EDITORIAL

Semantic Understanding of Human Behaviors in Image Sequences: From Video-Surveillance to Video-Hermeneutics

Jordi Gonzàlez

Department of Computer Science and Computer Vision Center

Universitat Autònoma de Barcelona

08193 Bellaterra (Barcelona), Catalonia, Spain

Jordi.Gonzalez@uab.cat

Thomas B. Moeslund

Department of Media Technology

Aalborg University

Niels Jernes Vej 14, 9220 Aalborg East, Denmark

tbm@create.aau.dk

Liang Wang

National Laboratory of Pattern Recognition (NLPR)

Institute of Automation, Chinese Academy of Sciences

95 Zhongguancun East Road, Haidian District, Beijing 100190, P.R.China

wangliang@nlpr.ia.ac.cn

Hermeneutics, defined as the art of interpreting a message, centered its study for many centuries in theorizing the interpretation process in written texts, mainly biblical ones. The appearance of the first recordings of image sequences during the last period of the 19th century expanded such a domain. The aesthetics of such videos was of interest in the early 20th century to great philosophers such as Ludwig Wittgenstein since, in that new format of communication, cinematographic texts became an interpretation game in which the language film was articulated in a network of multiple readings.

Hermeneutics in image sequences, or as we call it, Video-Hermeneutics (VH) involves explaining the subjective and social value of the human behavior observed in image sequences and, in general, of all multimedia content. So VH is not only related to what happens in a video, but also to understand the meaning of what is being described, i.e. what message is being transmitted to us as human observers.

This analysis of the human behavior in image sequences has also been modeled in computational terms within the field of Computer Science. This has been possible thanks to software and hardware advancements, particularly the reduction of camera costs, which entailed the expansion of their use for video-surveillance and thus the necessity of analyzing automatically and in real-time human behaviors observed through millions of cameras.

These factors have also created important technical-scientific contributions in the areas of Computer Vision and Artificial Intelligence. A collection of the most recent works in these fields is now published in this Special Issue of the *Computer Vision and Image Understanding* journal. Taking this opportunity we would like to thank all the authors who submitted their papers for their efforts and interest. The guest

editors would also like to deliver their sincere thanks to the reviewers for their careful and valuable comments on the submitted manuscripts and for their detailed and helpful suggestions about improving the manuscripts. This Special Issue would not have been possible otherwise. We are especially very grateful to the Editor-in-Chief Prof. Avi Kak and to the Elsevier Journal Manager Linda Shapiro for their support and editing guidance during the preparation and publication of this Special Issue.

In general terms, it can be stated that four levels of complexity in artificial vision systems have been the milestones achieved in the last two decades, which we identify as *movement*, *action*, *activity* and *behavior* descriptions. These four levels indicate the increasing in-depth analysis of the semantics of each image sequence by reducing uncertainty and errors in interpretation processes while improving the semantic utility of the explanation reported for those observed behaviors. This is the reason why image understanding is given more and more importance to the particular scenario where motion is detected, to the most semantically relevant scene regions, such as sidewalks or roads, and to those objects with which detected agents are interacting, such as bags, cars, etc.

The first semantic level, basic and critical at the same time, is the detection and tracking of (i) *movements*. Without movement detection, reasoning in video surveillance is highly restricted; but this basic task is plenty of problems: saturation, shadows, camouflage, background-in-motion, *ghosts*, etc. Examples of good contributions based on low-level movement analysis are found at the beginning of this Special Issue: Krausz and Bauckhage propose a method that learns those movement patterns from crowd analysis which best characterize disaster simulations, like stampedes; Bertini et al. present a novel unsupervised, real-time approach for anomaly detection in motion trajectories based on scene dynamic statistics together with appearance; and Daubney et al. propose a method for pose estimation based on Pictorial Structures using only motion.

The next level in semantics corresponds to the recognition of the (ii) *actions* performed by the detected agents, for example walking, bending over or jumping forward. In this Special Issue there are several contributions addressing this task, like Iosifidis et al. for action modeling in multi-view dyneme spaces; Liu et al. for action class representation based on Bag-of-Words and computed by clustering Pointwise Mutual Information features; Thi et al. for action retrieval and classification based on a wise combination of discriminative and generative action elements; and Holte et al. who present a novel approach for computing Spatio-Temporal Interest Points based on a surround suppression strategy together with local and temporal constraints.

The third level involves modeling the (iii) *activities*, defined as actions plus the analysis of their interactions with and/or reactions to other objects/agents around. Activity descriptions can establish, for example, that if the trajectories of a walking pedestrian and a car are approaching too fast, there might a danger of run over. In this context, excellent contributions can be found in this Special Issue: Drosou et al. present a novel biometric identification technique applied to humans interacting with office objects based on static anthropometric profiles and dynamic motion trajectories; Kosmopoulos et al. analyze the interactions of human operators with objects located in industrial plants based on a novel online re-adjustment framework which incorporates user's feedback; and Tran et al. propose a Hidden-Markov Model based system which analyzes the interactions of drivers' feet with brake and acceleration car pedals.

Finally, the level with more contextual load corresponds to (iv) *behaviors*, that situate activities in a specific scene and reacts according to inferred interpretations. Two papers addressing behavior analysis conclude this Special Issue: Dee et al. present an interesting unsupervised approach for building semantic scene models, like junctions, roads, and open squares, characterized by dominant patterns of movements clustered based on both proximity and local motion similarity; and finally Bellotto et al. describe a complete cognitive visual system which, based on the high-level interpretation of tracking data, is able to generate the most appropriate commands for semantically controlling pan-tilt-zoom camera.s

More than a century has passed since pioneers such as Muybridge and Marey made the first recordings of human movements with VH purposes. At that time, one of the most demanded applications was to establish the optimal distribution of the armament carried by soldiers for reducing their fatigue during long journeys. Nowadays, long term research will entail other promising (and civil) applications: one example would be the design of automatic video annotation strategies by applying more informative and efficient indexing algorithms of huge multimedia archives, by refining in an unsupervised way the results of search engines and, definitely, by finding a requested, semantically complex, visual content within an exponentially rising volume of image data.