ChaLearn Multi-Modal Gesture Recognition 2013: Grand Challenge and Workshop Summary

Sergio Escalera
Dept. Applied Mathematics, Universitat de Barcelona
Computer Vision Center, UAB
sergio@maia.ub.es

Jordi Gonzàlez
Dept. Computer Science, Univ. Autònoma de Barcelona
Computer Vision Center, UAB
poal@cvc.uab.es

Xavier Baró
EIMT at the Open University of Catalonia, Barcelona
Computer Vision Center, UAB
xbaro@uoc.edu

Miguel Reyes
Dept. Applied Mathematics, Universitat de Barcelona
Computer Vision Center, UAB
mreyese@gmail.com

Isabelle Guyon
ChaLearn, Berkeley, California
guyon@chalearn.org

Vassilis Athitsos
University of Texas
athitsos@uta.edu

Hugo J. Escalante
INAOE, Puebla, Mexico
hugojair@inaoep.mx

Leonid Sigal
Disney Research, Pittsburgh
lsigal@disneyresearch.com

Antonis Argyros
Institute of Computer Science, FORTH
argyros@ics.forth.gr

Cristian Sminchisescu
Lund University
sminchisescu@gmail.com

Richard Bowden
University of Surrey
r.bowden@surrey.ac.uk

Stan Sclaroff
Department of Computer Science, Boston University
sclaroff@bu.edu

ABSTRACT

We organized a Grand Challenge and Workshop on Multi-Modal Gesture Recognition.

The MMGR Grand Challenge focused on the recognition of continuous natural gestures from multi-modal data (including RGB, Depth, user mask, Skeletal model, and audio). We made available a large labeled video database of 13,858 gestures from a lexicon of 20 Italian gesture categories recorded with a Kinect^{TM} camera. More than 54 teams participated in the challenge and a final error rate of 12% was achieved by the winner of the competition. Winners of the competition published their work in the workshop of the Challenge.

The MMGR Workshop was held at ICMI conference 2013, Sidney. A total of 9 relevant papers with basis on multi-modal gesture recognition were accepted for presentation. This includes multi-modal descriptors, multi-class learning strategies for segmentation and classification of temporal data, as well as relevant applications in the field, including multi- modal Social Signal Processing and multimodal Human Computer Interfaces. Five relevant invited speakers participated in the workshop: Profs. Leonid Signal from Disney Research, Antonis Argyros from FORTH, Institute of Computer Science, Cristian Sminchisescu from Lund University, Richard Bowden from University of Surrey, and Stan Sclaroff from Boston University. They summarized their research in the field and discussed past, current, and future challenges in Multi-Modal Gesture Recognition.

ICMI'13, December 9–13, 2013, Sydney, Australia

http://dx.doi.org/10.1145/2522848.2532597.

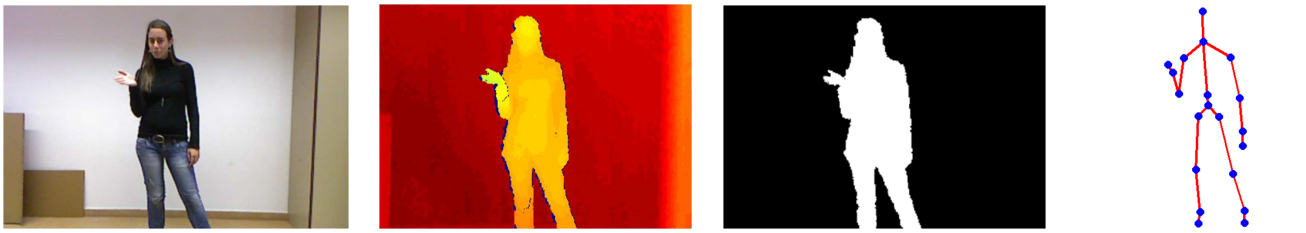
1. MMGR CHALLENGE SUMMARY

The focus of the challenge was on *user independent multiple gesture learning*. There are no resting positions and the gestures are performed in continuous sequences lasting 1-2 minutes, containing between 8 and 20 gesture instances in each sequence. In this challenge we focus on the recognition of a vocabulary of 20 Italian cultural/anthropological signs. As a result, the dataset contains around 1.720.800 frames. In addition to the 20 main gesture categories, 'distracter' gestures are included, meaning that additional audio and gestures out of the vocabulary are included. In all the sequences, a single user is recorded in front of a Kinect^{TM}, performing natural communicative gestures and speaking in fluent Italian. All the characteristics of the data are described in detail in [1]. The dataset is available at `http://sunai.uoc.edu/chalearn`. An example of the provided visual modalities is shown in Figure 1.

