TABLE OF CONTENTS

S1A - 01 - Optimal Feature Learning and Discriminative Framework for Polarimetric Thermal to Visible Face Recognition ................................................................. 1
Benjamin S. Riggan, Nathaniel J. Short, Shuowen Hu

S1A - 02 - Discovery of Facial Motions Using Deep Machine Perception ........................................ 8
Afshan Ghasemi, Simon Denman, Sridha Sridharan

S1A - 03 - Customized Expression Recognition for Performance-Driven Cutout Character Animation .............................................................. 15
Xiang Yu, Jianchao Yang, Linjie Luo, Wilmot Li, Jonathan Brandt, Dimitris N. Metaxas

S1A - 04 - Going Deeper in Facial Expression Recognition Using Deep Neural Networks .............. 24
Ali Mollahosseini, David Chan, Mohammad H. Mahoor

S1A - 05 - Discriminative FaceTopics for Face Recognition via Latent Dirichlet Allocation ............ 34
Tejas Indulal Dhonecha, Praneet Sharma, Richa Singh, Mayank Vatsa

S1A - 06 - Can We Still Avoid Automatic Face Detection? ........................................................ 42
Michael J. Wilber, Vitaly Shmatikov, Serge Belongie

S1A - 07 - OpenFace: An Open Source Facial Behavior Analysis Toolkit .................................... 51
Tadas Baltrušaitis, Peter Robinson, Louis-Philippe Morency

S1A - 08 - Correlation Filter Cascade for Facial Landmark Localization .................................... 61
Hamid Kiani Galoogahi, Terence Sun

S1A - 09 - Face Recognition Using Deep Multi-Pose Representations ........................................ 69
Wael Abdalmageed, Yue Wu, Stephen Rovdis, Tal Hassner, Iacopo Masì, Jongmoo Choi, Jatuporn Leksat, Jungyeon Kim, Prem Natarajan, Ryan Nevatia, Gerard Medioni

S1A - 10 - Effect of Illicit Drug Abuse on Face Recognition .................................................... 78
Daksha Yadav, Naman Kohli, Prateekshit Pandey, Richa Singh, Mayank Vatsa, Afeel Noore

S1A - 11 - Unconstrained Face Verification Using Deep CNN Features ................................... 85
Jun-Cheng Chen, Vishal M. Patel, Rama Chellappa

S1A - 12 - Frontal to Profile Face Verification in the Wild .......................................................... 94
Sounyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M. Patel, Rama Chellappa, David W. Jacobs

S1A - 13 - Capturing Facial Videos With Kinect 2.0: A Multithreaded Open Source Tool and Database .............................................................. 103
Daniel Merget, Tobias Eckl, Martin Schwörer, Philipp Tiefenbacher, Gerhard Rigoll

S1A - 14 - Naming TV Characters by Watching and Analyzing Dialogs .................................... 108
Monica-Laura Haurilet, Makarand Tapaswi, Ziad Al-Halah, Rainer Stiefelhagen

S1A - 15 - Direct Face Detection and Video Reconstruction From Event Cameras .................... 117
Yoshitaka Miyatani, Souptik Barua, Ashok Veeraraghavan

SIB - 01 - Object Detection in 20 Questions ................................................................................. 126
Xi Chen, He He, Larry S. Davis

SIB - 02 - Pose Tracking by Efficiently Exploiting Global Features ........................................ 135
Ratnesh Kumar, Dhruv Batra

SIB - 03 - Exploring Bounding Box Context for Multi-Object Tracker Fusion ......................... 144
Stefan Breuers, Shishan Yang, Markus Mathias, Bastian Leibe

SIB - 04 - OCPAD - Occluded Checkerboard Pattern Detector ............................................ 152
Peter Fürsattel, Sergiu Dotenco, Simon Placht, Michael Balda, Andreas Maier, Christian Riess

SIB - 05 - Leveraging Single for Multi-Target Tracking Using a Novel Trajectory Overlap Affinity Measure ................................................................................. 161
Santiago Munen, Rada Timofte, Dengxin Dai, Luc Van Gool

SIB - 06 - Procursatean Decomposition for Orthogonal Cascade Detection ................................ 170
Kun Duan, Wei Wang, Ting Yu

SIB - 07 - Region Graph Based Method for Multi-Object Detection and Tracking Using Depth Cameras ...................................................................................... 179
Sachin Mehta, Balakrishnan Prabhakaran

SIB - 08 - Online Tracking Using Saliency ................................................................................ 187
Mohammed A. Yousefhussien, N. Andrew Browning, Christopher Kanan

SIB - 09 - Robust Visual Tracking Using Template Anchors .................................................. 197
Luka Cehovin, Aleš Leonardis, Matej Kristan

SIB - 10 - A Structured Approach to Predicting Image Enhancement Parameters .................... 205
Parag Shridhar Chandakkar, Baoxin Li

