International Journal of Computer Vision (2018) 126:141–143
https://doi.org/10.1007/s11263-017-1058-y

Looking at People Special Issue

Sergio Escalera[1,2,3] · Jordi Gonzàlez[2,4] · Hugo Jair Escalante[3,5] · Xavier Baró[2,6] · Isabelle Guyon[3,7]

Published online: 31 January 2018

1 Introduction

The automatic analysis of humans in static images and video sequences, known as *Looking at People*, keeps making rapid progress with the regular improvement of existing methods and the proposal of new paradigms that constantly push the state-of-the-art. Applications are countless, including human computer interaction, affective computing, human robot interaction, communication, entertainment, security, commerce and sports, while having an important social impact in assistive technologies for the handicapped and the elderly. Because of this overwhelming spectrum of development and applications, we edited this special issue that aims at covering all aspects of Looking at People.

This compendium is a natural follow up of a series of events organized on this topic by ChaLearn Looking at People,[1] which focuses on challenge organization in the fields of computer vision and machine learning (Baró et al. 2015; Escalera et al. in "ChaLearn Looking at People: A Review of Events and Resources", https://doi.org/10.1109/IJCNN.2017.7966041; Escalera et al. 2017). During the last 7 years ChaLearn has witnessed an important evolution in LAP methodologies, in essence thanks to the use of depth sensors, the collection and annotation of large databases, the empowerment of deep neural networks, and their huge success in multiple fields like social science. Consequently, the characterization of human behavior in image sequences has become one of the most important research topics in areas like affective computing, human-machine interfaces, gaming, security, marketing, and e-health.

This special issue is composed by outstanding articles that are representative of recent developments in the LAP field. These manuscripts were not selected solely based on their reported performance or their participation in Chalearn workshops, but rather on their overall quality and significance to the field. Consequently, all the 28 submitted manuscripts went through the regular IJCV reviewing process. So we would like to thank all the referees for their careful and critical role and thank the authors for participating in the lengthy review process. The final list of 16 accepted articles can be grouped in three main categoties: Looking at Faces, Looking at Actions, and Looking at Gestures. This compilation provides a snapshot of the state of the art in the Looking at People domain. New data sets were released and extended versions of others have been made available by authors. Data is critical for the success of certain models in LAP tasks (e.g., deep-learning based techniques) and for the establishment of evaluation protocols. As such, contributed data sets in the included manuscripts will be decisive to the progress of LAP in the next few years. Regarding applications, articles approached virtually all aspects of LAP: from head and face analysis, to pose estimation, gesture and action recognition. Interestingly, deep learning did not rule as modeling framework, in fact, treebased learners, manifold learning and rank minimization shared credits with deep learning as modeling framework for LAP. In contrast, most of the recent winner solutions from ChaLearn LAP participants are based on deep learning strategies. This may be because of the nature of the problems addressed in some ChaLearn LAP competitions: classification of large datasets related to face and behavior analysis in semi-controlled environments (see Footnote 1). Multimodal information, including that captured with depth sensors was also effectively exploited by some authors of different published articles within this special issue. In addition, novel tasks were also presented and studied, including prediction of interactee and manipulation actions. We foresee that articles in this compendium will play a key role

[1] http://chalearnlap.cvc.uab.es/.

✉ Sergio Escalera
sergio.escalera.guerrero@gmail.com

1 University of Barcelona, Barcelona, Spain
2 Computer Vision Center, Catalonia, Spain
3 ChaLearn, Berkeley, CA, USA
4 Universitat Autònoma de Barcelona, Barcelona, Spain
5 INAOE, Puebla, Mexico
6 Universitat Oberta de Catalunya, Catalonia, Spain
7 Université Paris Saclay, Paris, France

in the progress of LAP in Computer Vision in years to come.

2 Looking at Faces

Seven papers were published in relation to face analysis within this special issue. They address major topics in face analysis, including face detection, alignment, tracking, pose extimation, face recognition, age estimation, and expression recognition, both considering RGB and 3D data.