The final evaluation of the challenge was defined in terms of the Levenshtein edit distance, where the goal was to indicate the real order of gestures within the sequence. 54 international teams participated in the challenge, and outstanding results were obtained by the first ranked participants [1].

1.1 Challenge schedule

The challenge consisted of two main components: a development phase (April 30th to Aug 1st) and a final evaluation



RGB Depth User mask Skeletal model

Figure 1: Different data modalities of the provided data set.

phase (Aug 2nd to Aug 15th). The submission and evaluation of the challenge entries was via the *Kaggle* platform [1]. The official participation rules were provided on the website of the challenge. In addition, publicity and news on the ChaLearn Multi-modal Gesture Recognition Challenge were published in well-known online platforms, such as LinkedIn, Facebook, Google Groups and the ChaLearn website.

During the development phase, the participants were asked to build a system capable of learning from several gesture samples a vocabulary of 20 Italian sign gesture categories. To that end, the teams received the development data to train and self-evaluate their systems. In order to monitor their progress they could use the validation data for which the labels were not provided. The prediction results on validation data could be submitted online to get immediate feed-back. A real-time leaderboard showed to the participants their current standing based on their validation set predictions.

During the final phase, labels for validation data are published and the participants performed similar tasks as those performed in previous phase, using the validation data and training data sets in order to train their system with more gesture instances. The participants had only few days to train their systems and upload them. The organizers used the final evaluation data in order to generate the predictions and obtain the final score and rank for each team. At the end, the final evaluation data was revealed, and authors submitted their own predictions and fact sheets to the platform.

1.2 Challenge results

The challenge attracted high level of participation, with a total of 54 teams and near 300 total number of entries. This is a good level of participation for a computer vision challenge requiring very specialized skills. Finally, 17 teams successfully submitted their prediction in final test set, while providing also their code for verification and summarizing their method by means of a fact sheet questionnaire.

After verifying the codes and results of the participants, the final scores of the top rank participants on both validation and test sets were made public. In the end, the final error rate on the test data set was around 12%. The details about the score achieved by top ranked participants as well as the analysis of the methods are described in [1].

2. MMGR WORKSGHOP SUMMARY

A total of 9 relevant papers with basis on multi-modal gesture recognition were accepted for presentation at the workshop. They include multi-modal descriptors, multi-class learning strategies for segmentation and classification in temporal data, as well as relevant applications in the field, including multi-modal Social Signal Processing, and multi-modal Human Computer Interfaces. Five relevant invited speakers participated in the workshop: Profs. Leonid Signal from Disney Research, Antonis Argyros from FORTH, Institute of Computer Science, Cristian Sminchisescu from Lund University, Richard Bowden from University of Surrey, and Stan Sclaroff from Boston University. They summarized their research in the field and discussed past, current, and future challenges in Multi-Modal Gesture Recognition. The list of accepted papers and summary of invited speaker talks are described next.

2.1 Grand Challenge and Workshop papers

The nine accepted papers were distributed into four main oral sessions as follows.

Oral session I: Multi-modal Gesture Recognition Challenge I

- Multi-modal Gesture Recognition Challenge 2013: Dataset and Results
Authors: Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Oscar Lopés, Isabelle Guyon, Vassilis Athistos, and Hugo J. Escalante
- Fusing Multi-modal Features for Gesture Recognition **(1st Challenge Prize)**
Authors: Jiaxiang Wu, Jian Cheng, Chaoyang Zhao, and Hanqing Lu
- A Multi Modal Approach to Gesture Recognition from Audio and Video Data **(3rd Challenge Prize)**
Authors: Immanuel Bayer and Thierry Silbermann

Oral session II: Multi-modal Gesture Recognition Challenge II

- Online RGB-D Gesture Recognition with Extreme Learning Machines
Authors: Xi Chen and Markus Koskela
- A Multi-modal Gesture Recognition System Using Audio, Video, and Skeletal Joint Data
Authors: Karthik Nandakumar, Kong-Wah Wan, Jian-Gang Wang, Wen Zheng Terence Ng, Siu Man Alice Chan, and Wei-Yun Yau

