapeandmorerarelythefacialtexture.Thispaperchallenges the above framework by showing that it is feasible to perform joint

landmarkslocalization(i.e.spatialalignment)andtemporalanalysisofbe-
haviouralsequencewiththeuseofasimplefacedetectorandasimpleshape
model. To do so, we propose a new component analysis technique, which
wecallAutoregressiveComponentAnalysis(ARCA),andweshowhowthe
parameters of a motion model can be jointly retrieved. The method does
not require the use of any sophisticated landmark tracking methodology
and simply employs pixelintensities for thetexturerepresentation.
Keywords: Facealignment,timeseriesalignment,slowfeatureanalysis.
1
****

Learning a Deep Convolutional Network
for Image Super-Resolution
Chao Dong1, Chen Change Loy1, Kaiming He2, and Xiaoou Tang1
1Department of Information Engineering,
The Chinese University of Hong Kong, China
2Microsoft Research Asia, Beijing, China
Abstract. We propose a deep learning method for single image super-
resolution (SR). Our method directly learns an end-to-end mapping be-tween the l
ow/high-resolution images. The mapping is represented as
a deep convolutional neural network (CNN) [15] that takes the low-
resolution image as the input and outputs the high-resolution one. Wefurther sho
w that traditional sparse-coding-based SR methods can also
be viewed as a deep convolutional network. But unlike traditional meth-
odsthat handleeach component separately,ourmethodjointly optimizesall layers. Ou
r deep CNN has a lightweight structure, yet demonstrates
state-of-the-art restoration quality, and achieves fast speed for practical
on-line usage.
Keywords: Super-resolution, deep convolutional neural networks.
1
****

Discriminative Indexing
for Probabilistic Image Patch Priors
Yan Wang1,■, Sunghyun Cho2,■■, Jue Wang2, and Shih-Fu Chang1
1Dept. of Electrical Engineering, Columbia University, USA
{yanwang,sfchang }@ee.columbia.edu
2Adobe Research, USA
sodomau@postech.ac.kr, juewang@adobe.com
Abstract. Newly emerged probabilistic image patch priors, such as Ex-
pected Patch Log-Likelihood (EPLL), have shown excellent performance
on image restoration tasks, especially deconvolution, due to its rich ex-
pressiveness. However, its applicability is limited by the heavy compu-
tation involved in the associated optimization process. Inspired by the
recent advances on using regression trees to index priors de■ned on a
Conditional Random Field, we propose a novel discriminative indexing
approach on patch-based priors to expedite the optimization process.
Speci■cally,weproposeane■cienttreeindexingstructureforEPLL,and
overcome its training tractability challenges in high-dimensional spaces
by utilizing special structures of the prior. Experimental results show
that our approach accelerates state-of-the-art EPLL-based deconvolu-
tion methods by up to 40 times, with very little quality compromise.
1
****

Modeling Video Dynamics
with Deep Dynencoder
Xing Yan, Hong Chang, Shiguang Shan, and Xilin Chen
Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CA
S),
Institute of Computing Technology, CAS, Beijing, 100190, China
{xing.yan,hong.chang,shiguang.shan,xilin.chen }@vipl.ict.ac.cn
Abstract. Videos always exhibit various pattern motions, which can

be modeled according to dynamics between adjacent frames. Previousmethods based on linear dynamic system can model dynamic textures
but have limited capacity of representing sophisticated nonlinear dynamics. Inspired by the nonlinear expression power of deep autoencoders, wepropose a novel model named dynencoder which has an autoencoder at
the bottom and a variant of it at the top (named as dynpredictor). It generates hidden states from raw pixel inputs via the autoencoder andthen encodes the dynamic of state transition over time via the dynpredictor. Deep dynencoder can be constructed by proper stacking strategy
and trained by layer-wise pre-training and joint ■ne-tuning.Experimentsverify that our model can describe sophisticated video dynamics andsyn-thesize endless video texture sequences with high visual quality. We also
design classi■cation and clustering methods based on our model and
demonstrate the e■cacy of them on tra■c scene classi■cation and mo-tion segmentation. ...
Keywords: Video Dynamics, Deep Model, Autoencoder, Time Series,
Dynamic Textures.
1
****

Good Image Priors for Non- blind Deconvolution:
Generic vs.Speci■c
Libin Sun1, Sunghyun Cho2,■, Jue Wang2,a n dJ a m e sH a y s1
1Brown University, Providence, RI 02912, USA
2Adobe Research, Seattle, WA 98103, USA
{lbsun,hays }@cs.brown.edu, sodomau@postech.ac.kr, juewang@adobe.com
Abstract. Most image restoration techniques build “universal” image
priors, trained on a variety of scenes, which can guide the restoration
of any image. But what if we have more speci■c training examples,
e.g.sharp images of similar scenes? Surprisingly, state-of-the-art image
priors don’t seem to bene■t from from context-speci■c training examples. Re-training generic image priors using ideal sharp example images provides minimal improvement in non-blind deconvolution. To help
understand this phenomenon we explore non-blind deblurring performance over a broad spectrum of training image scenarios. We discover
two strategies that become bene■cial as example images become more
context-appropriate: (1) locallyadapted priors trained from region level
correspondencesigni■cantly outperform globally trained priors,and(2)a
novel multi-scale patch-pyramid formulation is more successful at transferring mid and high frequency details from example scenes. Combining
these two key strategies we can qualitatively and quantitatively outperform leading generic non-blind deconvolution methods when context-appropriate example images are available. We also compare to recent
work which, like ours, tries to make use of context-speci■c examples.
Keywords: deblur, non-blind deconvolution, gaussian mixtures, image
pyramid, image priors, camera shake.
1
****

Image Deconvolution Ringing Artifact Detection
and Removal via PSF Frequency Analysis
Ali Mosleh1, J.M. Pierre Langlois1, and Paul Green2
1´Ecole Polytechnique de Montr´ eal, Canada
2Algolux, Canada
{ali.mosleh,pierre.langlois }@polymtl.ca, paul.green@algolux.com
Abstract. We present a new method to detect and remove ringing
artifacts producedbythedeconvolutionprocessinimagedeblurringtechniques. The method takes into account non-invertible frequency components of the blur kernel used in the deconvolution. E■cient Gabor
wavelets are produced for each non-invertible frequency and applied on
the deblurred image to generate a set of ■lter responses that reveal existing ringing artifacts. The set of Gabor ■lters is then employed in a

regularization scheme to remove th e corresponding artifacts from the deblurred image. The regularization scheme minimizes the responses of the reconstructed image to these Gabor ■lters through an alternating algorithm in order to suppress the artifacts. As a result of these steps we are able to signi■cantly enhance the quality of the deblurred images produced by deconvolution algorithms. Our numerical evaluations using a ringing artifact metric indicate the e■ectiveness of the proposed deringing method.

Keywords: deconvolution, image deblurring, point spread function, ringing artifacts, zero-magnitude frequency.

1

****

View-Consistent 3D Scene Flow Estimation over Multiple Frames

Christoph Vogel1,S t e f a nR o t h2, and Konrad Schindler1
1Photogrammetry and Remote Sensing, ETH Zurich, Switzerland
2Department of Computer Science, TU Darmstadt, Germany

Abstract. We propose a method to recover dense 3D scene ■ow from stereo video. The method estimat es the depth and 3D motion ■eld of ad y n a m i cs c e n ef r o m multiple consecutive frames in a sliding temporal window, such that the estimate is consistent across both viewpoints of all frameswithin the window. The observed scene is modeled as a collection ofplanarpatchesthatareconsistentacrossviews,eachundergoingarigid motion that is approximately constant over time. Finding the patchesand their mo tions is cast as minimization of an energy function over the continuous plane and motion parameters and the discrete pixel-to-plane assignment. We show that such a view-consistent multi-frame schemegreatly improv es scene ■ow computation in the presence of occlusions,and increases its robustn ess against adverse imaging conditions, such as specularities. Our method currently achieves leading performance on the KITTI benchmark, for both ■ow and stereo.

1

****

Hand Waving Away Scale

Christopher Ham1,S i m o nL u c e y2, and Surya Singh1
1Robotics Design Lab, The University of Queensland, Australia
2Robotics Institute, Carnegie Melon University, USA
{c.ham,spns }@uq.edu.au, slucey@cs.cmu.edu

Abstract. This paper presents a novel solution to the metric reconstruction of objects using any smart device equipped with a camera and an inertial measurement unit (IMU). We propose a batch, vision centric approach which only uses the IMU to estimate the metric scale of a scene reconstructed by any algorithm with Structure from Motion like (SfM) output. IMUs have a rich history of being combined with monocular vision for robotic navigation and odometry applications.These IMUs require sophisticated and quite expensive hardware rigs to perform well. IMUs in smart devices, however, are chosen for enhancing interactivity - a task which is more forgiving to noise in the measurements. We anticipate, however, that the ubiquity of these “noisy” IMUs makes them increasingly useful in modern computer vision algorithms. Indeed, we show in this work how an IMU from a smart device can help a face tracker to measure pupil distance, and anSfM algorithm to measure the metric size of objects. We also identify motions that produce better results, and develop a heuristic for estimating, in real-time, when enough data has been collected for an accurate scale estimation.