SIB - 11 - Active Contours for Selective Object Segmentation .................................................. 214
Jozef Molnar, Adam Istvan Szucs, Csaba Molnar, Peter Horvath
<table>
<thead>
<tr>
<th>Conference</th>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1B - 12</td>
<td>A Survey on Moving Object Detection for Wide Area Motion Imagery</td>
<td>Lars Wilko Sommer, Michael Teutsch, Tobias Schuchert, Jürgen Beyerer</td>
</tr>
<tr>
<td>S1B - 13</td>
<td>Dynamic Belief Fusion for Object Detection</td>
<td>Hyungtae Lee, Heesung Kwon, Ryan M. Robinson, William D. Nothwang, Amar M. Marathe</td>
</tr>
<tr>
<td>S1B - 14</td>
<td>Text Detection in Stores Using a Repetition Prior</td>
<td>Bo Xiong, Kristen Grauman</td>
</tr>
<tr>
<td>S1B - 15</td>
<td>Cutting Through the Clutter: Task-Relevant Features for Image Matching</td>
<td>Rohit Girdhar, David F. Fouhey, Kris M. Kitani, Abhinav Gupta, Martial Hebert</td>
</tr>
<tr>
<td>S1C - 01</td>
<td>A Multi-Modal Feature Fusion Framework for Kinect-Based Facial Expression Recognition Using Dual Kernel Discriminant Analysis (OKDA)</td>
<td>Sherin Aly, A. Lynn Abbot, Marwan Torki</td>
</tr>
<tr>
<td>S1C - 02</td>
<td>Learning Patch-Dependent Kernel Forest for Person Re-Identification</td>
<td>Wei Wang, Ali Taalimi, Kun Duan, Rui Guo, Hairong Qi</td>
</tr>
<tr>
<td>S1C - 03</td>
<td>Hide and Seek: Uncovering Facial Occlusion With Variable-Threshold Robust PCA</td>
<td>Wee Kheng Leow, Guodong Li, Juan Lai, Terence Sim, Vaishali Sharma</td>
</tr>
<tr>
<td>S1C - 04</td>
<td>Score Reliability Based Weighting Technique for Score-Level Fusion in Multi-Biometric Systems</td>
<td>Waziha Kabir, M. Omair Ahmad, M. N. S. Swamy</td>
</tr>
<tr>
<td>S1C - 05</td>
<td>HeHOP: Highly Efficient Head Orientation and Position Estimation</td>
<td>Anke Schwarz, Zhuang Lin, Rainer Stiefelhagen</td>
</tr>
<tr>
<td>S1C - 06</td>
<td>A Revisit to Human Action Recognition From Depth Sequences: Guided SVM-Sampling for Joint Selection</td>
<td>Michel Antunes, Djamila Aouada, Björn Ottersten</td>
</tr>
<tr>
<td>S1C - 07</td>
<td>Abstraction Hierarchy and Self Annotation Update for Fine Grained Activity Recognition</td>
<td>Song Cao, Kan Chen, Ram Nevatia</td>
</tr>
<tr>
<td>S1C - 08</td>
<td>Activity Recognition and Prediction With Pose Based Discriminative Patch Model</td>
<td>Song Cao, Kan Chen, Ram Nevatia</td>
</tr>
<tr>
<td>S1C - 09</td>
<td>Deep Tree-Structured Face: A Unified Representation for Multi-Task Facial Biometrics</td>
<td>Rui Guo, Liu Liu, Wei Wang, Ali Taalimi, Chi Zhang, Hairong Qi</td>
</tr>
<tr>
<td>S1C - 10</td>
<td>Recognition of Ongoing Complex Activities by Sequence Prediction Over a Hierarchical Label Space</td>
<td>Wenbin Li, Mario Fritz</td>
</tr>
<tr>
<td>S1C - 11</td>
<td>Linear-Time Online Action Detection From 3D Skeletal Data Using Bags of Gesturelets</td>
<td>Moustafa Mesrhy, Mohamed E. Hussein, Marwan Torki</td>
</tr>
<tr>
<td>S1C - 12</td>
<td>Efficient Video-Based Retrieval of Human Motion With Flexible Alignment</td>
<td>Ankur Gupta, John He, Julieta Martinez, James J. Little, Robert J. Woodham</td>
</tr>
<tr>
<td>S1C - 13</td>
<td>Combining Multiple Sources of Knowledge in Deep CNNs for Action Recognition</td>
<td>Eunhyung Park, Xufeng Han, Tamara L. Berg, Alexander C. Berg</td>
</tr>
<tr>
<td>S1C - 14</td>
<td>Person Re-Identification Using Deformable Patch Metric Learning</td>
<td>Slawomir Bak, Peter Carr</td>
</tr>
<tr>
<td>S1C - 15</td>
<td>Support Vector Machines With Time Series Distance Kernels for Action Classification</td>
<td>Mohammad Ali Bagheri, Qingguang Guo, Sergio Escalera</td>
</tr>
<tr>
<td>S1C - 16</td>
<td>Face Fiducial Detection by Consensus of Exemplars</td>
<td>Mallikarjun B R, Viseesh Chari, C. V. Jawahar, Akshay Asthana</td>
</tr>
<tr>
<td>S1C - 17</td>
<td>One-to-Many Face Recognition With Bilinear CNNs</td>
<td>Aruni Roychowdhury, Tsung-Yu Lin, Subhranshu Maji, Erik Learned-Miller</td>
</tr>
<tr>
<td>S1D - 01</td>
<td>Tooth Guard: A Vision System for Detecting Missing Tooth in Rope Mine Shovel</td>
<td>Ser Nam Lim, Joao Soares, Ning Zhou</td>
</tr>
<tr>
<td>S1D - 02</td>
<td>Multiscale Fully Convolutional Network With Application to Industrial Inspection</td>
<td>Xiao Bian, Ser Nam Lim, Ning Zhou</td>
</tr>
<tr>
<td>S1D - 04</td>
<td>Texture Classification for Rail Surface Condition Evaluation</td>
<td>Ke Ma, Tomáš F. Yago Vicente, Dimitris Samaras, Mike Petrucci, Daniel L. Magnus</td>
</tr>
<tr>
<td>S1D - 05</td>
<td>Visual Recognition of Paper Analytical Device Images for Detection of Falsified Pharmaceuticals</td>
<td>Sandipan Banerjee, James Sweet, Christopher Sweet, Marya Lieberman</td>
</tr>
<tr>
<td>S1D - 06</td>
<td>Deep Learning Architectures For Domain Adaptation in HOV/HOT Lane Enforcement</td>
<td>Safwan Whah, Bilee Xu, Orhan Bulan, Jayanti Kumar, Peter Paul</td>
</tr>
</tbody>
</table>
S1D - 08 - Detection of Cracks in Nuclear Power Plant Using Spatial-Temporal Grouping of Local Patches ................................................................. 462
S1D - 09 - An Elastic Functional Data Analysis Framework for Preoperative Evaluation of Patients With Rheumatoid Arthritis ................................................. 469
   Chaofan Sun, Sebastian Kurtek, Anuj Srivastava, Noé Borges
S1D - 10 - A Deep Convolutional Neural Network Trained on Representative Samples for Circulating Tumor Cell Detection ......................................................... 477
S1D - 11 - A New Computer Vision-Based System to Help Clinicians Objectively Assess Visual Pursuit With the Moving Mirror Stimulus for the Diagnosis of Minimally Conscious State ................................................................. 483
   Thomas Hoeyns, Sarah Wannez, Thomas Langohr, Jérémy Wertz, Steven Laureys, Jacques G. Verly
S1D - 12 - Weighted Atlas Auto-Context With Application to Multiple Organ Segmentation ......................................................... 491
   Telmo Amaral, Ilias Kyriazakis, Stephen J. McKenna, Thomas Plötz
S1D - 13 - Accurate 3D Bone Segmentation in Challenging CT Images: Bottom-Up Parsing and Contextualized Optimization ......................................................... 500
   Le Lu, Dijia Wu, Nathan Lay, David Liu, Isabella Nogues, Ronald M. Summers
S1D - 14 - Real-Time Road Traffic Density Estimation Using Block Variance ................................................................................................................................. 510
   Kratika Garg, Siew Kei Lam, Thambipillai Srikanthan, Vedika Agarwal
S1D - 15 - Monocular Obstacle Avoidance for Blind People Using Probabilistic Focus of Expansion Estimation ................................................................................................................................. 519
   Sebastian Stabinger, Antonio Rodriguez-Sánchez, Justus Piater
S1D - 16 - Atomic Scenes for Scalable Traffic Scene Recognition in Monocular Videos ................................................................................................................................. 528
   Chao-Yeh Chen, Wungun Choi, Mannohan Chandraker
S1D - 17 - Eye-CU: Sleep Pose Classification for Healthcare Using Multimodal Multiview Data ................................................................................................................................. 537
   Carlos Torres, Victor Fragozli, Scott D. Hammond, Jeffrey C. Fried, Bangalore S. Manjunath
S2A - 01 - Fashion Apparel Detection: The Role of Deep Convolutional Neural Network and Pose-Dependent Priors ................................................................................................................................. 546
   Kota Hara, Vignesh Jagadeesh, Robinson Piramuthu
S2A - 02 - Fixation Prediction With a Combined Model of Bottom-Up Saliency and Vanishing Point ................................................................................................................................. 555
   Mengyang Feng, Ali Borji, Huchuan Lu
S2A - 03 - Is Image Super-Resolution Helpful for Other Vision Tasks? ................................................................................................................................. 562
   Dengxin Dai, Yujian Wang, Yuhua Chen, Luc Van Gool
S2A - 04 - Energy-Efficient ConvNets Through Approximate Computing ................................................................................................................................. 571
   Bert Moons, Bert De Brabandere, Luc Van Gool, Marian Verhelst
S2A - 05 - Fine-Grained Sketch-Based Image Retrieval: The Role of Part-Aware Attributes ................................................................................................................................. 579
   Ke Li, Kaiyue Pang, Yi-Zhe Song, Timothy Hospedales, Honggang Zhang, Yichuan Hu
S2A - 06 - Toward Correlating and Solving Abstract Tasks Using Convolutional Neural Networks ................................................................................................................................. 588
   Kuan-Chuan Peng, Tsz Chun Chan
S2A - 07 - Assessing Tracking Performance in Complex Scenarios Using Mean Time Between Failures ................................................................................................................................. 597
   Peter Carr, Robert T. Collins
S2A - 08 - A Crowdsourced Approach to Student Engagement Recognition in e-Learning Environments ................................................................................................................................. 607
   Aditya Kamath, Aradhyaa Biswas, Vineeth Balasubramanian
S2A - 09 - Color Multi-Fusion Fisher Vector Feature for Fine Art Painting Categorization and Influence Analysis ................................................................................................................................. 616
   Ajit Pathanputhussery, Qingfeng Liu, Chengjun Liu
S2A - 10 - Constructing Image Mosaics Using Focus Based Depth Analysis ................................................................................................................................. 625
   Mohamed A. Helala, Faisal Z. Qureshi
S2A - 11 - Image Set Classification by Symmetric Positive Semi-Definite Matrices ................................................................................................................................. 634
   Masoud Faraki, Mehrshad T. Harandi, Fatih Porikli
S2A - 12 - Precise Deterministic Change Detection for Smooth Surfaces ................................................................................................................................. 642
   Simon Stent, Riccardo Gherardi, Björn Stenger, Roberto Cipolla
S2A - 13 - Unsupervised Saliency Estimation Based on Robust Hypotheses ................................................................................................................................. 651
   Fei Xu, Xin Xian, H. D. Cheng, Jianrui Ding, Yingtao Zhang
S2A - 14 - On the Importance of Normalisation Layers in Deep Learning With Piecewise Linear Activation Units ................................................................................................................................. 657
   Zhibin Liao, Gustavo Carneiro
S2A - 15 - Deep Learning the Dynamic Appearance and Shape of Facial Action Units ................................................................................................................................. 665
   Shashank Jaiswal, Michel Valstar
S2B - 01 - IPDC: Iterative Part-Based Dense Correspondence Between Point Clouds ................................................................................................................................. 673
   Rongqi Qiu, Ulrich Neumann
S2B - 02 - Categorizing Cubes: Revisiting Pose Normalization ................................................................. 682
Mohsen Hejrati, Deva Ramanan