In "Deep expectation of real and apparent age from a single image without facial landmarks", https://doi.org/10.1007/s11263-016-0940-3 propose a solution for real and apparent age estimation. The authors present a Deep EXpectation (DEX) formulation which builds upon a robust face alignment, the VGG-16 deep architecture and a classification followed by a expected value formulation of the age estimation problem. Another contribution is IMDB-WIKI, the largest public face images dataset to date with age and gender annotations.

In "Real-time Accurate 3D Head Tracking and Pose Estimation with Consumer RGB-D Cameras", https://doi.org/10.1007/s11263-017-0988-8 propose a head pose estimation algorithm from RGB-D video: the RGB image is used to detect the face and initialize a tracker, while the depth image is used to temporally track the head pose throughout the sequence using a random forest algorithm. The authors also propose a multi-camera system, which deals with the data acquired simultaneously from multiple RGB-D sensors.

In "Toward Personalized Modeling: Incremental and Ensemble Alignment for Sequential Faces in the Wild", https://doi.org/10.1007/s11263-017-0996-8 propose a novel approach for face alignment in unconstrained videos. Their approach incorporates motion models to perform ensemble initialization, which can effectively overcome the initialization-sensitivity issue of many existing alternatives. Among their main contributions, ensemble alignment is performed within a single frame, a rank minimization framework with group-sparse regularization is designed, and a novel drifting evaluation network is proposed to significantly alleviated the model drifting issue.

Chrysos et al. in "A Comprehensive Performance Evaluation of Deformable Face Tracking *In-the-Wild*" https://doi.org/10.1007/s11263-017-0999-5 present a comprehensive evaluation of multiple deformable face tracking pipelines. The authors show how current face detection and model free tracking technologies are advanced enough so that even a naive combination with landmark localization techniques is adequate to achieve state-of-the-art performance on deformable face tracking under arbitrary conditions.

In "Large scale 3D Morphable Models", https://doi.org/10.1007/s11263-017-1009-7 present a large scale facial model (LSFM) built from 9663 distinct facial identities, thus containing statistical information from a huge variety of the human population. The LSFM software pipeline and the 3D Morphable Models (3DMM) were made available. The authors explore the structure of high dimensional facial manifolds, revealing how age and ethnicity variations are clustered, and used for age prediction.

In "Confidence-Weighted Local Expression Predictions for Occlusion Handling in Expression Recognition and Action Unit detection", https://doi.org/10.1007/s11263-017-1010-1 propose to train random forests upon spatially-constrained random local subspaces of the face. The local predictions form a categorical expression-driven high-level representation called local expression predictions (LEPs). LEPs are combined to describe categorical facial expressions, as well as action units, by modeling the manifold around specific facial feature points using a hierarchical autoencoder network.

Lastly, the paper entitled "Unconstrained Still/Video-Based Face Verification with Deep Convolutional Neural Networks" by https://doi.org/10.1007/s11263-017-1029-3 presents the design details of a deep learning system for unconstrained face recognition, including modules for face detection, association, alignment and face verification. The system integrates DCNN-based face detection and fiducial point detection taking faces as input with different resolutions, while employing an efficient metric learning method which optimizes the embedding matrix, as opposed to norm-based constraints used in other methods.

3 Looking at Actions

Seven papers were published in relation to action recognition within this special issue. They address major topics in action recognition, including posebehavior analysis, action detection/localization and classification, human-object and human-human action recognition, and action recognition in uncontrolled and complex environments.

Chen and Grauman in "Subjects and Their Objects: Localizing Interactees for a Person-Centric View of Importance" https://doi.org/10.1007/s11263-016-0958-6 propose to predict the "*inter- actee*" in novel images, i.e. to localize the object of a persons action. So given an arbitrary image with a detected person, the goal is to produce a saliency map indicating the most likely positions and scales where that persons "*interactee*" would be found. The authors also introduce a new 10,147-image dataset of interaction annotations for all person images in COCO.

