LAYERED AND COLLABORATIVE GESTURE ANALYSIS IN MULTI-CAMERA NETWORKS

Hamid Aghajan, Chen Wu

Wireless Sensor Networks Lab
Department of Electrical Engineering
Stanford University, Stanford, CA 94305

ABSTRACT

A layered and collaborative architecture for gesture recognition in a multi-camera network is presented in this paper. The proposed approach is motivated by the diversity of gestures expressed in passive monitoring applications. It is based on the concept of opportunistic fusion of simple features within a single camera and active collaboration between multiple cameras in the decision making process. The decision process is pursued through mutual, assisted, and self correspondences, using features available at different cameras. The dynamics employed by the opportunistic fusion of different features within a single camera as well as those from multiple cameras offer the potential to address gesture recognition problems more efficiently and accurately across a variety of different applications.

Index Terms— Gesture analysis, Collaborative processing, Opportunistic data fusion, Camera networks

1. INTRODUCTION

The increasing interest in understanding human behaviors and events in a camera context has heightened the need for gesture analysis of image sequences. Gesture recognition problems have been extensively studied in Human Computer Interactions (HCI), where gestures are well-defined for delivering instructions to machines [1, 2]. However, "passive gestures" predominate in behavior descriptions of many applications. Some examples include surveillance and security applications, emergency detection in clinical environments, and video conferencing [3, 4]. Some approaches to analyzing passive gestures have been investigated in [5, 6].

Access to multiple sources of visual data often allows for making more comprehensive interpretation of events and gestures. Scalable implementation of multi-camera networks can be realized under a change of paradigm from centralized processing of raw data to distributed and collaborative implementation of vision-based reasoning algorithms at the network nodes. Besides access to different perspectives that can help determine a gesture, such approach to algorithm design also enables content-based employment of a variety of low-complexity algorithms instead of using computationally expensive techniques that need to universally run on various kinds of visual content. This paper sets forth a framework for analyzing human gesture based on opportunistic use of features available to the nodes of the network, and a layered and collaborative data analysis mechanism that systematically exploits the available information to achieve a description of the gesture.

An appropriate classification is essential towards a better understanding of the variety of passive gestures. Therefore, we propose a categorization of the gestures as follows:

- Static gestures, such as standing, sitting, lying;
- Dynamic gestures, such as waving arms, jumping;
- Interactions with other people, such as chatting;
- Interactions with the environment, such as dropping or picking up objects.

The proposed architecture for gesture recognition aims to accommodate the diversity of gestures and achieve efficient recognition in a multi-camera network. The layered structure consists of description layers and decision layers, which can be adapted to different subsets of gestures. This introduces flexibility for a variety of gesture applications. The collaborative decision process employs the concept of opportunistic data fusion of simple features within a single camera and active collaboration between multiple cameras in the decision making process. By employing different levels of collaboration, the proposed opportunistic feature fusion approach offers the potential to address gesture recognition problems more efficiently and accurately.

2. LAYERED AND COLLABORATIVE ARCHITECTURE

The overall architecture for the proposed gesture analysis approach is illustrated in Fig. 1. It consists of four description layers and three decision layers. From bottom to top, the four description layers are, layer 1 of images, layer 2 of features, layer 3 of gesture elements, and layer 4 of gestures. The three decision layers are the decision processes between neighboring description layers. With the layers going up, the abstraction of information contained in each description layer increases.

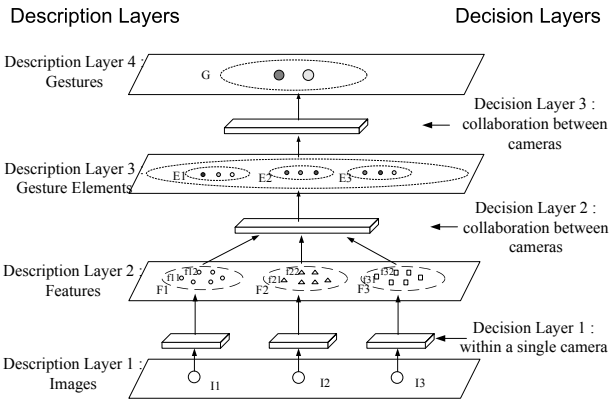

Fig. 1. The layered and collaborative architecture of the gesture recognition system. I_i stands for images taken from camera i; F_i is the feature set for I_i; E_i is the gesture element set in camera i; and G is the set of possible gestures.

In description layer 1, I_i is the image sequence of camera i. The images are processed and a feature set F_i in description layer 2 is obtained from I_i. Each F_i is a vector of several features $(f_{i1}, f_{i2}, \cdots)$. In description layer 3, gesture elements are selected for gesture models in description layer 4. They act as the bridge between low-level image features and high-level gesture descriptions. For each passive gesture category stated in Section 1, different gesture elements are of interest. Depending on the scope or diversity of the application, different subsets of a universal library of low-level features and gesture elements may be considered. Hence, the architecture has the flexibility to be applied to a variety of gestures.

The distinction between decision layer 1 and decision layers 2 and 3 is due to different characteristics of the data they process. Decision layer 1 takes in images from a camera, and outputs features based on these images. While in decision layers 2 and 3, the decision process takes in low-level elements from multiple cameras and outputs high-level elements. Elements in description layer 1 have a large volume since they are images. While in description layers 2 and above, elements have been reduced to abstract descriptions, which can be efficiently shared among neighboring cameras without imposing a heavy communication burden on the channel. Hence, in the proposed architecture, collaboration is provisioned in decision layers 2 and 3, and not in decision layer 1.