Oral session III: Challenge for Multimodal Mid-Air Gesture Recognition for close HCI

- A Challenge for Multimodal Mid-Air Gesture Recognition for close HCI
Authors: Simon Ruffieux, Denis Lalanne, and Elena Mugellini
- ChAirGest 2013 - Continuous Gesture Spotting and Recognition

[1] https://www.kaggle.com/c/multi-modal-gesture-recognition

Authors: Ying Yin and Randall Davis

Oral session IV: Multi-modal Gesture Recognition Applications

• Multi-modal Social Signal Analysis for Predicting Agreement in Conversation Settings
Authors: Víctor Ponce, Sergio Escalera, and Xavier Baró
• Multi-modal Descriptors for Multi-class Hand Pose Recognition in Human Computer Interaction Systems
Authors: Jordi Abella, Raúl Alcaide, Anna Sabaté, Joan Mas, Sergio Escalera, Jordi Gonzàlez, and Coen Antens

2.2 Invited speakers

Leonid Sigal, Disney Research,
Recognition and Understanding: Latest Challenges and Opportunities
The recognition and interpretation of human actions from video is a key enabling technology for variety of applications, including those in behavior analytics, video understanding, and human computer interactions. However, despite much research, and many advances along the way, the general problem remains challenging. I will outline the nature of the problem and describe some recent advances that focus on the ability to model action semantics, localize actions both spatially and temporally as well as understand actions at various levels of granularity (through hierarchical representations). I will also try to highlight potential directions for future research.

Antonis Argyros, FORTH, Institute of Computer Science,
Tracking the articulated motion of human hands
Humans use their hands in most of their everyday life activities. Thus, the development of technical systems that track the 3D position, orientation and full articulation of human hands from markerless visual observations can be of fundamental importance in supporting a number of diverse applications. In this talk, we provide an overview of our work on hand tracking. First, we describe methods for vision-based detection and tracking of hands and fingers in 2D, with emphasis on occlusions handling and illumination invariance. We also demonstrate hand posture recognition techniques and their use in Human Computer Interaction (HCI) and Human Robot Interaction (HRI). Then, we focus on a recently proposed framework for exploiting markerless visual observations to track the 3D position, orientation and full articulation of a human hand that moves in isolation in front of an RGBD camera. We treat this as an optimization problem that is effectively solved using a variant of Particle Swarm Optimization (PSO). Next, we show how the core of the tracking framework has been employed to provide state-of-the-art solutions for problems of even higher dimensionality and complexity, e.g., for tracking two strongly interacting hands or for tracking the state of a complex scene where a hand interacts with several objects. Finally, we demonstrate how the results of hand tracking have been used to recognize human actions and infer human intentions in the context of tabletop object manipulation scenarios.

Cristian Sminchisescu, Lund University,
Human Actions and 3D Pose in the Eye: From Perceptual Evidence to Accurate Computational Models
Recent progress in computer-based visual recognition, in particular image classification, object detection or action recognition heavily relies on machine learning techniques trained on large scale annotated datasets. While such data has made advances in model design and evaluation possible, it does not necessarily provide insights or constraints into those intermediate levels of computation, or deep structure, perceived as ultimately necessary in order to design highly reliable computer vision systems. This is noticeable in the accuracy of state of the art systems trained with such annotations, which still lags significantly behind human performance in similar tasks. Nor does the existing data make it immediately possible to exploit insights from a working system - the human eye - to potentially derive better features, models or algorithms. In this talk I will provide an overview of our research in human action recognition as well as 3d human pose estimation, which relies on large-scale human eye movement and 3d body motion capture datasets, collected in the context of visual recognition tasks[2]. I will show that: (1) the human fixation patters are stable, both statically, and sequentially as well as dynamically; (2) they are influenced by the task; (3) fixation detectors as well as scan-paths estimators can be effectively learned from human eye movement data, and (4) such learnt detectors can be used as interest point operators, leading to state of the art recognition results when used in end-to-end automatic recognition systems. I will also discuss perhaps non-intuitive quantitative empirical evidence regarding the human capability to perceive 3d articulated poses, indicating that humans are not particularly good at extracting (and re-enacting) metrically accurate 3d pose information from images. Time permitting I will also cover a moving pose, fast and accurate kinematic action detection and recognition framework, based on RGB-D sensors. This is joint work with S. Mathe, E. Marinoiu, D. Papava, V. Olaru, M. Leordeanu and M. Zanfir.