Keywords: Smart devices, IMU, metric, 3D reconstruction.

1

****

A Non-Linear Filter for Gyroscope-Based Video Stabilization

Steven Bell1, Alejandro Troccoli2, and Kari Pulli2

1Stanford University, Stanford, CA, USA
sebell@stanford.edu
2NVIDIA Research, Santa Clara, CA, USA
{atroccoli,karip }@nvidia.com
Abstract. We present a method for video stabilization and rolling-
shutter correction for videos captured on mobile devices. The method
uses the data from an on-board gyroscope to track the camera’s angular
velocity, and can run in real time within the camera capture pipeline. Weremove
small motions and rolling-shutter distortions due to hand shake,
creating the impression of a video shot on a tripod. For larger motions,
we ■lter the camera’s angular velocity to produce a smooth output. Tomeet the la
tency constraints of a real-time camera capture pipeline, our
■lter operates on a small temporal window of three to ■ve frames. Our
algorithm performs better than the previous work that uses a gyroscopeto stabili
ze a video stream, and at a similar level with respect to current
feature-based methods.
Keywords: video stabilization, rolling-shutter, gyroscopes.
1
****

Multi-modal and Multi-spectral Registration
for Natural Images
Xiaoyong Shen1,L iX u2, Qi Zhang1,a n dJ i a y aJ i a1
1The Chinese University of Hong Kong, China
2Image & Visual Computing Lab, Lenovo R&T,
Project Website, Hong Kong, China
http://www.cse.cuhk.edu.hk/leojia/projects/multimodal
Abstract. Images now come in di■erent forms – color, near-infrared,
depth, etc. – due to the development of special and powerful camerasin computer
vision and computational photography. Their cross-modal
correspondence establishment is however left behind. We address this
challenging dense matching problem considering structure variation pos-sibly exi
sting in these image sets and introduce new model and solution.
Our main contribution includes designing the descriptor named robust
selective normalized cross correlation (RSNCC) to establish dense pixelcorrespon
dence in input images and proposing its mathematical param-
eterization to make optimization tractable. A computationally robust
framework includingglobal and local matchingphases is also established.We build
a multi-modal dataset including natural images with labeled
sparse correspondence. Our method will bene■t image and vision appli-
cations that require accurate image alignment.
Keywords: multi-modal, multi-spectral, dense matching, variational
model.
1
****

Using Isometry to Classify
Correct/Incorrect 3D -2D Correspondences
Toby Collins and Adrien Bartoli
ALCoV-ISIT, UMR 6284 CNRS/Universit´ e d’Auvergne, Clermont-Ferrand, France
Abstract. Template-based methods havebeen successfully usedfor sur-
face detection and 3D reconstruction from a 2D input image, especiallywhen the s
urface is known to deform isometrically. However, almost all
such methods require that keypoint correspondences be ■rst matched
between the template and the input image. Matching thus exists as acurrent limit
ation because existing methods are either slow or tend to
perform poorly for discontinuous or unsmooth surfaces or deformations.
This is partly because the3D isometric deformation constraint cannot beeasily us
ed in the 2D image directly. We propose to resolve that di■culty
by detecting incorrect correspondences using the isometry constraint di-
rectly in 3D. We do this by embedding a set of putative correspondencesin 3D spa
ce, by estimating their depth and local 3D orientation in the

input image, from local image warps computed quicklyand accurately by means of Inverse Composition. We then relax isometry to inextensibilityto get a ■rst correct/incorrect classi■cation using simple pairwise constraints. This classi■cation is then e■ciently re■ned using higher-order constraints, which we formulate as the consistency between the corre-spondences' local 3D geometry. Our algorithm is fast and has only one free parameter governing the precision/recall trade-o■. We show experimentally that it signi■cantly outperforms state-of-the-art.
1
****

Bilateral Functions for Global Motion Modeling
Wen-Yan Daniel Lin1, Ming-Ming Cheng2, Jiangbo Lu1, Hongsheng Yang3, Minh N. Do4, and Philip Torr2
1Advanced Digital Sciences Center, Singapore
2Oxford University, UK
3University of North Carolina at Chapel Hill, USA
4University of Illinois at Urbana-Champaign, USA
Abstract. This paper proposes modeling motion in a bilateral domain that augments spatial information with the motion itself. We use the bilateral domain to reformulate a piecewise smooth constraint as continuous global modeling constraint. The resultant model can be robustly computed from highly noisy scattered feature points using a global minimization. We demonstrate how the model can reliably obtain large numbers of good quality correspondences over wide baselines, while keeping outliers to a minimum.
1
****

VCDB: A Large-Scale Database for Partial Copy Detection in Videos
Yu-Gang Jiang, Yudong Jiang, and Jiajun Wang
School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
ygj@fudan.edu.cn
Abstract. The task of partial copy detection in videos aims at ■nding if one or more segments of a query video have (transformed) copies in alarge dataset. Since collecting and annotating large datasets of real partial copies are extremely time-consuming, previous video copy detection research used either small-scale datasets or large datasets with simulatedpartial copies by imposing several pre-de■ned transformations (e.g., photometric or geometric changes). While the simulated datasets were useful for research, it is unknown how well the techniques developed on suchdata work on real copies, which are often too complex to be simulated. In this paper, we introduce a large-scale video copy database (VCDB) with over 100,000 Web videos, containing more than 9,000 copied segmentpairs found through careful manual annotation. We further benchmarka baseline system on VCDB, which has demonstrated state-of-the-art results in recent copy detection research. Our evaluation suggests that existing techniques—which have shown near-perfect results on the sim-ulated benchmarks—are far from satisfactory in detecting complex real copies. We believe that the release of VCDB will largely advance the research around this challenging problem.
Keywords: Video copy detection, benchmark dataset, frame matching, temporal alignment.
1
****

Single-Image Super-Resolution: A Benchmark
Chih-Yuan Yang1,C h a oM a1,2, and Ming-Hsuan Yang1
1University of California at Merced, USA
2Shanghai Jiao Tong University, China
{cyang35,cma26,mhyang }@ucmerced.edu

Abstract. Single-image super-resolution is of great importance for vi-
sion applications, and numerous algorithms have been proposed in recent
years. Despite the demonstrated success, these results are often gener-ated base
d on di■erent assumptions using di■erent datasets and met-
rics. In this paper, we present a systematic benchmark evaluation for
state-of-the-art single-image super-resolution algorithms. In addition toquantit
ative evaluations based on conventional full-reference metrics, hu-
man subject studies are carried out to evaluate image quality based on
visual perception. The benchmark evaluations demonstrate the perfor-mance and li
mitations of state-of-the-art algorithms which sheds light
on future research in single-image super-resolution.
Keywords: Single-image super-resolution, performance evaluation,
metrics, Gaussian blur kernel width.
1
****
Well Begun Is Half Done:
Generating High-Quality Seeds
for Automatic Image Dataset Construction from Web
Yan Xia1, Xudong Cao2, Fang Wen2,a n dJ i a nS u n2
1University of Science and Technology of China
2Microsoft Research Asia, Beijing, China
Abstract. We present a fully automatic approach to construct a large-scale, high
-
precision dataset from noisy web images. Within the entire pipeline, we focus on
generating high quality seed images for subsequent dataset growing. High quality
seeds are essential as we revealed, but they have received relatively less atten
tion
in previous works with respect to how to automatically generate them. In this
work, we propose a density score based on rank-order distance to identify positi
veseed images. The basic idea is images relevant to a concept typically are tigh
tly
clustered, while the outliers are widely scattered. Thr ough adaptive thresholdi
ng,
we guarantee the selected seeds as numerous and accurate as possible. Startingwi
th the high quality seeds, we grow a high quality dataset by dividing seeds and
conducting iterative negative and positive mining. Our system can automatically
collect thousands of images for one concept/class, with a precision rate of 95%o
r more. Comparisons with recent state-of-the-arts also demonstrate our method’s
superior performance.
1
****
Zero-Shot Learning via Visual Abstraction
Stanislaw Antol1, C. Lawrence Zitnick2, and Devi Parikh1
1Virginia Tech, Blacksburg, VA, USA
2Microsoft Research, Redmond, WA, USA
Abstract. Oneofthemainchallengesinlearning■ne-grainedvisualcat-
egories is gathering training images. Recent work in Zero-Shot Learning
(ZSL)circumventsthischallengebydescribingcategoriesviaattributesortext. However,
not all visual concepts, e.g., two people dancing, are easily
amenable to such descriptions. In this paper, we propose a new modal-
ity for ZSLusing visual abstraction to learn di■cult-to-describeconcepts.
Speci■cally, we explore concepts related to people and their interactions
withothers.Ourproposedmodalityallowsonetoprovidetrainingdataby
manipulating abstract visualizations, e.g., one can illustrate interactions
between two clipart people by manipulating each person’s pose, expres-
sion,gaze,andgender.Thefeasibilityofourapproachisshownonahuman
pose dataset and a new dataset containing complex interactions betweentwopeople,
whereweoutperformseveralbaselines.Tobettermatchacross
the two domains, we learn an explicit mapping between the abstract and
real worlds.

