Guest Editors' Introduction to the Special Issue on Multimodal Human Pose Recovery and Behavior Analysis

Sergio Escalera, Jordi Gonzàlez, Xavier Baró, and Jamie Shotton

HUMAN Pose Recovery and Behavior Analysis (HuPBA) is one of the most challenging topics in Computer Vision, Pattern Analysis and Machine Learning. It is of critical importance for application areas that include gaming, computer interaction, human robot interaction, security, commerce, assistive technologies and rehabilitation, sports, sign language recognition, and driver assistance technology, to mention just a few. In essence, HuPBA requires dealing with the articulated nature of the human body, changes in appearance due to clothing, and the inherent problems of clutter scenes, such as background artifacts, occlusions, and illumination changes. Given these inherent difficulties, the combination of alternative, complementary visual and non-visual modalities coming from different types of sensors has drawn a lot of attention in the literature: sensor data from visual cameras like RGB, time-of-flight (ToF), infrared, light field, multispectral, underwater, or thermal wavelengths cameras, together with another non-visual sensors like audio signals, inertial measurement unit (IMU) data, electrothermal activity responses, or electroglottograph signals, among others, have been exploited and combined to estimate the pose, gesture and behavior in both single images and image sequences. The combination of these visual and non-visual modalities has increased the accuracy of computer vision approaches, although gives rise to new challenges with feature extraction, synchronization of data coming from different sensors, data fusion, and temporal series analysis.

As Guest Editors of this Special Issue on Multimodal Human Pose Recovery and Behavior Analysis (M2HuPBA), we are happy to present 16 accepted papers that represent the most recent research in this field, including new methods considering still images, image sequences, depth data, stereo vision, 3D vision, audio, and IMUs, among others, while presenting new multimodal datasets, in addition to those proposed by the ChaLearn Looking at People series of workshops.[1] We would like to thank the authors of the 57 submissions we received, and above all the outstanding and timely work performed by the reviewers. All the 57 submissions followed a rigorous TPAMI review process, where at least three external reviewers provided reviews to each paper. We would also like to thank the Editor-In-Chief (EIC), David Forsyth, for making this special issue possible. We are also grateful to the editorial staff for managing the submission process and providing us with assistance.

The set of 16 accepted papers can be split into three main categories within M2HuPBA: (i) human pose recovery and tracking; (ii) action and gesture recognition; and (iii) datasets. We describe these next.

1 Human Pose Recovery and Tracking

The first group of 6 papers are centered on multimodal human pose analysis.

The paper "Spatio-Temporal Matching for Human Pose Estimation in Video" by F. Zhou and F. de la Torre formulates the problem of human detection in videos as spatio-temporal matching (STM) between a 3D motion capture model and trajectories in videos. The algorithm estimates the camera view and selects a subset of tracked trajectories that matches the motion of the 3D model. The STM is efficiently solved with linear programming, and it is robust to tracking mismatches, occlusions and outliers.

The work "3D Reconstruction of Human Motion from Monocular Image Sequences" by B. Wandt, H. Ackermann, and B. Rosenhahn, estimates non-rigid human 3D shape and motion from image sequences taken by uncalibrated cameras. Authors factorize 2D observations in camera parameters, base poses and mixing coefficients. The method is based on previously trained base poses. Strong periodic assumptions on the coefficients can be used to define an efficient and accurate algorithm for estimating periodic motion, and non-periodic motion is estimated by including a regularization term based on temporal bone length constancy.

The paper entitled "Real-Time Simultaneous Pose and Shape Estimation for Articulated Objects using a Single Depth Camera" by M. Ye, Y. Shen, C. Du, Z. Pan, and R. Yang presents a real-time algorithm for simultaneous pose and shape estimation of articulated objects, such as human beings and animals. The key to the pose estimation

- *S. Escalera is with the Mathematics and Informatics Department, University of Barcelona, and the Computer Vision Center, Catalonia, Spain (http://www.sergioescalera.com). E-mail: sergio@maia.ub.es.*
- *J. Gonzàlez is with the Universitat Autònoma de Barcelona and the Computer Vision Center, Catalonia, Spain. E-mail: jordi.gonzalez@cvc.uab.es.*
- *X. Baró is with the Universitat Oberta de Catalunya and the Computer Vision Center, Catalonia, Spain. E-mail: xbaro@uoc.edu.*
- *J. Shotton is with Microsoft Research, Cambridge, United Kingdom. E-mail: Jamie.Shotton@microsoft.com.*

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2557878

1. http://gesture.chalearn.org/

component is to embed the articulated deformation model with exponential-maps-based parametrization into a Gaussian mixture model. A shape adaptation algorithm based on this probabilistic model automatically captures the shape of the subjects during the dynamic pose estimation process.

The paper by G. Pons-Moll, T. von Marcard, and B. Rosenhahn entitled "Human Pose Estimation from Video and IMUs" presents an approach to fuse video with sparse orientation data obtained from inertial sensors to improve and stabilize full-body human motion capture. The authors propose a hybrid tracker that combines video with a small number of inertial units to compensate for the drawbacks of each sensor type.