In "Space–Time Tree Ensemble for Action Recognition and Localization", https://doi.org/10.1007/s11263-016-0980-8 present an approach to discover compact sets of hierarchical spacetime tree structures of human actions in video.

Using an ensemble of the discovered trees, or in combination with simpler action words and pairwise structures, the authors build action classifiers that achieve promising results on three challenging datasets: HighFive, UCF-Sports and Hollywood3D.

Georgakis et al. in "Dynamic Behavior Analysis via Structured Rank Minimization" https://doi.org/10.1007/s11263-016-0985-3 describe a framework for dynamic behavior analysis in real-world conditions. In essence, the framework is based on a novel structured rank minimization method to learn low-complexity models from time-varying data, in the presence of gross sparse noise and possibly missing data. Based on the e_cient Alternating-Directions Method of Multipliers (ADMM) algorithm, an structured rank minimization model along with a scalable version is proposed.

Fermüller et al. in "Prediction of Manipulation Actions" https://doi.org/10.1007/s11263-017-0992-z present an approach to action interpretation, which treats the problem as a continuous updating of beliefs and predictions. Their contribution is applied to two tasks: the prediction of perceived action from visual input, and the prediction of force values on the hand. In addition, new datasets of videos of dexterous actions and force measurements were created to be publicly shared.

In "Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos", https://doi.org/10.1007/s11263-017-1013-y extend the existing THUMOS dataset by introducing MultiTHUMOS, a new dataset of dense labels over unconstrained internet videos. The authors show that modeling multiple, dense labels benefits from temporal relations within and across classes. In addition, authors define a novel variant of long short-term memory deep networks for modeling temporal relations.

Wang et al. in "Transferring Deep Object and Scene Representations for Event Recognition in Still Images" https://doi.org/10.1007/s11263-017-1043-5 address the problem of image-based event recognition by presenting an architecture for event recognition in still images, which transfers deep representations from object and scene models to the event recognition task. Three transfer techniques are used, which turn out to be effective in reducing the effect of over-fitting and improving the generalization ability of the learned CNNs.

In "Joint Estimation of Human Pose and Conversational Groups from Social Scenes", https://doi.org/10.1007/s11263-017-1026-6 present a novel framework to jointly estimate the head and body pose of targets, and geometric formations involving interacting targets known as F-formations from social scenes. Their approach is based on a weakly-supervised learning algorithm for joint inference of head and body orientations, showing an increased efficacy over the state-of-the-art.

4 Looking at Gestures

Finally, two papers were published in relation to gesture recognition within this special issue. They address major topics in face analysis, including deep temporal analysis of gestures and gesture recognition from multimodal data sources.

In "Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video", https://doi.org/10.1007/s11263-016-0957-7 show that adding bidirectional recurrence and temporal convolutions improves frame-wise gesture recognition in video significantly. Adding temporal convolutions in all layers of an architecture has a notable impact on the performance, as they are able to learn hierarchies of motion features, unlike RNNs.

In "Deep Multimodal Fusion: A Hybrid Approach", https://doi.org/10.1007/s11263-017-0997-7 present a hybrid model comprising of temporal generative and discriminative models for classifying sequential data from multiple heterogeneous modalities. The model is based on Conditional RBMs (CRBMs), an extension of the RBM model, that takes into account short term temporal phenomena. So using a CRBM-based generative model enables modeling short-term multimodal data and allows to deal with missing data by generating it within or across modalities. A discriminative component is included in the model to define the Discriminative CRBMs or DCRBMs. Finally, Multimodal Discriminative CRBMs (MMDCRBMs) are proposed, which combines a collection of unimodal DCRBMs, one for each visible modality.

References

Baró, X., Gonzàlez, J., Fabian, J., Bautista, M. A., Oliu, M., Jair, E., et al. (2015). ChaLearn Looking at People 2015 challenges: Action spotting and cultural event recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2015.

Escalera, S., Guyon, I., & Athitsos, V. (Eds.) (2017). *Gesture recognition. Springer Series on Challenges in Machine Learning* (Vol. 2).