Keywords: zero-shot learning, visual abstraction, synthetic data, pose.
1
****
Zero-Shot Learning via Visual Abstraction 403
Moreover, our easy-to-use interface results in biases in the illustrations (e.g
.,
the interface does not allow for out-of-plane rotation). To account for these hu
-man tendencies, as well as interface biases, we learn an explicit mapping from
the features extracted from illustrations to the features extracted from real im
-
ages. This allows us to improve performance on instance-level ZSL. Our visualabs
traction interface, code, and datasets are publicly available.
2 Related Work
We discuss existing work on zero-shot learning, learning with synthetic data,
learning semantic relations, pose estimation, and action recognition.
Zero-Shot Learning (ZSL): Theproblemoflearningmodelsofvisualconcepts
without example images of the concepts is called Zero-Shot Lea rning. Attributes
(mid-level,visual,and semantic features)[9,10,15,16] providea naturalinterface
for ZSL [16], where an unseen class is descr ibed by a list of attributes. Equip
ped
with asetofpre-trainedattribute classi■ers,atest imagecanbe probabilistically
matched to each of these attribute signatures and be classi■ed as the category
with the highest probability. Instead of using a list of attributes, recent work
[7]has leveraged more general textual descriptions of categories to build visua
l
models of these categories. Our work takes a fundamentally di■erent approach
to ZSL. We propose a strictly visual modality to allow a supervisor to train a
model for visual concepts that may not be easily describable in semantic terms,
e.g., poses of people, interactions between people.
Learning With Synthetic Data: Our work introduces the use of abstract
visualizations as a modality to train visual models in a ZSL setting. Previously
,
papers have explored the use of synthetic data to aid in the training of vision
al-
gorithms.In many object recognitiontasks, it is commonto perturb the training
data using a■ne warps to augment the training data [14]. Computer-generated
scenes may also be used to evaluate recognition systems [13]. Shotton et al.[2
3]
usedsynthetically generateddepth datadepicting humansto learnahumanpose
detector from this depth data. Unlike these approaches, we are trying to learn
high-level, complex concepts where it i s not feasible to automatically generate
synthetic data, so we must rely on humans to create our synthetic data. Most
similar to our work, the problem of semantic scene understanding using abstract
scenes was studied in [31]. They use a dat aset of simple sentences correspondin
g
toabstractscenestolearnamappingfrom sentencestoabstractscenes.Recently,
sequences of abstract scenes were used to predict which objects will move in the
near future [11]. Unlike these works, we use abstraction to learn visual models
that can be applied to realimages. Sketch-based image retrieval [6,24] allows
users to search for an image by sketching t he concept. Sketching complex inter-
actions between people would be time consuming, and likely inaccurate for most
lay users. More importantly, our modality has the potential to augment the ab-
stract scenes with a large variety of visual cues (e.g., gender, ethnicity, clo
thing,
background) that would be cumbersome for users to convey via sketches
****
Zero-Shot Learning via Visual Abstraction 405
(e.g., somebodypartscanbecroppedout),whichmakesthisdatasetchallenging
(e.g., right side of Figure 1). This resulted in 3,600 initial images.
Real Image Annotations: We also used AMT to collect various image an-

notations that are needed for our featur es via di■erent custom interfaces. The pose annotation interface prompted the worker with one of our images and its corresponding sentence. We highlighted whether the worker should be annotating Person A or Person B in the sentence. The worker annotates the person's 14body parts (right side of Figure 3). The worker provides their best guess if the part is occluded and responds "not present" if it is not within the image border.

We had 5 workers annotate each person in each image and averaged them forthe ■nal ground truth pose annotations. In addition, workers annotated ground truth eyegaze (i.e., lookingto the imageleft orright),facialexpression(i.e., one of six prototypic emotionalexpressions[5] plus a neutral expression),and gender of each person via separate interfaces. W e selected the mode of their responses for our ■nal annotation. In addition to collecting the annotation of interest, t wointerfaces asked one additional question each. One asked if the prompted image contained exactly two main people or not and the other asked if the annotated pose overlaid on the prompted image was of good quality or not. We used thelast two annotation queries to remove poor quality work. Additionally, a GIST-based [20] image matching scheme was used to remove duplicates. Removing these images gave us our ■nal annotated dataset with 3,172 images (52.9 images per category on average). Some examples can be found in the bottom part of Figure 1 and the rightmost two columns of Figure 5. More details about ourinterfaces and our procedure can be found in the supplementary material.

3.2 PARSE

We also use a subset of the standard PARSE [21] dataset, which originally contains 305 images of individuals in various poses. We created a list of categories that frequently appear in the PARSE dataset (e.g., "is dunking," "is diving for an object"). From the images that belong to these categories, we removed those that were used to train the pose detector [26]. Some categories (e.g., "is standing") had disproportionately large number of images, so we removed images at random from these categories. This leaves us with 108 images in our dataset (7.7 images per category on average). We also collected the same annotations as in Section 3.1, except for pose (since ground truth pose annotations are already available with the dataset). See the supplementary material for more details.

4 Our Approach

In this section, we present our new modality for ZSL. We begin by introducingour user interface for collecting visual illustrations for training. We then describe the novel features that are extracted from our abstract illustrations and real images. Finally, we describe the approach used to train our models. The resultsof various experiments follow in Section 5

****

Fig. 2.User interface (with random initialization) used to collect abstract illustrations on AMT. Workerswere able to manipulate pose, expression, gaze direction, andgender.

4.1 Visual Abstraction Interface

For our domain of interest, we conjectured that our concepts depend primarily on four main factors: pose, eye gaze, facial expression, and gender. Some other factors that we do not model, but may also be important are clothing, thepresence of other objects, and scene conte xt. A screenshot of our user interface is shown in Figure 2. Initially, two people (one blond-haired and one brown-haired) are shown with random poses, gaze directions (i.e., "■ip"), expressions, and genders. We allow our subjects to continuously manipulate the poses (i.e., joint angles and positions) of both people by dragging on the various bodyparts. They may horizontally ■ip the people to change their perceived eye gaze