S2B - 03 - Unsupervised Categorical Shape Reconstruction Through Manifolds .......................................... 691
Kent Fujiwara, Minara Mori, Kunio Kashino

S2B - 04 - Mobile Phone and Cloud - A Dream Team for 3D Reconstruction .................................................. 699
Alex Locher, Michal Ferdioch, Hayko Riemenschnieder, Luc Van Gool

S2B - 05 - Online Inspection of 3D Parts via a Locally Overlapping Hierarchical Camera Network ......................... 707
Tolga Birdal, Emrah Bala, Tolga Eren, Slobodan Ilic

S2B - 06 - Omnidirectional Image Capture on Mobile Devices for Fast Automatic Generation of 2.5D Indoor Maps .................................................................................................................................................. 717
Giovanni Pintore, Valeria Garro, Fabio Gianvelli, Marco Agus, Enrico Gobbetti

S2B - 07 - Where Is That Pixel in the Oblique-View Video? ................................................................................. 726
Yin Li, Teck Kin Ng

S2B - 08 - Mono Camera Multi-View Diminished Reality ....................................................................................... 734
Philipp Tiefenbacher, Michael Sirch, Gerhard Rigoll

S2B - 09 - Mosaicing Scenes With a Quadcopter ................................................................................................ 742
Meghshyam G. Prasad, Sharath Chandran, Michael S. Brown

S2B - 10 - High Accuracy Model-Based Object Pose Estimation for Autonomous Recharging
Applications .......................................................................................................................................................... 750
Hanno Jaspers, Georg R. Mueller, Hans-Joachim Waensche

