Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2008, Article ID 676094, 3 pages
doi:10.1155/2008/676094

Editorial
Anthropocentric Video Analysis: Tools and Applications

Nikos Nikolaidis,[1,2] Maja Pantic,[3,4] and Ioannis Pitas[1,2]

[1] *Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece*
[2] *Informatics and Telematics Institute, CERTH, 57001 Thermi-Thessaloniki, Greece*
[3] *Department of Computing, Imperial College London, London SW7 2AZ, UK*
[4] *Department of Computer Science, University of Twente, 7522 NB Enschede, The Netherlands*

Correspondence should be addressed to Nikos Nikolaidis, nikolaid@aiia.csd.auth.gr

Received 23 April 2008; Accepted 23 April 2008

During the last two decades, we have witnessed an increasing research interest towards what one could call *anthropocentric video analysis*, namely, algorithms that aim to extract, describe, and organize information regarding the basic element of most videos: humans. This diverse group of algorithms processes videos from various sources (movies, home videos, TV programmes, surveillance videos, etc.) and extracts a wealth of useful information. A large cluster of algorithms target information related to the state or state transitions of individuals: presence and position/posture through face or body detection, body or body parts tracking and posture estimation; identity by means of face recognition/verification, full-body recognition, gait analysis, and so forth; emotional state through facial expression, body gesture, and/or posture analysis; performed actions or activities; and behavior through spatio-temporal analysis of various behavioral cues including facial/head/hand/body gestures and postures. Another smaller group of techniques focuses on detecting or recognizing interactions or communication modes by means of visual speech recognition, dialogue detection, social signals recognition such as head nods and gaze exchanges, recognition of activities or events in multiple-person environments (e.g., event analysis in sport videos or crowd-scene analysis, etc.). Finally, a number of techniques aim at deriving information regarding physical characteristics of humans, mainly in the form of 3D head or full-body models.

The interest of the scientific community for anthropocentric video analysis stems from the fact that the extracted information can be utilised in various important applications. First of all, it can be used to devise intuitive and natural paradigms of man-machine interaction, for example, through gesture-based interfaces, visual (or audio-visual) speech recognition, interfaces that understand and adapt to the emotional state of users, and interfaces between virtual characters and human users, which are governed by the same social rules as the human-human interaction. In the same wavelength, but in a considerably broader scope, anthropocentric video analysis techniques are some of the enabling technologies for the so-called ubiquitous computing trend (also known as pervasive computing or ambient intelligence) where a large number of small (or embedded), interconnected, and clever computing devices and sensors cooperate to assist people in their everyday life in an unobtrusive and natural way. An intelligent living space, that controls lighting, music, temperature, and home appliances according to the inhabitants' mood, location, habits, and behavioral patterns indicating their intention, is frequently used as an example of this trend. Moreover, techniques like person detection, tracking, recognition or verification, and activity recognition are already being integrated in smart surveillance systems, access-control systems, and other security systems capable of detecting access permission violations, abnormal behaviors, or potentially dangerous situations. In addition, data derived from anthropocentric video analysis techniques can be used to infer human-related semantic information for videos, to be utilised in video annotation, retrieval, indexing, browsing, summarisation, genre classification, and similar tasks. Highlights detection in sport videos, automatic generation of visual movie summaries, and content-based retrieval in video databases are only some of the applications in this category that can benefit from human-centric analysis of video. Finally, such algorithms are indispensable building blocks for a number

of other applications that include automatic diagnosis of neuromuscular and orthopaedic disorders, performance analysis of athletes, intelligent/immersive videoconferencing, automated creation of 3D models for animated movies, users' avatar animation in virtual environments and games, and so forth.

The papers that have been selected for publication in this special issue present interesting new ideas in a number of anthropocentric video analysis topics. Although not all areas mentioned above are represented, we do hope that the issue will give readers the opportunity to sample some state-of-the-art approaches and appreciate the diverse methodologies, research directions, and challenges in this hot and extremely broad field.

Most papers in this issue address either the problem of person detection and tracking or the problem of human body posture estimation.

In "Detection and tracking of humans and faces," by S. Karlsson et al., a framework for multi-object detection and tracking is proposed, and its performance is demonstrated on videos of people and faces. The proposed framework integrates a prior knowledge of object categories (in the form of a trained object detector) with a probabilistic tracking scheme. The authors experimentally show that the proposed integration of detection and tracking steps improves the state estimation of the tracked targets.

In "Integrated detection, tracking, and recognition of faces with omni video array in intelligent environments," by K. S. Huang and M. Trivedi, robust algorithms are proposed for face detection, tracking, and recognition in videos obtained by an omnidirectional camera. Skin tone detection and face contour ellipse detection are used for the face detection, a view-based face classification is applied to reject the nonface candidates, and Kalman filtering is applied for face tracking. For face recognition, the best results have been obtained by a continuous hidden Markov model-based method, where accumulation of matching scores along the video boosts the accuracy of face recognition.