direction. The facial expressions are chosen from the same selection as is used for the annotation of real images (Section 3.1). Finally, the subjects may selec tone of the two predominant genders for each clipart person.
To collect our training data for category-level ZSL, we prompt the user with a sentence to illustrate using the interface (e.g., “Person A is dancing with Person B.”, “A person is dunking.”). To promote diversity, we encouraged them to imagine any objects or background, as long as the poses are consistent withth e imagined scene (e.g., a worker can imagine a chair and illustrate someone sitting on it). The interface includes buttons to annotate which clipart person correspondsto which personin the sentence. Some illustrations areshown on thelef t side of Figure 1 and in the left three columns of Figure 5. For the PARSE concepts, the interface is the same ex cept that only one person is present.
For instance-level ZSL, we modify our prev iousinterface. Instead ofsentences, we ■rst (brie■y, for 2 seconds) show the user a real image and then they recreat e
it (from memory) as best they can. The stated goal is to recreate the real image so another person would be able to select the shown image from a collection of real images. This mimics the scenario wh en a person is searching for a speci■c image: they might be clear on the semantically important aspects while having af uzzier or skewed notion of other aspects. Another bias of the illustrations occu rs
when it is impossible to recreate the real image exactly due to the limitations of
the interface, such as not being able to change the height of the clipart people ,the interface not allowing for out-of-plane rotation, etc
****
Zero-Shot Learning via Visual Abstraction 407
4.2 Relation and Appearance Features
Using the annotations described in Section 3.1 (i.e., pose, gaze, expression, a nd
gender) for persons denoted by iandj, we compute a set of relation and ap-
pearance features. Some of our relation features are distance-basedand some are angle-based. All distance-based feature s use Gaussians placed at di■erent posi-tions to capture relative distance. The Gaussians’ σparameters are proportional to thescaleof each person. A person’s scale is de■ned as the distance between their head and the center of their shoulders and hips. Unless otherwise noted, all angles/orientations are w.r.t. the image frame’s x-axis. They are represente d
by 12 dimensional unit histograms with each bin corresponding to π/6 radians. Soft assignments are made to the histograms using linear weighting. The ■rst two sets of features, Basic and Gaze, a ccount for both people. The remaining ■ve feature sets are described for a sing le person and must be evaluated twice (swappingiandj) and concatenated. The feature sets are described below.
Basic:This feature set encodes basic rela tion properties between two people, such as relative orientation and distance. We calculate each person’s body angle (in the image frame). This is calculated from the image coordinates for the head andmid-pointbetweenshoulders.We placeGaussiansatthe centerofthe peopleand then u se the distance between them to evaluate the Gaussian functions. We
also calculate the angle (in the image frame) between the centers of the two people. This gives us a total of 2 ∗(12 + 1) + 12 = 38 features. They can be thought of as simplifying the people into two boxes (possibly having di■erent scale parameters) with certain orientations and looking at the relative position sand angle between their centers.
Gaze:The gaze feature set is encoded using 5 binary features, corresponding toilooking atj,jlooking ati, both people are looking at each other, both people are looking away from each other, and both people are looking in the same direction. To determine if iis looking at j,w ec h e c ki f j’s neck is in the
appropriate region oftheimage.Theimageisdividedintotwopartsbyextending the line between i’s head and neck and the appropriate region is de■ned to be

the area where iis looking (which depends on i’s gaze direction). Once we have bothilooking atjandvice versa features, we compute the remaining three gaze features via the appropriate logic operations (e.g., ifiis looking at jandjis looking ati, then the looking-at-each-other feature is true).

Global: This feature set encodes the general position of the joints in reference to a body. Three Gaussians are placed in a 3 ×1 grid on the image based on the body’s size and orientation (the blue circles in Figure 3). The positions of one person’s 8 joints (two for each limb) are evaluated using all Gaussians from both Gaussian sets (
i.e., personi’s joints relative to person i’s global Gaussians
and personj’s global Gaussians), giving us a total of 8 ∗3∗2 = 48 features.

Contact: This feature set encodes the speci■c location of the joints in referenc e
to other body parts. For each person, we place Gaussians at 13 positions:

****

if an image represents a speci■c concept, i.e., given a test real image, we wish to determine which speci■c abstract visu alization (instance) corresponds to the real image. For this, we use Nearest Neighbor matching. Since our features are from two di■erent domains, learning a mapping between them could improve the matching performance. This is described next.

4.4 Mapping From Abstract to Real for the Instance-level Model

We learn a mapping between the domain of abstract images and the domain of real images. To learn such a mapping, we need examples that correspond to the same thing in both domains. We use some of our instance-level illustrations (Section 4.1) as these abst ract-real pairs. The mappi ng can learn to correct f or
both user and interface biases discussed in Section 4.1.

Simpler techniques, such as Canonical Correspondence Analysis [12], did not learn a good mapping between the abstract and real worlds. We found that General Regression Neural Networks (GRNN) [25] did better. We also found that converting from our abstract features into “real” features performed better than converting real features into “abstract” features. Thus, the GRNN’s input is all of the abstract features and its output is all of the real features.

5 Experimental Results

In this section, we describe our experiments which show that our new modality for ZSL is able to create models that can l earn category-level (Section 5.1) an d
instance-level (Section 5.2) visual concepts. We perform an ablation study on di■erent feature sets, showing their performance contribution (Section 5.3). Fi-
nally,weutilizeastate-of-the-artposedetectoronbothINTERACTandPARSE
datasets to investigate our approach in a more automatic setting (Section 5.4).

5.1 Category-Level Zero-Shot Learning

We begin by experimenting with the ability of our novel modality to learn our category-level concepts, i.e., classifying images into one of the semantic descr ip-
tions, such as “A is kicking B.” To acquire the required training illustrations, we ran our visual abstraction interface with sentence prompts (described in Sec-
tion 4.1) on AMT. We had 50 workers create an abstract illustration for each oft he 60 semantic concepts from INTERACT (Section 3.1) and the 14 semantic concepts from PARSE (Section 3.2). Af ter removing poor quality work, we are left with 3,000 and 696 illustrations, respectively.

The setup for all category-level ZSL experiments (unless otherwise noted) is described here. Using the abstract illustrations, we train multiple one-vs-al llinear SVMs (liblinear [8]) with the cost parameter, C, set to 0.01, which work ed
reasonably well across all experiments. For INTERACT, there is ambiguity (at test time) as to which person is Person A and which person is Person B. Toaccoun tforthis,weevaluateeachoftheclassi■ersusingbothorderings,selectth

****

5.2 Instance-Level Zero-Shot Learning

We also test the ability of our new modality to learn instance-level concepts. To acquire the necessary training illustrations, we ran our visual abstractionin terface with image prompts (as described in Section 4.1) on AMT. We showed a real image (one of 3,172 from INTERACT and one of 305 from PARSE) for two seconds to the workers,w ho recreated it using the i nterface. Through a pil ot study, just as in [6], we found two seconds to be su■cient for people to capture the more salient aspects of the image. It is unlikely that a user would have eve rydetail of the instance in mind when trying to train a model for a speci■c conc ept and we wanted to mimic this in our set up. We had 3 workers recreate each of the images, and after manually removing work from problematic workers,we areleft with 8,916 and 914 illustrations for INTERACT and PARSE, respectively.

We perform classi■cation via nearest-neighbor matching. If the real image’s features match the features of any of the (up to) 3 illustration instances that workers created for it, we have found a correct label. We vary K, the number of nearestneighborsthat areconsidered,andevaluatethepercentageofrealimagesthat have a correctlabel within those Kn eighbors.We normalizedK by the total number of illustrations. We need a training dataset to learn a mapping between the abstract and real worlds, i.e., training the GRNN from Section 4.4. For IN-TERACT, we split the categories into 39 seen categories for training and 21 unseen categories for testing to minimize l earning biases speci■c to speci■c ca te-gories (i.e., verb phrases). The results are averaged over 10 random seen/unseen category splits. For PARSE, the training data corresponds to the 197 images

Fig. 5.The left columns show 5 random illustrations (of 50) used for classi■er t raining. Columns 6 and 7 contain the most con■dent true positive and false positive for a givencategory, respectively. Mistakes include choosing a semantically reasonabl e verb (top), choosing the incorrect preposition (middle), and incorrect prediction due to the pose similarity between two classes (bottom). More examples are in the supplement

****

0481216

Random

B

C

G

O

B+C

B+G

B+O

B+E+Z+S

B+C+G+O

B+C+G+O+E

C+G+O+E+Z+S

B+G+O+E+Z+S

B+C+O+E+Z+S

B+C+G+E+Z+S

B+C+G+O+Z+S

B+C+G+O+E+S

B+C+G+O+E+Z

B+C+G+O+E+Z+SPP

YR-BB

Random

1

1Mean of Class-wise
Raw Accuracies (%) INTERACT
Features:
(B)asic
(C)ontact
(G)lobal
(O)rientation
(E)xpression
Ga(Z)e
S for Gender
Fig. 7.Weplotclassi■cation performanceforINTERACTusingdi■erentsubsetsoffeatures. Some features, like Global, are more informative than others. Of the appearance-
based features, Expression turns out to be most informative, presumably when bodypose features are similar (e.g., “wrestling” vs. “hugging”).
5.4 Automatic Pose Evaluation
In this section, we do an evaluation of our category-level ZSL task using the current state-of-the-art pose detector developed by Yang and Ramanan [27]. We utilized the pre-trained PARSE mod el and detected the pose on both the INTERACT and the PARSE datasets. For the expression, gaze, and gender features, we continue to use human annotations. These results (YR) are shown in Figures 4, 6, and 7. As expected, due to the pose detector being developed for PARSE, automatic detection on the PARSE dataset yields reasonable performance (compared to perfect pose). Th e results on INTERACT do not perform nearly as well, although it still outperforms the baselines. To boost the perfor-
manceofthe posedetectoron INTERACT, wealsoexperimented with providing ground truth bounding boxes (YR-BB), which results in better performance. INTERACT is signi■cantly more challengingthan PARSE for automatic pose detection. Thus, it is not surprising that incorrectly detected poses confuse our
models. Properties that make INTERACT particularly challenging include: im-ages from arbitrary perspectives, mor e di■cult (for the detector) poses (e.g., “crawling,” “lying”), overlapping people (e.g., “hugging,” “standing in front of”), and incomplete poses (i.e., not all body parts are present). We investigated
this latter point by selecting images from INTERACT based on the number of parts present in the image. There are 14 parts per person and we ensure that both people have at least a certain number of parts. Requiring all parts to be withintheimagereducesINTERACTto1,689images(from3,172).91 .5%ofour
images contain at least 7 parts per person. More of these details can be found inthe supplementary material. We re-evaluate our category-levelZSL performance (at 50 trainingillustrationsper category)as wevary the part threshold and show our results in Figure 8. Although there is some noise, both the perfect pose and automatic pose detection methods show an increase in accuracy as we requir
****

