Proc. of the 2nd International Workshop on Visual Surveillance, Fort Collins, USA, 1999

The analysis of human motion and its application for visual surveillance

D.M. Gavrila
DaimlerChrysler Research and Technology
Wilhelm Runge St. 11, 89081 Ulm, Germany
dariu.gavrila@DaimlerChrysler.com

Abstract

"Looking at People" is currently one of the most active application area in computer vision. This contribution provides a short overview of existing work on human motion as far as whole-body motion and gestures are concerned. The overview is based on a more extensive survey article [10]; here, the emphasis lies on surveillance scenarios.

1 "Smart" surveillance systems

The "Looking at People" area involves building systems capable of interacting intelligently and effortlessly with a human-inhabited environment. A number of promising application domains can be identified: virtual reality, advanced user interfaces, motion analysis, model-based coding and finally "smart" surveillance systems, our focus here. Traditional vision-based surveillance systems have a human operator continuously monitor a wall of CCTV screens for specific events. Not only is this a quite tedious activity, but with increased demand for area coverage, this mode of operation quickly becomes unfeasible due to information overload. Systems are needed to automatically filter out spurious information and present the operator only those parts of data which are likely security relevant. Current commercial vision systems take a step in this direction by offering rudimentary capabilities for motion-detection, based on image differencing and background subtraction. However, their use is restricted to areas off-limits to people, and their performance is often plagued by false alarms due to external environmental effects (e.g. wind blowing, lighting changes, animals wandering around).

"Smart" surveillance systems incorporate specific knowledge about human shape and appearance to decrease false alarms; they might even be able to distinguish between simple authorized and non-authorized activities using methods referenced in the next Section. In case of wide area coverage, they provide an integrated functionality across multiple sensors. For areas covered by multiple sensors, this might involve sensor fusion to increase reliability. In addition, an automatic tracking capability across different areas can be provided. This is important, because in many applications it is the temporal succession of events, which determines whether a particular activity is to be classified as security relevant or not. Furthermore, in case an alarm is triggered, the originating object is focussed upon and tracked while countermeasures are taken; this involves active sensor control (e.g. [1]). While short response time is generally the norm for surveillance systems, in some applications sensor data is simply stored; here, relevant information must be retrieved by allowing the user to formulate queries; this assumes some means for scene description. Finally, a desirable feature is remote-access; one might want to transmit relevant surveillance information to a users' personal digital assistant or place it directly on the internet for remote inspection.

Here are a few scenarios these "smart" surveillance systems might handle:

- access control for buildings and other assets
- surveillance of parking lots, vending machines and ATMs
- detection of pedestrians for driver assistance and/or vehicle control
- battlefield awareness

Among the major surveillance projects currently in progress are the VSAM project on battlefield awareness in the U.S. and the IMPROOFS project on forensics in Europe. For more information, visit the WWW pages http://www.cs.cmu/~vsam/ and http://www.esat.kuleuven.ac.be/mi2/visics.html respectively.

2 Methods

Previous work on gestures and whole-body movement can be broadly classified into the following three groups [10]:

- 2-D approaches without explicit shape models
- 2-D approaches with explicit shape models
- 3-D approaches

The first approach bypasses a pose recovery step altogether and describes human movement in terms of simple low-level, 2-D features from a region of interest. Polana and Nelson [22] refer to "getting your man without finding his body parts". Models for human action are described in statistical terms based on low-level features. Foreground regions are typically obtained by skin-color detection or background subtraction from which features based on shape [2] [6] [12], texture [7] [8] [21], or motion [4] [9] [22] are extracted. In some cases [12] [21], the requirement of a separate foreground segmentation is relaxed by the employment of window search procedures.

The second approach uses explicit a priori knowledge of how the human body (or hand) appears in 2-D, taking essentially a model- and view-based approach to segment, track and label body parts. Since self-occlusion makes the problem quite hard for arbitrary movements, many systems assume a priori knowledge of the type of movement or the viewpoint under which it is observed (e.g. lateral human gait [15] [18] [20] [22]); some choose to consider more unconstrained movements [16] [19] [26]. All in all, 2-D approaches with explicit shape models use stick figures, wrapped around with ribbons [16] [19] or, alternatively, "blobs" [26].

3-D approaches aim to recover 3-D articulated pose over time, i.e. joint angles with respect to an object-centered coordinate system. Once obtained, the resulting features have the advantage to be viewpoint invariant and thus directly linked to pose. The general problem of 3-D motion recovery from 2-D images remains difficult. For 3-D human tracking, however, one can take advantage of a priori knowledge about the kinematic and shape properties of the human body to make the problem tractable. Tracking can also be supported by the use of 3-D shape models which can predict events such as (self) occlusion and (self) collision. See for example [11] [14] [23] [24].