The work "Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications" by C. Corneanu, M. Oliu, J. Cohn, and S. Escalera presents an up to date survey on facial expression analysis by considering different visual modalities, defining a new taxonomy encompassing all steps from face detection to facial expression recognition. Important datasets and future lines of research in the field are also discussed.

Finally, the work "Full-Body Pose Tracking—the Top View Reprojection Approach" by M. Sigalas, M. Pateraki, and P. Trahanias presents a model-based approach for markerless articulated full body pose extraction and tracking in RGB-D sequences. A cylinder-based model is employed to represent the human body. For each body part a set of hypotheses is generated and tracked over time using a particle filter. To evaluate each hypothesis, authors employ a novel metric that considers the reprojected top view of the corresponding body part. The latter, in conjunction with depth information, effectively copes with difficult and ambiguous cases, such as severe occlusions.

2 Action and Gesture Recognition

The second group of nine papers in this special issue are focused on multimodal action and gesture recognition.

The paper "Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition" by D. Wu, L. Pigou, P. J. Kindermans, N. Le, L. Shao, J. Dambre, and J. M. Odobez uses deep neural networks for multimodal gesture recognition. A semi-supervised hierarchical dynamic framework based on a hidden Markov model (HMM) is proposed for simultaneous gesture segmentation and recognition where skeleton joint information, depth and RGB images, are the multimodal input observations. The system learns high-level spatio-temporal representations using deep neural networks: a Gaussian-Bernouilli deep belief network to handle skeletal dynamics, and a 3D convolutional neural network to manage and fuse batches of depth and RGB images. This is achieved through the modeling and learning of the emission probabilities of the HMM required to infer the gesture sequence.

The paper "Semantic Event Fusion of Different Visual Modality Concepts for Activity Recognition" by C. Crispim-Junior, K. Avgerinakis, V. Buso, G. Meditskos, A. Briassouli, J. Benois-Pineau, Y. Kompatsiaris, and F. Bremond proposes a hybrid framework between knowledge-driven and probabilistic-driven methods for event representation and recognition. It introduces an algorithm for sensor alignment that uses semantic data similarity as a surrogate for the inaccurate temporal information of real life scenarios. It also proposes the combined use of an ontology language and a probabilistic interpretation for ontological models. The system is evaluated in multimodal scenarios considering different visual sources, such as RGB and depth maps.

The work "Nonparametric Feature Matching Based Conditional Random Fields for Gesture Recognition from Multimodal Video" by J. Y. Chang presents a gesture recognition method based on a conditional random field (CRF) using multiple feature matching. The approach determines gesture categories and their temporal ranges at the same time. A generative probabilistic model is formalized and probability densities are non-parametrically estimated by matching input features with a training dataset. Frame-wise recognition results can then be obtained by applying an efficient dynamic programming technique. To estimate the parameters of the proposed CRF model, the structured support vector machine (SSVM) framework is incorporated.

The paper "Explore Efficient Local Features from RGB-D Data for One-Shot Learning Gesture Recognition" by J. Wan, G. Guodong Guo, and S. Li proposes a spatio-temporal feature extracted from RGB-D data, namely Mixed Features around Sparse Keypoints (MFSK). The proposed MFSK feature is robust and invariant to scale, rotation and partial occlusions. It shows high categorization capability in both one-shot and classic gesture recognition problems.

In the paper "Labeled Graph Kernel for Behavior Analysis" by A. Martinez and R. Zhao, it is proposed to model behavior using a labeled graph, where the nodes define behavioral features and the edges are labels specifying their order. In this approach, classification reduces to a simple labeled graph matching. Graph kernel is derived to quickly and accurately compute graph similarity.

The work "Structure-Preserving Binary Representations for RGB-D Action Recognition" by L. Shao, M. Yu, and L. Liu proposes a binary local representation for RGB-D video data fusion with a structure-preserving projection (SPP). The authors convert the problem to describing the gradient fields of RGB and depth information of video sequences. With the local fluxes of the gradient fields, which include the orientation and the magnitude of the neighborhood of each point, a new kind of continuous local descriptor called Local Flux Feature (LFF) is obtained. The LFFs from RGB and depth channels are fused into a Hamming space via the Structure Preserving Projection, and a bipartite graph structure of data is taken into consideration as a high level connection between samples and classes.

The work "Robust Correlated and Individual Component Analysis" by Y. Panagakis, M. Nicolaou, S. Zafeiriou, and M. Pantic addresses (i) the presence of gross non-Gaussian noise, and (ii) temporally misaligned data. The authors propose a method for the Robust Correlated and Individual Component Analysis (RCICA) of data and two suitable optimization problems are solved. The generality of the proposed methods is demonstrated by applying them onto 4 applications, namely (i) heterogeneous face recognition and (ii) audio-visual feature fusion for prediction of interest and conflict, (iii) face clustering, and (iv) the temporal alignment of facial expressions.