References
1. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and
multiple instance learning. PAMI (2010)
2. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose
annotations. In: ICCV (2009)
3. Chakraborty, I., Cheng, H., Javed, O.: 3d visual proxemics: Recognizing human interactions in 3d from a single image. In: CVPR (2013)
4. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recog-
nition. In: CVPR (2011)
5. Darwin, C.: The Expression of the Emotions in Man and Animals. Oxford

University Press (1998)
6. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: Sketch-based image retrie val:
Benchmark and bag-of-features descriptors. IEEE Transactions on Visualizationand Computer Graphics (2011)
7. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classi■er: Zero-shot learning using
purely textual descriptions. In: ICCV (2013)
8. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR:
A library for large linear classi■cation. JMLR (2008)
9. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: ICCV (2009)
10. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS (2007)11. F ouhey,D.F., Zitnick, C.L.: Predicting object dynamics in scenes. In: CVPR (2014)
12. Hotelling, H.: Relations between two sets of variates. Biometrika (1936)
13. Kaneva, B., Torralba, A., Freeman, W.T.: Evaluation of image features using a
photorealistic virtual world. In: ICCV (2011)
14. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classi■cation with deep convo-
lutional neural networks. In: NIPS (2012)
15. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile cl assi-
■ers for face veri■cation. In: ICCV (2009)
16. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
17. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for
contextual group activities. In: NIPS (2010)
18. Larochelle, H., Erhan, D., Bengio, Y.: Zero-data learning of new tasks. In: AAAI (2008)
19. Marin-Jimenez, M., Zisserman, A., Eichner, M., Ferrari, V.: Detecting people look-
ing at each other in videos. IJCV (2013)
20. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic represe ntation
of the spatial envelope. IJCV (2001)
21. Ramanan, D.: Learning to parse images of articulated bodies. In: NIPS (2007)
22. Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: CVPR (2011)
23. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from singledept h images. In: CVPR (2011)
24. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-ba sed
image retrieval at the end of the early years. PAMI (2000)
25. Specht, D.F.: The general regression neural network-rediscovered. Neural Net works
(1993
****

Discovering Groups of People in Images

Wongun Choi1, Yu-Wei Chao2, Caroline Pantofaru3, and Silvio Savarese4
1NEC Laboratories, USA
2University of Michigan, Ann Arbor, USA
3Google, Inc, USA
4Stanford University, USA

Abstract. Understanding group activities from images is an important yet chal-
lenging task. This is because there is an exponentially large number of semantic and geometrical relationships among individuals that one must model in order to effectively recognize and localize the gr oup activities. Rather than focusing o

n di-
rectly recognizing group act ivities as most of the previous works do, we advoca te
the importance of introducing an intermediate representation for modeling groups of humans which we call structure groups. Such groups de■ne the way people spatially interact with each other. People might be facing each other to talk, w hileothers sit on a bench side by side, and some might stand alone. In this pape r wecontribute a method for identifying and localizing these structured groups i n a
single image despite their varying viewpoints, number of participants, and oc-
clusions. We propose to learn an ensemble of discriminative interaction patterns to encode the relationships between people in 3D and introduce a novel ef■cient iterative augmentation algorithm for solving this complex inference problem. A nice byproduct of the inference scheme is an approximate 3D layout estimate ofth e structured groups in the scene. Finally, we contribute an extremely challeng-
ing new dataset that contains images each showing multiple people performing multiple activities. Extensive evaluation con■rms our theoretical ■ndings.
Keywords: Group discovery, Social interaction, Activity recognition.
1
****

Untangling Object-View Manifold
for Multiview Recognition and Pose Estimation
Amr Bakry and Ahmed Elgammal
Department of Computer Science, Rutgers University
Piscataway, NJ, USA
Abstract. The problem of multi-view/view-invariant recognition
remains one of the most fundamental challenges to the progress of the
computer vision. In this paper we consider the problem of modeling thecombined o bject-viewpoint manifold. The shape and appearance of an
object in a given image is a function of its category, style within category,
viewpoint, and several other factors. The visual manifold (in any cho-sen featur e representation space) given all these variability collectively
is very hard and even impossible to model. We propose an e■cient com-
putational framework that can untangle such a complex manifold, andachieve a mod el that separates a view-invariant category representation,
from category-invariant pose representation. We outperform the state of
the art in the three widely used multiview dataset, for both categoryrecognition , and pose estimation.
1
****

Parameterizing Object Detectors
in the Continuous Pose Space
Kun He1, Leonid Sigal2, and Stan Sclaro■1
1Computer Science Department, Boston University, USA
2Disney Research Pittsburgh, USA
{hekun,sclaroff }@cs.bu.edu, lsigal@disneyresearch.com
Abstract. Object detection and pose estimation are interdependent
problems in computer vision. Many past works decouple these problems,either by d iscretizing the continuous pose and training pose-speci■c ob-
ject detectors, or by building pose estimators on top of detector outputs.
In this paper, we propose a structured kernel machine approach to treatobject de tection and pose estimation jointly in a mutually beni■cial way.
In our formulation, a uni■ed, continuously parameterized, discriminative
appearance model is learned over the entire pose space. We propose acascaded dis crete-continuous algorithm for e■cient inference, and give
e■ective online constraint generation strategies for learning our model
using structural SVMs. On three standard benchmarks, our method per-forms better than, or on par with, state-of-the-art methods in the com-
bined task of object detection and pose estimation.
Keywords: object detection, continuous pose estimation.

1
****
Jointly Optimizing 3D Model Fitting
and Fine-Grained Classi■cation
Yen-Liang Lin1, Vlad I. Morariu2, Winston Hsu1, and Larry S. Davis2
1National Taiwan University, Taipei, Taiwan
2University of Maryland, College Park, MD, USA
yenliang@cmlab.csie.ntu.edu.tw, whsu@ntu.edu.tw,
{morariu,lsd }@umiacs.umd.edu
Abstract. 3D object modeling and ■ne-grained classi■cation are often
treated as separate tasks. We propose to optimize 3D model ■tting and■ne-grained
classi■cation jointly. Detailed 3D object representations en-
code more information (e.g., precise part locations and viewpoint) than
traditional 2D-based approaches, and can therefore improve ■ne-grainedclassi■cat
ion performance. Meanwhile, the predicted class label can also
improve 3Dmodel ■tting accuracy, e.g., by providingmore detailed class-
speci■c shape models. We evaluate our method on a new ■ne-grained3D car dataset
(FG3DCar), demonstrating our method outperforms sev-eral state-of-the-art approa
ches. Furthermore, we also conduct a series
of analyses to explore the dependence between ■ne-grained classi■cation
performance and 3D models.
1
****
Pipelining Localized Semantic Features
for Fine-Grained Action Recognition
Yang Zhou1, Bingbing Ni2, Shuicheng Yan3, Pierre Moulin4, and Qi Tian1
1University of Texas at San Antonio, USA
2Advanced Digital Sciences Center, Singapore
3National University of Singapore, Singapore
4University of Illinois at Urbana-Champaign, USA
myh511@my.utsa.edu, bingbing.ni@adsc.com.sg, eleyans@nus.edu.sg,
moulin@ifp.uiuc.edu, qi.tian@utsa.edu
Abstract. In ■ne-grained action (object manipulation) recognition, it
is important to encode object semantic (contextual) information, i.e.,which obje
ct is being manipulated and how it is being operated. How-
ever, previous methods for action rec ognition often represent the seman-
tic information in a global and coarse way and therefore cannot copewith ■ne-gra
ined actions. In this work, we propose a representation and
classi■cation pipeline which seamlessly incorporates localized semantic
information into everyprocessing step for ■ne-grained action recognition.In the
feature extraction stage, we e xplore the geometric information
between local motion features and the surrounding objects. In the fea-
ture encoding stage, we develop a semantic-grouped locality-constrainedlinear co
ding (SG-LLC) method that captures the joint distributions
between motion and object-in-use information. Finally, we propose a
semantic-aware multiple kernel learning framework (SA-MKL) by uti-lizing the emp
irical joint distribution between action and object type
for more discriminative action classi■cation. Extensive experiments are
performed on the large-scale and di■cult ■ne-grained MPII cooking ac-tion datase
t. The results show that by e■ectively accumulating localizedsemantic informatio
n into the action representation and classi■cation
pipeline, we signi■cantly improve the ■ne-grained action classi■cation
performance over the existing methods.
1
****
Robust Scene Text Detection with Convolution
Neural Network Induced MSER Trees
Weilin Huang1,2, Yu Qiao1, and Xiaoou Tang2,1
1Shenzhen Key Lab of Comp. Vis and Pat. Rec.,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