Richard Bowden, University of Surrey,
Recognising spatio-temporal events in video
Learning to recognise patterns in video is a "balancing act" between representing the signal with sufficient complexity to accurately describe the subtleties of appearance and motion while employing a simple representation that can generalise or provide invariance to those aspects of the video which do not convey meaning. A standard approach is to employ a fixed, high dimensional representation and employ statistics to identify which features are important, but this requires sufficient training data. The notion of sufficient is dependent on the complexity of the representation. An alternative approach is to perform spatio-temporal feature selection to try and identify the simplest signature required for classification. However, the search space has high combinatorial complexity and it is therefore a computationally demanding task. This talk will discuss a generic solution to identifying spatio-temporal patterns in video called Sequential Pattern Hyper Trees and we will demonstrate their application to a number of problems including lip-reading and Sign Language Recognition. The talk will also discuss the use of 3D information in recognition and ongoing work into action recognition in movies using linguistic information as weak annotation in the learning process.

[2] `http://vision.imar.ro/human3.6m/` and `http://vision.imar.ro/eyetracking/`

Stan Sclaroff, Boston University,
Linguistically-Motivated and Data-Driven: Improved Recognition of Handshapes in Videos of Sign Language
This talk presents a linguistically motivated approach for improved recognition of handshapes in videos of sign language. Handshape recognition in videos of sign language is challenging. Many handshapes share similar 3D configurations but are indistinguishable for some hand orientations in 2D image projections. Additionally, significant differences in handshape appearance are induced by the articulated structure of the hand and variants produced by different signers. Linguistic rules involved in the production of signs impose strong constraints on the articulations of the hands, yet little attention has been paid to exploiting these constraints in previous works on sign recognition. In our work, we propose a Handhsape Bayesian Network (HSBN) formulation that can model handshape co-occurrence constraints in the production of monomorphic lexical signs - the largest class of signs in American Sign Language (ASL). Our formulation is data-driven in the sense that the HSBN parameters, state space, and observation model are all learned from a large, annotated corpus of ASL videos. This corpus is being prepared for public dissemination: video for three thousand signs, each from up to six native signers of ASL, annotated with linguistic information such as glosses and morpho-phonological properties and variations, including the start/end handshapes associated with each ASL sign production. In our experiments, we demonstrate that leveraging linguistic constraints on handshapes results in improved handshape recognition accuracy. This research is the result of a collaboration involving linguists and computer scientists, including Carol Neidle (Boston U.), Ashwin Thangali (Boston U.), Joan Poole-Nash (Boston U.), Christian Vogler (Gallaudet), and Vassilis Athitsos (U. Texas at Arlington). The research has been supported through grants from the US National Science Foundation.

Acknowledgments

The Multi-modal Gesture Recognition Challenge and Workshop were organized thanks to the support of ChaLearn [3], the University of Barcelona Mathematics Faculty, the Universitat Autònoma de Barcelona, the Computer Vision Center, the Universitat Oberta de Catalunya, and the Human Pose Recovery and Behavior Analysis Group[4]. We thank the Kaggle submission website for wonderful support, together with the committee members and participants of the ICMI 2013 Multi-modal Gesture Recognition workshop for their support, reviews and contributions. This work has also been partially supported by Pascal2 network of excellence, and the Spanish projects TIN2009-14501-C02-02, TIN2012-39051, and TIN2012-38187-C03-02.

3. REFERENCES

[1] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopés, I. Guyon, V. Athitsos, and H. J. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ChaLearn Multi-Modal Gesture Recognition Grand Challenge and Workshop, 15th ACM International Conference on Multimodal Interaction*, 2013.