In "Monocular 3D tracking of articulated human motion in silhouette and pose manifolds," F. Guo and G. Qian propose a system that is capable of tracking the human body in 3D from a single camera. The authors construct low-dimensional human body silhouette and pose manifolds, establish appropriate mappings between these two manifolds through training, and perform particle filter tracking over the pose manifold.

The paper "Multi-view-based cooperative tracking of multiple human objects" by C.-L. Huang and K.-C. Lien presents a multiple person tracking approach that utilises information from multiple cameras in order to achieve efficient occlusion handling. The idea is that the tracking of a certain target in a view where this target is fully visible can assist the tracking of the same target in a view where occlusion occurs. Particle filters are employed for tracking, whereas two hidden Markov processes are employed to represent the tracking and occlusion status of each target in each view.

The paper "3D shape-encoded particle filter for object tracking and its application to human body tracking" by H. Moon and R. Chellappa proposes a method for tracking and estimating object motion by using particle propagation and the 3D model of the object. The measurement update is carried out by particle branching according to weights computed by shape-encoded filtering. This shape filter has the overall form of the predicted projection of the 3D model, where the 3D model is designed to emphasise the changes in 2D object shape due to motion. Time update is handled by minimising the prediction error and by adaptively adjusting the amount of random diffusion. The authors experimentally show that the method is able to effectively and efficiently track walking humans in real-life videos.

In their paper entitled "Human posture tracking and classification through stereo vision and 3D model matching," S. Pellegrini and L. Iochhi present a method for human body posture recognition and classification from data acquired from a stereo camera. A tracking algorithm operating on these data provides 3D information regarding the tracked body. The proposed method uses a variant of ICP to fit a simplified 3D human body model and then tracks characteristic points on this model using Kalman filtering. Subsequently, body postures are classified through a hidden Markov model to a limited number of basic postures.

The seventh paper, "Compression of human motion animation using the reduction of inter-joint correlation" by S. Li et al. is closely related to the papers outlined above since it deals with the important issue of compressing human body motion data derived either through video-based motion tracking or motion capture equipment (e.g., magnetic trackers). Two different approaches for the compression of such data, represented as joint angles in a hierarchical structure, are proposed. The first method combines the wavelet transform with forward kinematics and allows for progressive decoding. The second method, which provides better results, is based on prediction and inverse kinematics.

The following two papers deal with human activity recognition. An algorithm based on motion and color information, is presented by A. Briassouli et al. in "Combination of accumulated motion and color segmentation for human activity analysis." The algorithm accumulates optical flow estimates and processes their higher-order statistics in order to extract areas of activity. MPEG-7 descriptors extracted for the activity area contours are used for comparing subsequences and detecting or analysing the depicted actions. This information is complemented by mean shift colour segmentation of the moving and static areas of the video, that provides information about the scene where the activity occurs and also leads to accurate object segmentation.

The paper "Activity representation using 3D shape models," by M. Abdelkader et al., presents a method for human activity representation and recognition that is based on 3D shapes generated by the target activity. Motion trajectories of points extracted from objects (e.g., human body parts) involved in the activity are used to build these 3D shape models for each activity, which are subsequently used for classification and detection of either target or unusual activities.

Finally, each of the last two papers in this special issue deal with a different problem. The paper "Comparison of

image transform based features for visual speech recognition in clean and corrupted video" authored by R. Seymour et al. deals with the important problem of visual speech recognition. More specifically, the paper studies and compares the performance of a number of transform-based features (including novel features extracted using the discrete curvelet transform) as well as feature set selection methods for visual speech recognition of isolated digits. Both clean video data and data corrupted by compression, blurring and jitter are used to assess the features' robustness to noise.

On the other hand, the paper "Athropocentric video segmentation for lecture webcasts" by G. Friedland and R. Rojas describes an interesting application of person detection and segmentation. The challenge addressed is that of recording and transmission of lectures in high quality and in a bandwidth-efficient way. An electronic whiteboard is used to record in vector format the handwritten content of the board whereas the lecturer is segmented in real time from the background by constructing, through a clustering approach, a colour signature for the background and by suppressing the changes introduced to the background due to the lecturer's handwriting. The segmented lecturer is then pasted semitransparently on the whiteboard content, and the synthesised sequence is played back or transmitted as MPEG-4 video.

ACKNOWLEDGMENTS

The guest editors of this issue wish to thank the reviewers who have volunteered their time to provide valuable feedback to the authors. They would also like to express their gratitude to the contributors for making this issue an important asset to the existing body of literature in the field. Many thanks to the editorial support of the EURASIP Journal on Image and Video Processing for their help during the preparation of this issue.

Nikos Nikolaidis,
Maja Pantic
Ioannis Pitas