S2B - 11 - CoRBS: Comprehensive RGB-D Benchmark for SLAM Using Kinect v2 ............................................. 757
Oliver Wasenmüller, Marcel Meyer, Didier Stricker

S2B - 12 - Sky Segmentation in the Wild: An Empirical Study .................................................................................. 764
Radu P. Mihail, Scott Workman, Zach Bessinger, Nathan Jacobs

S2B - 13 - Semantic Segmentation of Modular Furniture ...................................................................................... 770
Tobias Pohlen, Ishrat Badami, Markus Mathiass, Bastian Leibe

S2B - 14 - Simultaneous Semantic Segmentation of a Set of Partially Labeled Images .............................................. 779
Qiongjie Tian, Baoxin Li

S2C - 01 - Direct 3D Pose Estimation of a Planar Target .......................................................................................... 788
Hung-Yu Tseng, Po-Chen Wu, Ming-Hsuan Yang, Shao-Yi Chien

S2C - 02 - Forget the Checkerboard: Practical Self-Calibration Using a Planar Scene ............................................... 797
Daniel Herrera Castro, Juho Kannala, Janne Heikkilä

S2C - 03 - 6DOF Point Cloud Alignment Using Geometric Algebra-Based Adaptive Filtering .............................. 806
Anas Al-Nuaimi, Wilder B. Lopes, Eckehard Steinbach, Cassio G. Lopes

S2C - 04 - Unifying Diffuse and Specular Reflections for the Photometric Stereo Problem .................................. 815
Roberto Mecca, Yvain Quéau

S2C - 05 - Underwater 3D Capture Using a Low-Cost Commercial Depth Camera ................................................... 824
Sundara Tejasvi Digumarti, Aparna Taneja, Amber Thomas, Gaurav Chaurasia, Roland Siegwart, Paul Beardsley

S2C - 06 - Resolution Enhancement in Single Depth Map and Aligned Image .......................................................... 833
Yang Xian, Yingli Tian

S2C - 07 - Geometric Calibration for Mobile, Stereo, Autofocus Cameras ............................................................... 842
Stephen DiVerdi, Jonathan T. Barron

S2C - 08 - Lifting GIS Maps Into Strong Geometric Context for Scene Understanding ........................................... 850
Raül Díaz, Minhvuong Lee, Jochen Schubert, Charless C. Fowlkes

S2C - 09 - Heat Propagation Contours for 3D Non-Rigid Shape Analysis ............................................................... 859
Xupeng Wang, Ferdous Sohel, Mohammed Bennamoun, Hang Lei

S2C - 10 - Dealing With Small Data and Training Blind Spots in the Manhattan World ........................................... 866
Muhammad Wajahat Hussain, Javier Civera, Luis Montano, Martial Hebert

S2C - 11 - Automatic 3D Reconstruction of Manifold Meshes via Delaunay Triangulation and Mesh Sweeping ................................................................. 875
Andrea Romanoni, Amael Delaunoy, Marc Pollefeys, Matteo Matteucci

S2C - 12 - Half Hypersphere Confinement for Piecewise Linear Regression ........................................................... 883
Eduardo Pérez-Pellittero, Jordi Salvador, Javier Ruiz-Hidalgo, Bodo Rosenbahn

S2C - 13 - 3D Shape Retrieval Using a Single Depth Image From Low-Cost Sensors ............................................... 892
Jie Feng, Yan Wang, Shih-Fu Chang

S2C - 14 - Architectural Decomposition for 3D Landmark Building Understanding ................................................ 901
Nikolay Kohshev, Hayko Riemenschnieder, András Bódis-Szomorú, Luc Van Gool