[3] ChaLearn: `http://chalearn.org`
[4] HuPBA research group: http://www.maia.ub.es/∼sergio/

Sergio Escalera obtained the P.h.D. degree on Multi-class visual categorization systems at Computer Vision Center, Universitat Autònoma de Barcelona. He obtained the 2008 best Thesis award on Computer Science at UAB. He leads the Human Pose Recovery and Behavior Analysis Group[4]. He is a lecturer of the Department of Applied Mathematics and Analysis, Univ. of Barcelona. He is a partial time professor at Univ. Oberta de Catalunya. He is a member of the Perceptual Computing Group and a consolidated research group of Catalonia. He is also a member of the Computer Vision Center at Campus UAB. He is Editor-in-Chief of American Journal of Intelligent Systems and editorial board member of more than 5 international journals. He is advisor of ChaLearn Challenges in Machine Learning. He is an active member of the Cluster de Salud Mental de Cataluña. He is also member of the AERFAI Spanish Association on Pattern Recognition and ACIA Catalan Association of Artificial Intelligence. He is co-founder of the PhysicalTech company. He has been program committee, organizing committee, session chair, and invited speaker of different conferences, including ICCV2011, AMDO2012, and CVPR2012. In 2014 he edited a Special Topic on Gesture Recognition at Journal of Machine Learning Research. His group won the 1st Prize of Pascal VOC Human Layout Challenge in 2010 and achieved the 3rd Prize of the ChaLearn Demonstration challenge 2012, sponsored by Microsoft, at ICPR 2012. He has more than 150 scientific publications and edited different books. His research interests include, between others, statistical pattern recognition, visual object recognition, and HCI systems, with special interest in human pose recovery and behavior analysis.

Jordi Gonzàlez received the Ph.D. in Computer Engineering in 2004 from Un. Autònoma de Barcelona (UAB). He is Associate Professor in Computer Science at the Computer Science Dept., UAB. He is also a research fellow at the Computer Vision Center, where he co-founded 2 spin-offs and the Image Sequence Evaluation (ISE Lab) research group. Prior to this he was a postdoctoral fellow at the Institut de Robòtica i Informàtica Industrial (IRI), a Joint Research Center of the Technical University of Catalonia (UPC) and the Spanish Council for Scientific Research (CSIC). His research interests lie on pattern recognition and machine learning techniques for the computational interpretation of human behaviours in image sequences, or Video Hermeneutics. He has co-organized the THEMIS (BMVC2008 and ICCV2009) and ARTEMIS workshops (ACMMM 2010, ECCV 2012, ACMMM 2013) related to the video-based analysis of human motion. He has served as Area Chair at ICPR2012; Publicity Chair at AVSS2012; Workshop Chair and Local Arrangement Chair at ICCV2011; and Tutorial Chair at ibPRIA2011. He has co-organized Special Issues in IJPRAI (2009), CVIU (2012) and MVA (2013) journals. He is member of the Editorial Board of CVIU and IET-CVI. He is also member of IEEE, Spanish Association on Pattern Recognition (AERFAI) and Catalan Association for Artificial Intelligence (ACIA).

Xavier Baró received his B.S. degree in Computer Science at the Universitat Autònoma de Barcelona (UAB) in 2003. In 2005 he obtained his M.S. degree in Computer Science at UAB, and in 2009 the Ph.D degree in Computer Engineering. At the present he is a lecturer and researcher at the IT, Multimedia and Telecommunications department at Universitat Oberta de Catalunya (UOC). He is involved on the teaching activities of the Computer Science, Telecommunication and Multimedia degrees of the UOC, and collaborates as professor assistant on the teaching activities of the Computer Science degree at the Applied Mathematics and Analysis of the Universitat de Barcelona (UB). In addition, he is involved on the Interuniversity master on Artificial Intelligence (UPC-UB-URV). He is co-founder of the Scene Understanding and Artificial Intelligence (SUNAI) group at the Internet Interdisciplinary Institute (IN3) of the UOC, and collaborates with the Computer Vision Center of the UAB, as member of the Human Pose Recovery and Behavior Analysis (HUPBA) group. His research interests are related to machine learning, evolutionary computation, and statistical pattern recognition, specially their applications to generic object recognition over huge cardinality image databases.

Miguel Reyes received his Bachelor degree in Computer Science at Universitat Autònoma de Barcelona (UAB) in 2010, and his master degree in Artificial Intelligence at Universitat Politècnica de Catalunya (UPC) in 2011. In October 2011 he joined the University of Barcelona, where he is currently Math Ph.D. student, within the area of Computer Science and Artificial Intelligence. He is an assistant professor at the Universitat of Barcelona. He is co-founder of the PhysicalTech company. He is member of the Computer Vision Center at UAB and member of the Human Pose Recovery and Behavior Analysis Group. His research interests include pattern recognition, signal processing and visual object recognition, and their application to health care systems.