2Department of Information Engineering,
The Chinese University of Hong Kong, China
Abstract. Maximally Stable Extremal Regions (MSERs) have achieved
great success in scene text detection. However, this low-level pixel opera-tioni
nherentlylimits itscapability for handlingcomplex textinformation
e■ciently (e. g. connections between text or background components),
leading to the di■culty in distinguishing texts from background compo-nents.In t
his paper,we propose anovel framework to tackle thisproblem
by leveraging the high capability of convolutional neural network (CNN).
In contrast to recent methods using a set of low-level heuristic features,the CN
N network is capable of learning high-level features to robustly
identify text components from text-like outliers (e.g. bikes, windows,
or leaves). Our approach takes advantages of both MSERs and sliding-window based
 methods. The MSERs operator dramatically reduces the
number of windows scanned and enhances detection of the low-quality
texts. While the sliding-window with CNN is applied to correctly sepa-rate the c
onnections of multiple characters in components. The proposedsystem achieved str
ong robustness against a numberof extreme textvari-
ations and serious real-world problems. It was evaluated on the ICDAR
2011 benchmark dataset, and achieved over 78% in F-measure, which issigni■cantly
 higher than previous methods.
Keywords: Maximally Stable Extremal Regions (MSERs), convolu-
tional neural network (CNN), text-like outliers, sliding-window.
1
****

Deep Features for Text Spotting
Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman
Visual Geometry Group, Department of Engineering Science,
University of Oxford, UK
Abstract. The goal of this work is text spotting in natural images.
This is divided into two sequential tasks: detecting words regions in the
image, and recognizing the words within these regions. We make the
following contributions: ■rst, we develop a Convolutional Neural Net-work (CNN)
classi■er that can be used for both tasks. The CNN has
a novel architecture that enables e■cient feature sharing (by using a
number of layers in common) for text detection, character case-sensitiveand inse
nsitive classi■cation, and bigram classi■cation. It exceeds the
state-of-the-art performance for all of these. Second, we make a number
of technical changes over the traditional CNN architectures, includingno downsam
pling for a per-pixel sliding window, and multi-mode learn-
ing with a mixture of linear models (maxout). Third, we have a method
of automated data mining of Flickr, that generates word and characterlevel annot
ations. Finally, these components are used together to form
an end-to-end, state-of-the-art text spotting system. We evaluate the
text-spotting system on two standard benchmarks, the ICDAR RobustReading data se
t and the Street View Text data set, and demonstrate
improvements over the state-of-the-art on multiple measures.
1
****

Improving Image-Sentence Embeddings Using
Large Weakly Annotated Photo Collections
Yunchao Gong1,L i w e iW a n g2,M i c a hH o d o s h2, Julia Hockenmaier2,
and Svetlana Lazebnik2
1University of North Carolina at Chapel Hill, USA
yunchao@cs.unc.edu
2University of Illinois at Urbana-Champaign, USA
{lwang97,mhodosh2,juliahmr,slazebni }@illinois.edu
Abstract. This paper studies the problem of associating images with
descriptive sentences by embedding them in a common latent space. Weare interest
ed in learning such embeddings from hundreds of thousands

or millions of examples. Unfortunately, it is prohibitively expensive to fully annotate this many training images with ground-truth sentences.Instead, we ask whether we can learn better image-sentence embeddings by augmenting small fully annotated training sets with millions of images that have weak and noisy annotations (titles, tags, or descriptions).After investigating several state-of-the-art scalable embedding methods, we introduce a new algorithm called Stacked Auxiliary Embedding that can successfully transfer knowledge from millions of weakly annotatedimages to improve the accuracy of retrieval-based image description.

1

****

Strengthening the E■ectiveness of Pedestrian Detection with Spatially Pooled Features

Sakrapee Paisitkriangkrai, Chunhua Shen*, and Anton van den Hengel
The University of Adelaide, Australia
chunhua.shen@adelaide.edu.au

Abstract. We propose a simple yet e■ective approach to the problem of pedestrian detection which outperforms the current state-of-the-art. Our new features are built on the basis of low-level visual features and spatial pooling. Incorporating spatial pooling improves the translational invariance and thus the robustness of the detection process. We then directly optimise the partial area under the ROC curve (pAUC) measure, which concentrates detection performance in the range of most practical importance. The combination of these factors leads to a pedestrian detector which outperforms all competitors on all of the standard benchmark datasets. We advance state-of-the-art results by lowering the average miss rate from 13% to 11% on the INRIA benchmark, 41% to 37% on the ETH benchmark, 51% to 42% on the TUD-Brussels benchmark and 36% to 29% on the Caltech-USA benchmark.

1

****

Selecting In■uential Examples: Active Learning with Expected Model Output Changes

Alexander Freytag■, Erik Rodner■, and Joachim Denzler
Computer Vision Group, Friedrich Schiller University Jena, Germany
{firstname.lastname }@uni-jena.de
http://www.inf-cv.uni-jena.de

Abstract. In this paper, we introduce a new general strategy for active learning. The key idea of our approach is to measure the expected change of model outputs, a concept that generalizes previous methods based on expected model change and incorporates the underlying data distribution. For each example of an unlabeled set, the expected change of model predictions is calculated and marginalized over the unknown label. This results in a score for each unlabeled example that can be used for active learning with a broad range of models and learning algorithms. In particular, we show how to derive very ef■cient active learning methods for Gaussian process regression, which implement this general strategy, and link them to previous methods. We analyze our algorithms and compare them to a broad range of previous active learning strategies in experiments showing that they outperform state-of-the-art on well-established benchmark datasets in the area of visual object recognition.

Keywords: active learning, Gaussian processes, visual recognition, exploration-exploitation trade-off.

1

****

Ef■cient Sparsity Estimat ion via Marginal-Lasso Coding

Tzu-Yi Hung1,J i w e nL u2, Yap-Peng Tan1, and Shenghua Gao3
1School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore
2Advanced Digital Sciences Center, Singapore
3ShanghaiTech University, Shanghai, China

Abstract. This paper presents a generic optimization framework for ef■cient feature quantization using sparse coding which can be applied to many computer vi-s ion tasks. While there are many works working on sparse coding and dictionary learning, none of them has exploited the advantages of the marginal regression and the lasso simultaneously to provide more ef■cient and effective solutions. I nour work, we provide such an approach with a theoretical support. Therefore, th e
computational complexity of the proposed method can be two orders faster than that of the lasso with sacri■cing the inevitable quantization error. On the othe rhand, the proposed method is more robust than the conventional marginal regres-sion based methods. We also provide an adaptive regularization parameter se-lection scheme and a dictionary learning method incorporated with the proposedsp arsity estimation algorithm. Experimental results and detailed model analysis are presented to demonstrate the ef■cacy of our proposed methods.

Keywords: Sparsity estimation, marginal regression, sparse coding, lasso, dic-tionary learning, adaptive regularization parameter.

****

Continuous Conditional Neural Fields for Structured Regression

Tadas Baltru■ saitis1, Peter Robinson1, and Louis-Philippe Morency2
1Computer Laboratory, University of Cambridge, UK
{tadas.baltrusaitis,peter.robinson }@cl.cam.ac.uk
2Institute for Creative Technologies, University of Southern California, CA
morency@ict.usc.edu

Abstract. An increasing number of computer vision and pattern recog-nition problems require structured regression techniques. Problems likehuman pos e estimation, unsegmented action recognition, emotion pre-diction and facial landmark detection have temporal or spatial output dependencies that regular regression techniques do not capture. In thispaper we present continuous conditional neural ■elds (CCNF) – a novel structured regression model that can learn non-linear input-output de-pendencies, and model temporal and spatial outputrelationships of vary-ing lengt h sequences. Wepropose two instances of our CCNF framework: Chain-CCNF for time series modelling, and Grid-CCNF for spatial rela-tionship modelling. We evaluate our model on ■ve public datasets span-ning three di■erent regression problems: facial landmark detection in the wild, emotion prediction in music and facial action unit recognition. Our CCNF model demonstrates state-of-the-art performance on all of thedatasets used.