S2C - 15 - Voting-Based 3D Object Cuboid Detection Robust to Partial Occlusion From RGB-D Images ....................... 911
Sangdoo Yun, Hawook Jeong, Soowan Kim, Jin Young Choi
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3A - 10 - Crowd Density Estimation Based on Rich Features and Random Projection Forest</td>
<td>1143</td>
</tr>
<tr>
<td>Bolei Xu, Guoping Qiu</td>
<td></td>
</tr>
<tr>
<td>S3A - 11 - People Detection in Crowded Scenes by Context-Driven Label Propagation</td>
<td>1151</td>
</tr>
<tr>
<td>Jingjing Liu, Qianfu Fan, Sharath Pankanti, Dimitris N. Metaxas</td>
<td></td>
</tr>
<tr>
<td>S3A - 12 - Efficient Unsupervised Abnormal Crowd Activity Detection Based on a Spatiotemporal Salience Detector</td>
<td>1160</td>
</tr>
<tr>
<td>Yilin Wang, Qiang Zhang, Baosun Li</td>
<td></td>
</tr>
<tr>
<td>S3A - 13 - Vision-Based Counting of Pedestrians and Cyclists</td>
<td>1169</td>
</tr>
<tr>
<td>Mehmet Kemal Kocamaz, Jian Gong, Bernardo R. Pires</td>
<td></td>
</tr>
<tr>
<td>S3A - 14 - Static Action Recognition by Efficient Greedy Inference</td>
<td>1177</td>
</tr>
<tr>
<td>Shaukat Abidi, Massimo Picardi, Mary-Anne Williams</td>
<td></td>
</tr>
<tr>
<td>S3A - 15 - An Analysis of 1-to-First Matching in Iris Recognition</td>
<td>1185</td>
</tr>
<tr>
<td>Andrey Kueblakeng, Kevin W. Boyer</td>
<td></td>
</tr>
<tr>
<td>S3B - 01 - Self-Taught Object Localization With Deep Networks</td>
<td>1193</td>
</tr>
<tr>
<td>Loris Bazzani, Alessandro Bergamo, Dragomir Anguelov, Lorenzo Torresani</td>
<td></td>
</tr>
<tr>
<td>S3B - 02 - Learning a Structured Dictionary for Video-Based Face Recognition</td>
<td>1202</td>
</tr>
<tr>
<td>Hongyu Xu, Jingjing Zheng, Azadeh Alavi, Rama Chellappa</td>
<td></td>
</tr>
<tr>
<td>S3B - 03 - Kernel Auto-Encoder for Semi-Supervised Hashing</td>
<td>1211</td>
</tr>
<tr>
<td>Behnam Gholami, Abolfazl Hajisami</td>
<td></td>
</tr>
<tr>
<td>S3B - 04 - Higher-Order Class-Specific Priors for Semantic Segmentation of 3D Outdoor Scenes</td>
<td>1219</td>
</tr>
<tr>
<td>Bingsheng Tang, Yu Zhou, Yao Yu, Sidan Du</td>
<td></td>
</tr>
<tr>
<td>S3B - 05 - Multi-View Dynamic Texture Learning</td>
<td>1228</td>
</tr>
<tr>
<td>Thanh Minh Nguyen, Q. M. Jonathan Wu</td>
<td></td>
</tr>
<tr>
<td>S3B - 06 - Automatic and Quantitative Evaluation of Attribute Discovery Methods</td>
<td>1237</td>
</tr>
<tr>
<td>Liangchen Liu, Arnold Willem, Shaokang Chen, Brian C. Lovell</td>
<td></td>
</tr>
<tr>
<td>S3B - 07 - Deep Recursive and Hierarchical Conditional Random Fields for Human Action Recognition</td>
<td>1246</td>
</tr>
<tr>
<td>Tianliang Liu, Xincheng Wang, Xiubin Dai, Jiebo Luo</td>
<td></td>
</tr>
<tr>
<td>S3B - 08 - A New Image Transformation Method Using Linear Discriminant Analysis (LDA) and Kernel Mapping (k-vector)</td>
<td>1255</td>
</tr>
<tr>
<td>Sephehr Damavandi, Nejadmonfared, Vijay Varadharajan</td>
<td></td>
</tr>
<tr>
<td>S3B - 09 - Think Big, Solve Small: Scaling Up Robust PCA With Coupled Dictionaries</td>
<td>1262</td>
</tr>
<tr>
<td>Jian Lai, Wei Kheng Leow, Terence Sim, Vaishali Sharma</td>
<td></td>
</tr>
<tr>
<td>S3B - 10 - Discriminative Training of CRF Models With Probably Submodular Constraints</td>
<td>1270</td>
</tr>
<tr>
<td>Wojciech Zaremba, Matthew B. Blasichko</td>
<td></td>
</tr>
<tr>
<td>S3B - 11 - Efficient Transductive Semantic Segmentation</td>
<td>1277</td>
</tr>
<tr>
<td>Jose M. Alvarez, Mathieu Salzmann, Nick Barnes</td>
<td></td>
</tr>
<tr>
<td>S3B - 12 - Unsupervised Network Pretraining via Encoding Human Design</td>
<td>1286</td>
</tr>
<tr>
<td>Ming-Yu Liu, Arun Malliya, Oncel Tuzel, Xi Chen</td>
<td></td>
</tr>
<tr>
<td>S3B - 13 - Coupled Depth Learning</td>
<td>1295</td>
</tr>
<tr>
<td>Mohammad Haris Baig, Lorenzo Torresani</td>
<td></td>
</tr>
<tr>
<td>S3B - 14 - Fine-Grained Classification via Mixture of Deep Convolutional Neural Networks</td>
<td>1305</td>
</tr>
<tr>
<td>Zong Yuan Ge, Alex Bewley, Christopher McCool, Peter Corke, Ben Upercroft, Conrad Sanderson</td>
<td></td>
</tr>
<tr>
<td>S3B - 15 - An End-To-End Generative Framework for Video Segmentation and Recognition</td>
<td>1311</td>
</tr>
<tr>
<td>Hilde Kuehne, Juergen Gall, Thomas Serre</td>
<td></td>
</tr>
<tr>
<td>S3C - 01 - Detecting Temporally Consistent Objects in Videos Through Objects Class Label Propagation</td>
<td>1319</td>
</tr>
<tr>
<td>Subarna Tripathi, Serge Belongie, Younghwa Hwang, Truong Nguyen</td>
<td></td>
</tr>
<tr>
<td>S3C - 02 - A Mid-Level Representation of Visual Structures for Video Compression</td>
<td>1328</td>
</tr>
<tr>
<td>Georgios Georgiadis, Stefano Soatto</td>
<td></td>
</tr>
<tr>
<td>S3C - 03 - Video Summarization for Remote Invigilation of Online Exams, Melissa Cote</td>
<td>1336</td>
</tr>
<tr>
<td>Frédéric Jean, Alexandra Brunzan Albu, David Capson</td>
<td></td>
</tr>
<tr>
<td>S3C - 04 - Automatic Video Editing for Sensor-Rich Videos</td>
<td>1345</td>
</tr>
<tr>
<td>Wesley Taylor, Faisal Z. Qureshi</td>
<td></td>
</tr>
<tr>
<td>S3C - 05 - Tag-Based Video Retrieval by Embedding Semantic Content in a Continuous Word Space</td>
<td>1354</td>
</tr>
<tr>
<td>Arnav Agharwal, Rama Kovvari, Ram Nevatia, Cees G. M. Snoek</td>
<td></td>
</tr>
<tr>
<td>S3C - 06 - Discovering Picturesque Highlights From Egocentric Vacation Videos</td>
<td>1362</td>
</tr>
<tr>
<td>Vinay Bettadapura, Daniel Castro Chin, Ifran Essa</td>
<td></td>
</tr>
<tr>
<td>S3C - 07 - Compact CNN for Indexing Egocentric Videos</td>
<td>1371</td>
</tr>
<tr>
<td>Yair Poleg, Ariel Ephrat, Shmuel Poleg, Chetan Arora</td>
<td></td>
</tr>
<tr>
<td>S3C - 08 - Transition Hough Forest for Trajectory-Based Action Recognition</td>
<td>1380</td>
</tr>
<tr>
<td>Guillermo Garcia-Hernando, Hyung Jin Chang, Ismael Serrano, Oscar Deniz, Taekyun Kim</td>
<td></td>
</tr>
</tbody>
</table>
Transition Hough Forest for Trajectory-based Action Recognition

Guillermo Garcia-Hernando† Hyung Jin Chang†
†Imperial College London

Ismael Serrano‡ Oscar Deniz‡ Tae-Kyun Kim‡
‡University of Castilla-La Mancha

{ggarciah, hj.chang, tk.kim}@imperial.ac.uk
{ismael.serrano, oscar.deniz}@uclm.es

Abstract

In this paper, we propose a new discriminative framework based on Hough forests that enables us to efficiently recognize and localize sequential data in the form of spatio-temporal trajectories. Contrary to traditional decision forest-based methods where predictions are made independently of its output temporal context, we introduce the concept of "transition", which enforces the temporal coherence of estimations and further enhances the discrimination between action classes. We start applying our proposed framework to the problem of recognizing and localizing fingertip written trajectories in mid-air using an egocentric camera. To this purpose, we present a new challenging dataset that allows us to evaluate and compare our method with previous approaches. Finally, we apply our framework to general human action recognition using local spatio-temporal trajectories obtaining comparable to state-of-the-art performance on a public benchmark.