Isabelle Guyon is an independent consultant, specialized in statistical data analysis, pattern recognition and machine learning. Her areas of expertise include computer vision and bioinformatics. Her recent interest is in applications of machine learning to the discovery of causal relationships. Prior to starting her consulting practice in 1996, Isabelle Guyon was a researcher at AT&T Bell Laboratories, where she pioneered applications of neural networks to pen computer interfaces and co-invented Support Vector Machines (SVM), a machine learning technique, which has become a textbook method. She is also the primary inventor of SVM-RFE, a variable selection technique based on SVM. The SVM-RFE paper has thousands of citations and is often used as a reference method against which new feature selection methods are benchmarked. She also authored a seminal paper on feature selection that received thousands of citations. She organized many challenges in Machine Learning over the past few years supported by the EU network Pascal2, NSF, and DARPA, with prizes sponsored by Microsoft, Google, and Texas Instrument. Isabelle Guyon holds a Ph.D. degree in Physical Sciences of the University Pierre and Marie Curie, Paris, France. She is president of ChaLearn, a non-profit dedicated to organizing challenges, vice-president of the Unipen foundation, adjunct professor at New-York University, action editor of the Journal of Machine Learning Research, and editor of the Challenges in Machine Learning book series of Microtome.

Vassilis Athitsos received the BS degree in mathematics from the University of Chicago in 1995, the MS degree in computer science from the University of Chicago in 1997, and the PhD degree in computer science from Boston University in 2006. In 2005-2006 he worked as a researcher at Siemens Corporate Research, developing methods for data base guided medical image analysis. In 2006-2007 he was a postdoctoral research associate at the Computer Science department at Boston University. In August 2007 he joined the Computer Science and Engineering department at the University of Texas at Arlington, where he currently serves as associate professor. His research interests include computer vision, machine learning, and data mining. His recent work has focused on gesture and sign language recognition, detection and tracking of humans using computer vision, efficient similarity-based retrieval in multimedia databases, shape modeling and detection, and medical image analysis. His research has been supported by the National Science Foundation, including an NSF CAREER award.

Hugo Jair Escalante obtained his degree of PhD in computer science from the Instituto Nacional de Astrofisica, Optica y Electronica (INAOE) at Mexico, where he is now associate researcher. Dr. Escalante is member of the Mexican System of Researchers (SNI) since 2011, and director of ChaLearn, the Challenges in Machine Learning organization (2011-2013). Hugo Escalante obtained the 2010 Best PhD thesis on AI award from the Mexican Society on artificial intelligence (SMIA). His main research interests are on pattern recognition, machine learning and computational intelligence with applications in text mining and high-level computer vision.

Leonid Sigal is a Research Scientist at Disney Research Pittsburgh and an adjunct faculty at Carnegie Mellon University. Prior to this he was a postdoctoral fellow in the Department of Computer Science at University of Toronto. He completed his Ph.D. at Brown University in 2008; he received his B.Sc. degrees in Computer Science and Mathematics from Boston University (1999), his M.A. from Boston University (1999), and his M.S. from Brown Univ. (2003). From 1999 to 2001, he worked as a senior vision engineer at Cognex Corporation, where he developed industrial vi-

sion applications for pattern analysis and verification. His work received best paper awards at Articulate Motion and Deformable Object Conference in 2006 and 2012. He is co-editor of Springer Verlag book on "Visual Analysis of Humans: Looking at People". Leonid's research interests mainly lie in the areas of computer vision, machine learning, and computer graphics, but also borderline fields of psychology and humanoid robotics. His current research spans articulated pose estimation, action recognition, domain adaptation, latent variable models, data-driven simulation, controller design for animated characters and perception of human motion.