Keywords: Structured regression, Landmark detection, Face tracking.

****

Learning to Rank Using High-Order Information

Puneet Kumar Dokania1, Aseem Behl2,C . V .J a w a h a r2, and M. Pawan Kumar1
1Ecole Centrale de Paris
INRIA Saclay, France
2IIIT Hyderabad, India

Abstract. The problem of ranking a set of visual samples according to their relevance to a query plays an important role in computer vision. The traditional approach for ranking is to train a binary classi■er such asa sup port vector machine (svm). Binary classi■ers su■er from two main de■ciencies: (i) they do not optimize a ranking-based loss function, for example, the average precision (ap) loss; and(ii) theycannot incorporate

high-order information such as the ap r i o r icorrelation between the relevance of two visual samples (for example, two persons in the same image tend to perform the same action). We propose two novel learning formu-lations th at allow us to incorporate high-order information for ranking.
The ■rst framework, called high-order binary svm(hob-svm), allows for a structured input. The parameters of hob-svm are learned by minimizing a convex upper bound on a surrogate 0-1 loss function. In order to obtain the ranking of the samples that form the structured input, hob-svmsorts the samples according to their max-marginals . The second framework, called high-order average precision svm(hoap-svm), also allows for a structured input and uses the same ranking criterion. However, in contrast to hob-svm , theparameters of hoap-svm are learned by minimizing a di■erence-of-convex upper bound on the aploss. Using a standard, publicly available dataset for the challenging problem of action classi■cation, we show thatboth hob-svm andhoap-svm outperform the baselines that ignore high-order information.
1
****

Support Vector Guided Dictionary Learning
Sijia Cai1,3, Wangmeng Zuo2,L e iZ h a n g3,■, Xiangchu Feng4, and Ping Wang1
1School of Science, Tianjin University, China
2School of Computer Science and Technology, Harbin Institute of Technology, Chin a
3Dept. of Computing, The Hong Kong Polytechnic University, China
4Dept. of Applied Mathematics, Xidian University, China
cssjcai@gmail.com, cslzhang@comp.polyu.edu.hk
Abstract. Discriminative dictionary learning aims to learn a dictionary from training samples to enhance the discriminative capability of their coding vectors. Several discrimination terms have been proposed by assessing the prediction loss (e.g., logistic regression) or class separation criterion (e.g., Fisher discrimination criterion) on the coding vectors. In this paper, we provide a new insight on discriminative dictionary learning. Speci■cally, we formulate the discrimination term as the weighted summation of the squared distances between all pairs of coding vectors. Thediscrimination term in thestate-of-the-art Fisherdiscrimination dictionary learning (FDDL)methodcanbeexplainedas aspecial case ofour model, where the weights are simply determined by the numbers of samples of each class. We then propose a parameterization method to adaptively determine the weight of each coding vector pair, which leads to a support vector guided dictionary learning (SVGDL) model. Compared with FDDL, SVGDL can adaptively assign di■erent weights to di■erent pairs of coding vectors. More importantly, SVGDL automatically selects only a few critical pairs to assign non-zero weights, resulting in better generalization ability for pattern recognition tasks. The experimental results on a series of benchmark databases show that SVGDL outperforms many state-of-the-art discriminative dictionary learning methods.
Keywords: Dictionary learning, support vector machine, sparse representation, Fisher discrimination.
1
****

Video Object Discovery and Co-segmentation
with Extremely Weak Supervision
Le Wang1,■,G a n gH u a2, Rahul Sukthankar3, Jianru Xue1, and Nanning Zheng1
1Xi’an Jiaotong University, China
2Stevens Institute of Technology, USA
3Google Research, USA
Abstract. Video object co-segmentation refers to the problem of simultaneously segmenting a common category of objects from multiple videos. Most existing video co-segmentation methods assume that all frames from all videos contain the target objects. Unfortunately, this assumption is rarely true in practice, p

ar-
ticularly for large video sets, and existing methods perform poorly when the assumption is violated. Hence, any practical video object co-segmentation algorithm needs to identify the relevant frames containing the target object from all videos, and then co-segment the object only from these relevant frames. We present a spatiotemporal energy minimization formulation for simultaneous video object discovery and co-segmentation across multiple videos. Our formulation incorporates a spatiotemporal auto-context model, which is combined with appearance modeling for superpixel labeling. The superpixel-level labels are propagated
to the frame level through a multiple inst ance boosting algorithm with spatial rea-
soning (Spatial-MILBoosting), based on which frames containing the video object are identi■ed. Our method only needs to be bootstrapped with the frame-level
labels for a few video frames (e.g., usually 1 to 3) to indicate if they contain the
target objects or not. Experiments on three datasets validate the ef■cacy of our proposed method, which compares favorably with the state-of-the-art.
Keywords: video object discovery, video object co-segmentation, spatiotemporal auto-context model, Spatial-MILBoosting.
1
****

Supervoxel-Consistent
Foreground Propagation in Video

Suyog Dutt Jain and Kristen Grauman
University of Texas at Austin, USA

Abstract. A major challenge in video segmentation is that the foreground object may move quickly in the scene at the same time its ap-pearance and shape evolves over time. While pairwise potentials used
in graph-based algorithms help smooth labels between neighboring (super)pixels in space and time, they o■er only a myopic view of consis-tency and can be misled by inter-frame optical ■ow errors. We propose
ah i g h e ro r d e r supervoxel label consistency potential for semi-supervised foreground segmentation. Given an initial frame with manual annota-tion for the foreground object, our approach propagates the foreground
region through time, leveraging bottom-up supervoxels to guide its estimates towards long-range coherent regions. We validate our approachon three challenging datasets and achieve state-of-the-art results.
1
****

Clustering with Hypergraphs:
T h eC a s ef o rL a r g eH y p e r e d g e s

Pulak Purkait1,T a t - J u nC h i n1, Hanno Ackermann2, and David Suter1
1The University of Adelaide, Australia
2Leibniz Universit¨ at Hannover, Germany

Abstract. The extension of conventional clustering to hypergraph clustering, which involves higher order similarities instead of pairwise similarities, is increasingly gaining attention in computer vision. This is due
to the fact that many grouping problems require an a■nity measure thatmust involve a subset of data of size more than two, i.e., a hyperedge .A l -
most all previous works, however, have considered the smallest possible hyperedge size, due to a lack of study into the potential bene■ts of large hyperedges and e■ective algorithms to generate them. In this paper, weshow that large hyperedges are better from both theoretical and empir-
ical standpoints. We then propose a novel guided sampling strategy for large hyperedges, based on the concept of random cluster models .O u r method can generate pure large hyperedges that signi■cantly improve
grouping accuracy without exponential increases in sampling costs. In the important applications of face clustering and motion segmentation,our method

demonstrates substantially better accuracy and e■ciency.
Keywords: Hypergraph clustering, model ■tting, guided sampling.
1
****
Person Re-identi■cation by Video Ranking
Taiqing Wang1, Shaogang Gong2, Xiatian Zhu2, and Shengjin Wang1
1Dept. of Electronic Engineering, Tsinghua University, China
2School of EECS, Queen Mary University of London, UK
Abstract. Currentpersonre-identi■cation(re-id)methodstypicallyrely onsingle-frame imagery features, andignore space-timeinformation from image sequences. Single-frame (single-shot) visual appearance matchingis inherently limited for person re-id in public spaces due to visual ambiguity arising from non-overlapping camera views where viewpoint and lighting changes can cause signi■cant appearance variation. In this work,we present a novel model to automatically select the most discriminative video fragments from noisy image sequences of people where more reliable space-time features can be extracted, whilst simultaneously to learna video ranking function for person re-id. Also, we introduce a new image sequencere-iddataset(iLIDS-VID)basedonthei-LIDSMCT benchmark data. Using theiLIDS-VIDandPRID2011 sequence re-iddatasets,weextensivelyconductedcomparativeevaluationstodemonstratethead-vantagesoftheproposedmodelovercontemporarygaitrecognition,holistic image sequence matching and state-of-the-art single-shot/multi-shotbased reid methods.
1
****
Bayesian Nonparametric
Intrinsic Image Decomposition■
Jason Chang, Randi Cabezas, and John W. Fisher III
CSAIL, MIT, USA
Abstract. We present a generative, probabilistic model that decomposes an image into re■ectance and shading components. The proposed approach uses a Dirichlet process Gaussian mixture model where the mean parameters evolve jointly according to a Gaussian process. In contrast to prior methods, we eliminate the Retinex term and adopt more general smoothness assumptions for the shading image. Markov chain Monte Carlo sampling techniques are used for inference, yielding state-of-the-art results on the MIT Intrinsic Image Dataset.
Keywords: Intrinsic images, Dirichlet process, Gaussian process, MCMC.
1
****
Face Detection without Bells and Whistles
Markus Mathias1, Rodrigo Benenson2, Marco Pedersoli1, and Luc Van Gool1,3
1ESAT-PSI/VISICS, iMinds, KU Leuven, Belgium
2MPI Informatics, Saarbrücken, Germany
3D-ITET/CVL, ETH Zürich, Switzerland
Abstract. Face detection is a mature problem in computer vision. While diverse high performing face detectors have been proposed in the past, we present two surprising new top performance results. First, we show thata properly trained vanilla DPM reaches top performance, improving over commercial and research systems. Second, we show that a detector based onrigidtemplates - similar in structure to the Viola&Jones detector - can reach similar top performance on this task. Importantly, we discuss issues with existing evaluation benchmark and propose an improved procedure.
Fig. 1. Our proposed HeadHunter detector at the Oscars. Can you spot the one false positive, and one false negatives ? (hint: ■rst rows).
1
****