The paper "Probabilistic Social Behavior Analysis by Exploring Body Motion-Based Patterns" by K. Khoshhal, U. Nunes, and J. Dias explores interrelations between body part motions in scenarios with people doing a conversation. The authors analyze body motion-based features in frequency domain to estimate different human social patterns: interpersonal behaviors and a social role. To analyze the dynamics and interrelations of people's body motions, a human movement descriptor is used to extract discriminative features, and a multi-layer dynamic Bayesian network technique is proposed to model the existent dependencies.

Lastly, the work "Moddrop: Adaptive Multimodal Gesture Recognition" by N. Neverova, C. Wolf, G. Taylor, and F. Nebout presents a method for gesture detection and localization based on multi-scale and multimodal deep learning. The training strategy performs gradual fusion involving random dropping of separate channels (dubbed ModDrop) for learning cross-modality correlations while preserving uniqueness of each modality-specific representation. The proposed training technique ensures robustness of the classifier to missing signals in one or several channels to produce meaningful predictions from any number of available modalities.

3 New Datasets

The final paper in this special issue describes a new dataset.

The work "SALSA: A Novel Dataset for Multimodal Group Behaviour Analysis" by X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe presents a dataset facilitating multimodal and Synergetic sociAL Scene Analysis. SALSA records social interactions among 18 participants in a natural, indoor environment for over 60 minutes. Challenges in the data include low-resolution images, lighting variations, numerous occlusions, reverberations and interfering sound sources. The authors also facilitate multimodal analysis by recording the social interplay using four static surveillance cameras and sociometric badges worn by each participant, comprising the microphone, accelerometer, bluetooth and infrared sensors. Additional annotations are also provided: individuals' personality, position, head, body orientation and F-formation information over the entire event duration.

4 Research Outlook

This special issue summarizes the state of the art in the field of M2HuPBA, and presents interesting work that push the state of the art in the field, opening new lines of research and areas of applications. But we foresee many novel contributions still needed in order to solve some of the current requirements related to the generalization of human pose recovery and behavior analysis to uncontrolled environments and in the presence of arbitrary artifacts. We expect that new non-conventional modalities and fusion strategies with the support of new datasets for their evaluation will be important for further such advances.

Sergio Escalera received the PhD degree in multi-class learning from the Computer Vision Center, UAB. He received the 2008 best Thesis award. He leads the Human Pose Recovery and Behavior Analysis Group at UB, CVC, and BGSMath. He is an associate professor at the Mathematics and Informatics Department, UB. He is a partial time professor at UOC. He is a member of the Visual and Computational Learning consolidated research group of Catalonia. He is member of the CVC. He is the Editor-in-Chief and editorial member of more than five journals. He is the director of ChaLearn. He is the co-founder of PhysicalTech and Care Respite companies. He is a member of BGSCMath, AERFAI, ACIA, and the vice-chair of IAPR TC-12: Multimedia and visual information systems. He has different patents and registered models. He has published more than 150 research papers and organized scientific events, including CCIA2004, CCIA2014, ICCV2011, and workshops at ICCV2011, ICMI2013, ECCV2014, CVPR2015, ICCV2015, CVPR2016. He has been the guest editor at IJCV, Neural Computing, TPAMI, JMLR. He has been AC of WACV2016 and NIPS2016.

Jordi Gonzàlez received the PhD degree in computer engineering from University Autnoma de Barcelona (UAB) in 2004. He is an associate professor in computer science at the Computer Science Department, UAB. He is also a research fellow at the Computer Vision Center, where he has co-founded three spin-offs and the Image Sequence Evaluation (ISE Lab) research group. His research interests include machine learning techniques for the computational interpretation of human behaviors in social images, or visual hermeneutics. He is a member of the Editorial Board of the journals Computer Vision and Image Understanding and IET Computer Vision.

Xavier Baró received the PhD degree in computer engineering from the Universitat Autnoma de Barcelona in 2009. He is currently a lecturer and researcher at the IT, Multimedia and Telecommunications Department at Universitat Oberta de Catalunya (UOC). He is the co-founder of the Scene Understanding and Artificial Intelligence group of the UOC, and collaborates with the Computer Vision Center of the UAB, as a member of the Human Pose Recovery and Behavior Analysis Group. His research interests include machine learning, evolutionary computation, and statistical pattern recognition, specially their applications to generic object recognition over huge cardinality image databases.

Jamie Shotton studied computer science at the University of Cambridge. He leads the Machine Intelligence & Perception Group at Microsoft Research Cambridge. He joined Microsoft Research in 2008 where he is currently a Principal Researcher. His research interests include the intersection of computer vision, AI, machine learning, and graphics, with particular emphasis on systems that allow people to interact naturally with computers. He has received multiple Best Paper and Best Demo awards at top academic conferences. His work on machine learning for body part recognition for Kinect was awarded the Royal Academy of Engineering's gold medal MacRobert Award 2011, and he shares Microsoft's Outstanding Technical Achievement Award for 2012 with the Kinect engineering team. In 2014, he received the PAMI Young Researcher Award, and in 2015 the MIT Technology Review Innovator Under 35 Award.