Antonis A. Argyros is an Associate Professor at the Computer Science Department, University of Crete and a researcher at the Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas (FORTH) in Heraklion, Crete, Greece. He has been a postdoctoral fellow at the Computational Vision and Active Perception Laboratory (CVAP) at the Royal Institute of Technology in Stockholm, Sweden. Since 1999, as a member of the Computational Vision and Robotics Laboratory (CVRL) of FORTH-ICS, he has been involved in many RTD projects in computer vision, image analysis and robotics. Antonis Argyros is an area editor for the CVIU Journal, member of the Editorial Board of the IET Image Processing Journal and one of the general chairs of the 11th European Conference in Computer Vision (ECCV'2010, Heraklion, Crete). The research interests of Antonis fall in the areas of computer vision with emphasis on tracking, human gesture and posture recognition, 3D reconstruction and omnidirectional vision. He is also interested in applications of computational vision in the fields of robotics and smart environments.

Cristian Sminchisescu is a Professor in the Department of Mathematics, Faculty of Engineering, at Lund University. He has obtained a doctorate in computer science and applied mathematics with emphasis on imagining, vision and robotics at INRIA, France, under an Eiffel excellence doctoral fellowship, and has done postdoctoral research in the AI Laboratory at the Univ. of Toronto. He holds a Professor equivalent scientific title at the Romanian Academy and a Professor status appointment at Toronto, and conducts research at both institutions. During 2004-07, he has been a Faculty member at the Toyota Tech. Institute, a philanthropically endowed computer science institute located at the Univ. of Chicago, and during 2007-2012 on the Faculty of the Institute for Numerical Simulation in the Mathematics Depart. at Bonn Univ. Cristian Sminchisescu is a member in the program committees of the main conferences in CV and machine learning (CVPR, ICCV, ECCV, NIPS), an Area Chair for ICCV 2007-13, and a member of the Editorial Board (Associate Editor) of IEEE Transactions for Pattern Analysis and Machine Intelligence (PAMI). He has offered tutorials on 3d tracking, recognition and optimization at ICCV and CVPR, the Chicago Machine Learning Summer School, the AEFRAI Vision School in Barcelona, and the CV summer school at ETH in Zurich. Over time, his work has been funded by the United States National Science Foundation, the Romanian Science Foundation, the German Science Foundation, and the European Commission. Cristian Sminchisescu's research goal is to train computers to 'see' and interact with the world seamlessly, as humans do. His research interests are in the area of CV (articulated objects, 3d reconstruction, segmentation, and object and action recognition) and machine learning (optimization, structured prediction and kernel methods). Recent work in his group was the winner of the PASCAL VOC object segmentation and labeling challenge in four editions, 2009 - 2012. Three recent datasets constructed in the group, and available online, one containing 3.6 million 3D human poses (Human3.6M), and the other two containing millions of human fixations obtained under the constraint of action recognition tasks (Actions in the Eye), are the largest ever collected in those domains.

Richard Bowden leads the Cognitive Vision Group within the Centre for Vision Speech and Signal Processing at the Univ. of Surrey. His research centres on the use of computer vision to locate, track & understand humans with contributions across a range of topics including surveillance, sign and gesture recognition, lip-reading, facial expression recognition, cognitive robotics & tracking. His recent interests are in weakly supervised learning and video mining to automatically learn from large quantities of data. He has published over 130 papers, held over 20 research grants worth in excess of č5M and supervised over fifteen PhD students. His research has been recognised by prizes, plenary talks & media/press coverage including the Sullivan thesis prize in 2000 and best paper awards. He was a visiting research fellow at the Univ. of Oxford 2001-2004 working with Zisserman and Brady and was awarded a Royal Society Leverhulme Trust Senior Research Fellowship in 2013. He is a member of the BMVA, a senior member of IEEE and a fellow of the Higher Education Academy, UK.

Stan Sclaroff founded the Image and Video Computing group at Boston University. He received the PhD degree from MIT in 1995. He joined the faculty at Boston University in 1995, where he is now a Professor in the Department of Computer Science. He has coauthored numerous scholarly publications in the areas of tracking, video-based analysis of human motion and gesture, surveillance, deformable shape matching and recognition, as well as image/video database indexing, retrieval and data mining methods. Professor Sclaroff has received an ONR Young Investigator Award and an NSF Faculty Early Career Development Award. He has served on the program committees of over 80 computer vision conferences and workshops. He has served as an Associate Editor for IEEE Transactions on Pattern Analysis (T-PAMI), 2000-2004, and 2006-2011, and he currently serves as a T-PAMI Associate Editor in Chief. He is a Senior Member of the IEEE.