On Image Contours of Projective Shapes
Jean Ponce1,■and Martial Hebert2
1Department of Computer Science,
Ecole Normale Sup´ erieure, France
2Robotics Institute,
Carnegie-Mellon University, USA
Abstract. This paper revisits classical properties of the outlines of solid shap
es
bounded by smooth surfaces, and shows that they can be established in a purely
projective setting, without appealing to Euclidean measurements such as normals
or curvatures. In particular, we give new synthetic proofs of Koenderink’s famou
s
theorem on convexities and concavities of the image contour, and of the fact tha
t
the rim turns in the same direction as the viewpoint in the tangent plane at a
convex point, and in the opposite direction at a hyperbolic point. This suggests
that projective geometry should not be viewed merely as an analytical device
for linearizing calculations (its main role in structure from motion), but as th
e
proper framework for studying the relation between solid shape and its perspec-
tive projections. Unlike previous work in this area, the proposed approach does
not require an oriented setting, nor does it rely on any choice of coordinate sy
stem
or analytical considerations.
1
****
Programmable Automotive Headlights
Robert Tamburo1, Eriko Nurvitadhi2, Abhishek Chugh1, Mei Chen2,
Anthony Rowe1, Takeo Kanade1, and Srinivasa G. Narasimhan1
1Carnegie Mellon University, Pittsburgh, PA, USA
2Intel Labs, Pittsburgh, PA, USA
Abstract. The primary goal of an automotive headlight is to improve
safety in low light and poor weather conditions. But, despite decades ofinnovati
on on light sources, more than half of accidents occur at night
even with less tra■c on the road. Recent developments in adaptive light-
ing have addressed some limitations of standard headlights, however,they have li
mited ■exibility - switching between high and low beams,
turning o■ beams toward the opposing lane, or rotating the beam as
the vehicle turns - and are not designed for all driving environments.This paper
 introduces an ultra-low latency reactive visual system that
can sense, react, and adapt quickly to any environment while moving at
highway speeds. Our single hardware design can be programmed to per-form a varie
ty of tasks. Anti-glare high beams, improved driver visibility
during snowstorms, increased contrast of lanes, markings, and sidewalks,
and early visual warning of obstacles are demonstrated.
Keywords: Adaptive headlights, reactive visual system, computational
illumination.
1
****
ROCHADE: Robust Checkerboard Advanced
Detection for Camera Calibration
Simon Placht1,2, Peter F¨ ursattel1,2, Etienne Assoumou Mengue2,
Hannes Hofmann1, Christian Schaller1, Michael Balda1,
and Elli Angelopoulou2
1Metrilus GmbH, Erlangen, Germany
2Pattern Recognition Lab, University of Erlangen, Nuremberg, Germany
Abstract. We present a new checkerboard detection algorithm which is
able to detect checkerboards at extreme poses, or checkerboards whichare highly
distorted due to lens distortion even on low-resolution images.
On the detected pattern we apply a surface ■tting based subpixel re-

■nement speci■cally tailored for checkerboard X-junctions. Finally, weinvestigat
e how the accuracy of a checkerboard detector a■ects the over-
all calibration result in multi-camera setups. The proposed method is
evaluated on real images captured with di■erent camera models to showits wide ap
plicability. Quantitative comparisons to OpenCV’s checker-
board detector show that the proposed method detects up to 80% more
checkerboards and detects corner points more accurately, even understrong perspe
ctive distortion as often present in wide baseline stereo
setups.
Keywords: CheckerboardDetection,Saddle-BasedSubpixelRe■nement,
Multi Camera Calibration, Low Resolution Sensors, Lens Distortion.
1
****

Correcting for Duplicate Scene Structure
in Sparse 3D Reconstruction
Jared Heinly, Enrique Dunn, and Jan-Michael Frahm
The University of North Carolina at Chapel Hill, USA
Abstract. Structure from motion (SfM) is a common technique to re-
cover 3D geometry and camera poses from sets of images of a com-
mon scene. In many urban environments, however, there are symmetric,repetitive,
or duplicate structures that pose challenges for SfM pipelines.
The result of these ambiguous structures is incorrectly placed cameras
and points within the reconstruction. In this paper, we present a post-processin
g method that can not only detect these errors, but successfully
resolve them. Our novel approach proposes the strong and informative
measure of con■icting observations, and we demonstrate that it is robustto a lar
ge variety of scenes.
Keywords: Structurefrom motion, duplicatestructuredisambiguation.
1
****

Total Moving Face Reconstruction
Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M. Seitz
University of Washington, USA
Fig. 1.Given a YouTube video of a person’s face our method estimates high detail
geometry (full 3D ■ow and pose) in each video frame completely automatically
Abstract. We present an approach that takes a single video of a per-
son’s face and reconstructs a high detail 3D shape for each video frame.
Wetargetvideostakenunderuncontrolledanduncalibratedimagingcon-
ditions, such as youtube videos of celebrities. In the heart of this work isa ne
w dense 3D ■ow estimation method coupled with shape from shad-
ing. Unlike related works we do not assume availability of a blend shape
model, nor require the person to participate in a training/capturing pro-cess. I
nstead we leverage the large amounts of photos that are available
per individual in personal or internet photo collections. We show results
for a variety of video sequences that include various lighting conditions,head p
oses, and facial expressions.
Keywords: 3D reconstruction, faces, non-rigid reconstruction.
1
****

Automatic Single-View Calibration
and Recti■cation from Parallel Planar Curves
Eduardo R. Corral-Soto and James H. Elder
Centre for Vision Research, York University, Toronto, Canada
Abstract. Typical methods for camera calibration and image recti■ca-
tion from a single view assume the existence of straight parallel lines
from which vanishing points can be computed, or orthogonal structureknown to exi
st in the scene. However, there are practical situations where
these assumptions do not apply. Moreover, from a single family of paral-
lel lines on the ground plane there is insu■cient information to recovera comple
te recti■cation. Here we study a generalization of these meth-

ods to scenes known to contain parallel curves. Our method is based on establishing an association between pairs of corresponding points lyingon the image projection of these curves. We show how this method can be used to compute a least-squares estimate of the focal length and the camera pose from the tangent lines of the associated points, allowingcomplete rectification of the image. We evaluate the method on high-way and sports track imagery, and demonstrate its accuracy relative to a state-of-the-art vanishing point method.

Keywords: camera calibration, projective rectification, contour grouping, traffic surveillance.

1

****

On Sampling Focal Length Values to Solve the Absolute Pose Problem

Torsten Sattler1, Chris Sweeney2,■, and Marc Pollefeys1
1Department of Computer Science, ETH Z¨ urich, Z¨urich, Switzerland
2University of California Santa Barbara, Santa Barbara, USA

Abstract. Estimating the absolute pose of a camera relative to a 3D representation of a scene is a fundamental step in many geometric Computer Vision applications. When the camera is calibrated, the pose can be computed very efficiently. If the calibration is unknown, the problem becomes much harder, resulting in slower solvers or solvers requiring more samples and thus significantly longer run-times for RANSAC. In this paper, we challenge the notion that using minimal solvers is always optimal and propose to compute the pose for a camera with unknown focal length by randomly sampling a focal length value and using an efficient pose solver for the now calibrated camera. Our main contribution is a novel sampling scheme that enables us to guide the sampling process towards promising focal length values and avoids considering all possible values once a good pose is found. The resulting RANSAC variant is significantly faster than current state-of-the-art pose solvers, especially for low inlier ratios, while achieving a similar or better pose accuracy.

Keywords: RANSAC, n-point-pose (P nP), camera pose estimation.

1

****