1. Introduction

A human action can be seen as an ensemble of spatio-temporal trajectories that describe human motion. Trajectories can have different levels of abstraction (see Figure 1): from low-level trajectories describing local motion of parts of a human body to high-level trajectories such as handwritten characters that have a meaning by themselves. However, different kind of trajectories have a common and important property: they are time-structured patterns.

With the recent introduction of wearable cameras a new chapter in computer vision called egocentric vision has emerged where the user is the center of the action. A distinctive characteristic of this new paradigm relative to the classic third-person vision is that hands are very present in the scene [6, 14]. As these wearable sensors lack a keyboard, an interesting way to communicate with them would involve using our hands. A natural way of doing this would consist in writing with our fingertip in front of the camera. We can think of the fingertip motion as a spatio-temporal trajectory in mid-air which can represent, for instance, a handwritten character in the English alphabet (see Figure 2). This could lead to many different new applications in the domains of human-computer interaction or augmented reality. If we are able to recognize the fingertip written trajectories in mid-air we can use them as text input for the device. If we are able to not only recognizing them but localize them, we could, for instance, write notes in a virtual blackboard. Note that these two applications need real-time performance. Motivated by the aforementioned challenges, we address the problem of recognizing and localizing fingertip written characters in mid-air. We propose a new framework based on Hough forests [8] that we evaluate presenting the first public dataset of fingertip written characters in mid-air recorded with an egocentric sensor.

Decision forests-based methods have been very successful and popular in many computer vision tasks because their efficiency both in training and testing, their inherently multi-class handling ability and their capacity to handle overfitting. Their efficiency in prediction comes with the cost of assuming independence in the output variables, which is not always the case for sequential data. In the interest of enforcing temporal coherence in our forest framework, we introduce the concept of transition. We define a transition as the probability of observing the current output of the forest taking into account previous observations. Thus, our current prediction will consider what has been previously observed. Based on Hough forests, our method inherits the benefits of a decision forest model while enforc-
ing the temporal coherence of predictions. Finally, we show that our framework formulation is general enough to deal with different types of sequential data proving its suitability for general human action recognition.

In summary, our main contributions are three-fold:

- A new general framework based on Hough forests which can simultaneously recognize and localize spatio-temporal trajectories.
- Introduction of temporal context in a decision forest-based classifier in the form of transitions.
- The first public dataset containing fingertip written characters in mid-air in egocentric viewpoint.

2. Related Work

Decision forests for structured prediction: There has been some preliminary work using decision forest methods for spatio-temporal data modeling in diverse applications. Spatio-temporal relational probability trees [18] were proposed for weather process understanding, however the nature of their data is very different from ours. More related to our work, some approaches used tree-based methods for human action recognition [20, 35, 8, 10]. [20] proposed simultaneous action recognition and localization using local motion-appearance features method and clustering trees. [35] also used codebooks for building spatio-temporal histograms and matching them using histogram intersections and a SVM classifier. On the other hand, decision forests methods have been also used for directly mapping spatio-temporal features to space-time location and class label. In [8], dense spatio-temporal cuboids were extracted and each of them voted independently for a hypothesis in space, time and class in the Hough space. [10] proposed a spatio-temporal forest for detecting the action of finger clicking from an egocentric viewpoint, but they only considered one simple action with not much temporal structure. These approaches rely on the premise that spatio-temporal structure is adequately embedded in the feature level. In practice, noisy and incoherent labels are observed mainly caused by the output independence assumption. This is a general problem in structured prediction using decision forests and some authors have proposed solutions in other computer vision areas such as semantic image segmentation [21, 29, 30, 23, 12]. [21, 29] exploited the hierarchical nature of the trees in order to cluster similar samples and extract context information. [30, 23] used graphical models in top of decision forest predictions, while [12] proposed directly modelling the context within the forest in order to have smooth pixel-wise labellings. [3] introduced temporal context in a decision forest framework by warping map confidences using optical flow for body pose estimation.

Figure 3. Dense cuboid patches in Hough forest (orange) and trajectory patches in Transition Hough forest (blue) in two different scenes of punching and kicking from UT-interaction dataset [27].

Fingertip writing in mid-air: Recognizing fingertip written trajectories in mid-air has been previously explored in the last decades highly depending on the available technology and mainly from a third person viewpoint [24, 1, 28, 7, 31]. From an egocentric point of view the problem remains quite unexplored; however there are some early approaches related to our application [16, 11, 9]. [16] with the help of a wearable computer recognized fingertip trajectories using a spline-based matching algorithm. [11] and [9] recorded fingertip writing gestures with a webcam pointing a desktop and used DTW-based classifiers, but no realtime performance was achieved and localization was not performed.

Trajectories for human action recognition: Trajectory-based methods [17, 19, 32, 33, 34] for human action recognition have been very popular in the last years mainly to its good results in a variety of datasets. [19] extracted trajectories using an interesting point detector and proposed a graphical model to model the velocities of those trajectories. [17] extracted trajectories using a KLT tracker and clustered them in a bag of words fashion [22]. [32] densely sampled the and tracked trajectories extracting local descriptors such as HOGHOF [13] and MBH [5] along them. Densely sampling trajectories leads to the problem of capturing non meaningful trajectories; a problem that can be attenuated modelling the camera motion [33] or automatically learning the feature representation [34]. After extracting trajectory-sampled features, [32, 33, 34] use a bag of words model [22] losing important structural information.

3. Overview of the method

3.1. Hough forests for spatio-temporal trajectories

A spatio-temporal trajectory is a set of time-ordered space tracked points \( P_t = (x_t, y_t) \) where \( t \) is the frame number. A trajectory can be written as \( (P_t, P_{t+1}, ..., P_{t+L-1}) \) where \( L \) is the length, in frames, of the trajectory. We store our trajectory data in the form of patches \( \{P_t = (f(P_t), c_t, d_t)\} \) where \( f(P_t) \) are appearance
and motion features for a given point of a trajectory, \( c_i \) is the class label and \( d_i \) is a vector pointing to the spatio-temporal center of the trajectory in the fingertip writing problem and to the spatio-temporal center of the action in the human action recognition problem.