2-D approaches have generally been more natural for applications in surveillance than their 3-D counterparts. They are better suited for applications where image resolution is low, where single sensors are used (or uncalibrated multiple sensors) and where precise 3-D pose recovery is not needed. 3-D approaches make more sense for applications in controlled indoor environments where one has well-calibrated (multiple) sensors and one desires a high level of discrimination between various unconstrained and complex (multiple) human movements. This would typically be the case in virtual reality applications.

The prevalent view towards action recognition has been to consider it as a classification problem involving time-varying feature data; the feature data is derived from an earlier segmentation stage, using of the three approaches mentioned before. Recognition then consists of matching an unknown test sequence with a library of labeled sequences which represent the prototypical actions. A complementary problem is how to learn the reference sequences from training examples. Both learning and matching methods have to be able to deal with small spatial and time scale variations within similar classes of movement patterns. Previous work has used dynamic time warping [3] [8], hidden markov models [4] [25] [27] and neural networks [15] [17] to match time-varying feature data. Other work has used higher-level descriptions of scene activity based on symbolic reasoning [5] or scenarios [13].

For a more detailed discussion of work on gestures and whole-body movement, see [10].

3 Conclusion

Various desirable features of "smart" surveillance systems have been discussed: high information filtering capability, multi-sensor systems for saturated area coverage, integrated functionality, query capability, active sensor control and remote access. In terms of methods, 2-D approaches with and without explicit shape models for feature extraction were compared alongside 3-D approaches. Action recognition was considered as classification of time-varying feature data, or alternatively, facillitated by queries in high-level, symbolic constructs.

References

[1] J. Batista, P. Peixoto, and H. Araujo. Real-time active visual surveillance by integrating peripheral motion detection with foveated tracking. In *Proc. of First IEEE Int. Workshop on Visual Surveillance*, pages 18–25, Bombay, India, 1998.

[2] A. Baumberg and D. Hogg. An efficient method for contour tracking using active shape models. In *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 194–199, Austin, 1994.

[3] A. Bobick and A. Wilson. A state-based technique for the summarization and recognition of gesture. In *International Conference on Computer Vision*, pages 382–388, Cambridge, 1995.

[4] C. Bregler. Learning and recognizing human dynamics in video sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997.

[5] F. Bremond and M. Thonnat. Issues of representing context illustrated by video-surveillance applications. *Int. Journal of Human-Computer*, 1998.

[6] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models - their training and applications. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[7] R. Cutler and L. Davis. Real-time periodic motion detection, analysis and applications. In *IEEE Conference on Computer Vision and Pattern Recognition*, Fort Collins, U.S.A., 1999.

[8] T. Darrell and A. Pentland. Space-time gestures. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 335–340, New York, 1993.

[9] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997.

[10] D. Gavrila. The visual analysis of human movement: A survey. *Computer Vision Image Understanding*, 73(1):82–98, 1999.

[11] D. Gavrila and L. Davis. 3-D model-based tracking of humans in action: a multi-view approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–80, San Francisco, U.S.A., 1996.

[12] D. Gavrila and V. Philomin. Real-time object detection for "smart" vehicles. In *Submitted to International Conference on Computer Vision*, 1999.

[13] N. Goddard. Incremental model-based discrimination of articulated movement direct from motion features. In *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 89–94, Austin, 1994.

[14] L. Goncalves, E. Di Benardo, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3-D. In *International Conference on Computer Vision*, pages 764–770, Cambridge, 1995.

[15] Y. Guo, G. Xu, and S. Tsuji. Understanding human motion patterns. In *International Conference on Pattern Recognition*, pages 325–329, 1994.

[16] I. Haritaoglu, D. Harwood, and L. Davis. W4: Who, when, where, what: A real time system for detecting and tracking people. *Image and Vision Computing*, January, 1999.

[17] B. Heisele and C. Woehler. Motion-based recognition of pedestrians. In *International Conference on Pattern Recognition*, 1998.

[18] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parametrized model of articulated image motion. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 38–44, Killington, 1996.

[19] M. Leung and Y. Yang. First Sight: A human body outline labeling system. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 17(4):359–377, 1995.

[20] S. Niyogi and E. Adelson. Analyzing gait with spatiotemporal surfaces. In *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 64–69, Austin, 1994.

[21] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 193–199, San Juan, Puerto Rico, 1997.

[22] R. Polana and R. Nelson. Low level recognition of human motion. In *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–82, Austin, 1994.

[23] J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *International Conference on Computer Vision*, pages 612–617, Cambridge, 1995.

[24] K. Rohr. Towards model-based recognition of human movements in image sequences. *Computer Vision, Graphics and Image Processing: Image Understanding*, 59(1):94–115, 1994.

[25] T. Starner and A. Pentland. Real-time American Sign Language recognition from video using Hidden Markov Models. In *International Symposium on Computer Vision*, pages 265–270, Coral Gables, 1995.

[26] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 19(7):780–785, 1997.

[27] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–385, 1992.