We formulate the spatio-temporal trajectory recognition as a multi-class classification problem and action center localization as regression. To perform them simultaneously, we build upon Hough forest [8]. An important difference between [8] framework and ours is that we extract patches by sampling along trajectories instead of dense sampling cuboids (see Figure 3).

A Hough Forest is an ensemble of decision trees that, in addition to classification, they also perform regression. Each tree in the forest is constructed from a set of patches extracted along the trajectories \( \{ P_i \} \). Tree training starts at the root and input data is divided and rooted left or right following a split function. Split candidates are generated randomly and the best split is chosen based on an objective function that is minimized. This objective function is randomly chosen between Shannon entropy, which minimizes class uncertainty, and variance of displacement vectors, a regression measure that tends to group similar vectors. If the current node reaches a certain depth or a good split cannot be found, it becomes a leaf node. At each leaf node \( l_i \), a class histogram \( p(c|l) \) is estimated by the proportion of trajectory patches per class that reached that node. Both histogram and displacement vectors \( d_i \) are stored.

During inference, patches are passed through each tree in the trained forest. Starting at the root of the forest the patch traverses the tree, branching left or right according to the split node function, until reaching a leaf node. Using the stored class distribution and vector displacements at the leaf nodes, each leaf node votes for its corresponding class label and spatio-temporal center location. Each patch votes in a 4D Hough space and the most likely hypothesis can be found searching the maxima. We refer the reader to [8] for further details.

### 3.2. Transition Hough forest

A major drawback of using a decision forest based classifier for sequential data is that the forest produces each estimation independently of its temporal context. This assumption can be too strong in the problem of recognizing spatio-temporal trajectories. For instance, the human action of punching involves the movement of an arm in a particular direction and speed. This movement follows a certain temporal order that makes it different from similar actions such as pulling.

As presented in the previous section, a Hough Forest reduces both class and displacement uncertainty throughout the tree. The leaf nodes will contain similar patches both in displacement, feature-space and class, thus it can be seen as a clusters of similar patches. Such idea of using a decision forest framework for clustering is not new and it has been explored in other areas such as semantic image segmentation (e.g. [21, 29]), but relatively less for action recognition [35]. From this perspective, we can see a spatio-temporal trajectory as a time-indexed sequence of codebook values.

Based on this, we introduce our concept of transition. Our hypothesis is that different classes of spatio-temporal trajectories will have a different temporal dynamics within the forest. For example, if we observe that in a given frame \( t \) the trajectory patch \( P_t \) has reached the node \( i \) while in the previous frame \( t - 1 \) the corresponding patch \( P_{t - 1} \) reached the node \( j \), we can quantify how likely is the transition from node \( j \) to node \( i \) or, more formally, \( p(n(t) = i | n(t - 1) = j) \) a certain class. We name this last term as transition probability borrowed from HMM literature [25]. We define our transitions for one time step, thus we will ignore the time index in the following sections. [29] showed that adding non-terminal nodes while constructing codebooks captured the hierarchical structure of the tree leading to a better performance. Accordingly, we consider transitions between both leaf and split nodes. Although in practice trees are not balanced and transitions can be observed between different levels of the tree, we ignore them maintaining its hierarchical nature considering only same level transitions. In order to compact this information, we define a transi-
tion matrix $A(c,l)$ that encode all transitions between nodes for a given class $c$ and level $l$ in one time step. Rows of $A(c,l)$ encode transition probabilities from node $i$ to all the rest of the nodes $j$ in a particular level $l$ of the tree and they are normalized defining a probability distribution $(\sum_j p(n = j | n = i)) = 1, n \in l$. See Figure 4 for further details. Note that we will have a transition matrix for every tree in the forest, however we omitted this to make the notation clear.

In order to incorporate this temporal information into our predictions, we treat this transition probability as a prior probability $p(c)$ in a similar way to [29]. We want transitions to emphasize classes that are likely in a temporal context and reject unlikely ones. Given two temporal consecutive patches from a trajectory, $P_t$ and $P_{t-1}$, we pass both patches through the forest and each of them reaches different nodes through each tree of the forest. We ponder the prediction for $P_t$, $p(c|l_t)$, with the (almost independent) prior probability:

$$p'(c|l_t) = p(c|l_t)p(c)$$

(1)

with $p(c)$ defined as the averaged transition probability $p_{\text{trans}}$, of all $T$ trees in the forest soften by a constant $\alpha$:

$$p(c) = \frac{1}{T} \sum_{k=0}^{T} p_{\text{trans}}^c(c,k)$$

(2)

$p_{\text{trans}}$ is calculated from our transition matrices defined above:

$$p_{\text{trans}}^c(c,k) = \frac{1}{W} \sum_{l=0}^{D_k} A^k(c,l)$$

(3)

where $D_k$ is the maximum level reached in the $k$-th tree by $P_{t-1}$ or $P_t$ and $W$ is a factor that ensures probability normalization. $D_k$ is not necessarily the total depth of the tree $D$ as leaf nodes can be found at any level of the tree.

### 3.3. Implementation details

#### 3.3.1 Fingertip writing trajectories

We extract fingertip writing spatio-temporal trajectories by tracking the index fingertip in space and time. Instead of standard RGB video, we decided to use a depth sensor. Using depth data makes the problems of hand segmentation and fingertip detection easier. We detect and track the fingertip using the approach of [15] where hand contour is obtained and fingertips are tracked using a distance transform and a particle filter respectively. Once obtained the spatio-temporal trajectories, we extract local features that we encode in $f(P_t)$. These features are extracted using a temporal sliding window of $n_\tau$ frames. This parameter defines the length of strokes encoded in patches. Small values of $n_\tau$ may not allow us to properly capture motion, while large ones could give us non meaningful information. At each temporal window, we concatenate the following local features: displacement vector between points, curvature, distance and velocity. Distance and velocity are defined only between the first and the last point in the window and we considered both euclidean distance and geodesic distance. Their temporal derivatives provide us complementary information about the writing stroke. Note that for these trajectories $L$ does not have a fixed size and it will depend on their category.

#### 3.3.2 Action recognition trajectories

For tracking and extracting features along spatio-temporal trajectories on video data, we follow the approach from [33]. We chose this method mainly because its excellent results and its publicly available code, however other spatio-temporal trajectory representation would be also valid. In [33], each trajectory point is tracked at different scales using optical flow. Tracked points are sampled in small volumes of size $n_\sigma \times n_\sigma \times n_\tau$ and rich feature descriptors, HOGHOF [13] and MBH [5], are extracted. We encode all
this information in our trajectory patches \( f(P_i) \). In contrast to [33], we do not concatenate the descriptors of all the points in the trajectory nor we average them. Instead, we treat each point of the trajectory independently and we store it as a patch. Our trajectories are defined as ensembles of \( L \) independent patches.

4. New dataset: Egocentric fingertip writing

As there exists no public dataset of fingertip written trajectories in mid-air using an egocentric sensor, we recorded our own one, which we plan to make public for further research. Our dataset is composed of depth video sequences containing fingertip written trajectories that represent the 26 English alphabet characters (from ‘a’ to ‘z’). We attached a depth sensor (Creative Interactive Gesture Camera) to a cap in order to be able to record gestures in egocentric viewpoint. In total, 10 sequences of 26 different characters performed by a single actor have been recorded (making a total of 260). Furthermore, we fully annotated the sequences with \((x, y, t)\) fingertip positions after our detection and tracking stage to help research on this direction as well. See Figure 5 for some examples of our recorded sequences and Table 1 for further details.

| Classes | 26 | Total frames | 15792 |
| Clips | 260 | Clips per class | 10 |
| Mean clip frames | 60.74 | Resolution | 320x240 |
| Min clip frames | 27 | Max clip frames | 154 |

Table 1. Characteristics of the dataset

5. Experiments

5.1. Egocentric fingertip writing

Character recognition: In table 2 we present the performance of different methods on our new dataset. All the results have been obtained performing 10 leave-one-out cross validation (234 sequences for training and 26 for testing). We chose empirically a sliding window size of \( n_r = 7 \). We first show the results of two classical algorithms for sequential data recognition, Hidden Markov Model (HMM) [25] and Dynamic Time Warping (DTW) [31, 4]. Although neither of these methods is suitable for our application since they do not perform localization, we include them for completeness of this work. Next, we show the results for decision forest-based classifiers: a conventional decision forest [2], our framework without transition term and the full framework. For all forest based algorithms we fixed \( T = 8 \) and \( D = 25 \), optimizing cross validated results. We see that our proposed Transition Hough forest outperform all the other approaches. Introducing the transition term slightly improves the accuracy in a 1.5%. Comparing with the conventional random forest, we note that including localization also helped classification, as it was already pointed in [8].

<table>
<thead>
<tr>
<th>Recognition</th>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character recognition</td>
<td>HMM [25]</td>
<td>66.4</td>
</tr>
<tr>
<td></td>
<td>DTW [31]</td>
<td>78.5</td>
</tr>
<tr>
<td></td>
<td>Decision forest [2]</td>
<td>79.6</td>
</tr>
<tr>
<td></td>
<td>Trajectory Hough forest</td>
<td>90.4</td>
</tr>
<tr>
<td></td>
<td>Transition Hough forest</td>
<td><strong>91.9</strong></td>
</tr>
</tbody>
</table>

Table 2. Recognition performance of fingertip writing.

From the confusion matrix (Figure 6), we can observe that most errors came from similar characters such as ‘a-d’, ‘m-n’, ‘g-q’ and ‘v-w’, which are all of them very similar and sometimes difficult to recognize even for humans. We believe that adding a broader temporal context could help on these cases.

Character center localization: Our method can also correctly localize spatio-temporal center of each character writing by spatio-temporal offset Hough voting. Figure 7 shows some localization results in spatio-temporal space. As we can see, estimated centers are similar to ground-truth ones. The writing center information of each character can be used as an important clue for segmenting each character in a word or to anchor where the user wrote in an augmented reality scenario.

5.2. Action Recognition: UT-Interaction dataset

To demonstrate the effectiveness of our proposed method for human action recognition, we conducted experiments on a public benchmark: UT-interaction dataset [27]. The UT-interaction dataset consists in 6 different classes of human-human interactions in a surveillance scenario: shake-hands,
point, hug, push, kick and punch (an example of kick and punch is shown in Figure 3). We used the segmented set 1 of the dataset which contains 10 sequences per each class. We followed the methodology recommended by the authors and we performed 10-fold leave-one-out cross validation to find the average performance.

In table 3 we present the performance of our method compared to baseline and other state-of-the-art methods. The parameters for extracting trajectories were the recommended by [33] \( n_r = 2 \) and \( L = 15 \). We used \( n_r = 1 \), meaning that a patch was generated for every frame in the trajectory. We defined our baseline as the Hough forest using trajectory-based patches and we also compared to the conventional Hough forest using dense cuboid sampling [8]. Forests parameters are \( T = 4 \) and \( D = 35 \).

We observe that using trajectory sampled descriptors instead of dense cuboids slightly improves the recognition accuracy in a 2%, which is in line with what was reported in [32]. In top of that, we show that adding our transition term further improves the performance in a 3.3% from the baseline making it comparable to state-of-the-art performances.

Our approach offers a similar performance to [26] even if they use more sophisticated medium-level features encoding the pose of humans in scene, which are usually hard to annotate and obtain. Compared to [36], our method performs not as well as theirs. The main reason for this is that we rely on very local spatio-temporal context while in [36] they also consider long range spatio-temporal relations. Finally, we also show the result from [35] where they also used the clustering capability of a decision forest, however important spatio-temporal information is lost when histogram quantization is performed.

### 6. Conclusion and Future Work

We have presented a novel framework for recognizing and localizing spatio-temporal trajectories using a Hough forest-based classifier and showed that it is general enough to be applied in different scenarios. We have introduced a new concept of transition that makes forest predictions sensitive to their output temporal context without losing efficiency. As a future work, we plan to investigate how can we make these transitions more discriminative within the forest, enforcing the transitions at feature level or designing a novel split criteria that privileges transitions. Furthermore, we also plan to explore the introduction of long-range temporal context in contrast to only exploiting the context of time-consecutive patches.
References