3. OPPORTUNISTIC APPROACH IN DECISION PROCESS

The underlying concept set forth through the decision making process is one of opportunistic fusion consisting of two aspects. First, within a single camera a number of simple features are adaptively aggregated, with processes that occur in decision layer 1. Second, between multi-view cameras, collaboration is actively pursued in different levels to employ the available pieces of information to reduce decision uncertainty. Collaboration occurs in decision layers 2 and 3.

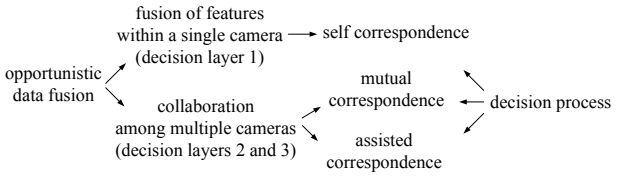

Fig. 2. Relationship of opportunistic data fusion, decision layers, and different kinds of correspondences through which decisions are made.

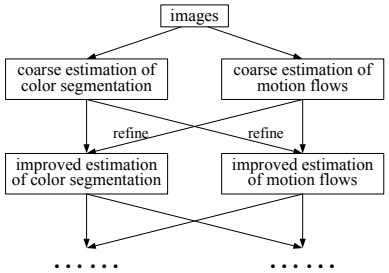

Fig. 3. Opportunistic fusion of features. Estimates of color-based segmentation and motion flow can be used to refine one another.

In terms of implementation, collaboration is achieved primarily through analysis of correspondence. We categorize decision processes into three kinds of correspondences, mutual correspondence, assisted correspondence, and self correspondence. Mutual and assisted correspondences involve collaboration among cameras, while self correspondence is the decision process within a single camera. Fig. 2 outlines the relationship of the two aspects of opportunistic fusion, the three kinds of correspondences in decision making, and the three decision layers.

3.1. Fusion of Features in A Single Camera

To illustrate the concept of opportunistic fusion of features, we choose the estimation of color segmentation and motion flows as two features in the feature vector F_i obtained in decision layer 1 for camera i. A first-step estimation is made for each feature. However, this estimation is coarse since we confine our algorithm with low complexity due to two considerations. One is that a complex algorithm may be costly in terms of image processing computation and time efficiency. And that in many applications it is difficult to obtain sufficient descriptions from images based only on one feature, even by employing complex algorithms. On the other hand, different features often complement each other. For example, it is possible to use results from color segmentation to refine those of motion flows, and vice versa. The algorithmic flow presenting this notion is shown in Fig. 3.

Two examples for opportunistic fusion of features in a single camera are given in Figs. 4 and 5. In Fig. 4(a), a prelimi-

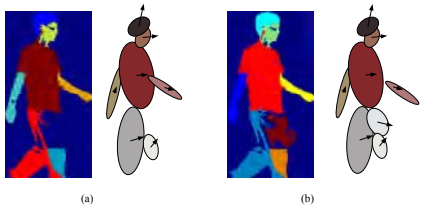

Fig. 4. Motion estimation assists segmentation. (a) is the result of segmentation based on color information; in (b), motion is used to find missing parts in (a).

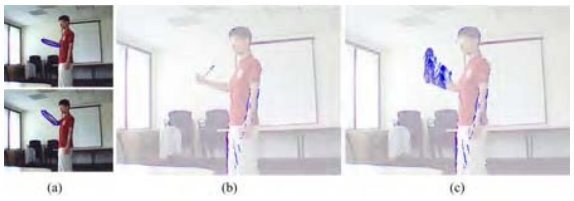

Fig. 5. Segmentation assists motion estimation. (a) shows ellipses fitted to the arm using segmentation results. (b) shows motion estimation without segmentation information. (c) uses angles of ellipses as parameters for motion estimation, and shows much improved performance.

nary color segmentation is obtained, in which a part of the leg is missing due to having a similar color to the background. In Fig. 4(b), motion flows have been used to segment the missing parts, since some of those parts have strong motions and are thus identifiable through optical flow estimation. In Fig. 5(a), a person is raising his arm. The ellipses fit to the arm from segmentation results are overlaid on the images. We are using a fast two-frame feature based motion estimation algorithm, which we have developed for translational motions. So in Fig. 5(b), the motion of the arm cannot be correctly detected without using the rotation information. In Fig. 5(c), the orientation angles of the ellipses are used as refinement parameters to obtain correct optical flow results for the arm.

3.2. Collaboration among Multiple Cameras

Collaboration among multiple cameras offers three major advantages in analyzing gestures. First, employment of multi-view cameras can help circumvent occlusions and provide an opportunity to identify the best view, especially when the human body itself is self-occlusive. Second, even without occlusions, gesture elements obtained from a single camera may be ambiguous for decision making, whereas a combination of gesture elements from multiple views may convey a higher-confidence interpretation of the gesture. Finally, since the image of a certain camera is a projection of the real gesture onto the image plane, gesture elements from different perspectives are correlated with each other. This correlation itself may impose helpful constraints on the gesture recognition process.

An illustration of how collaboration is implemented through

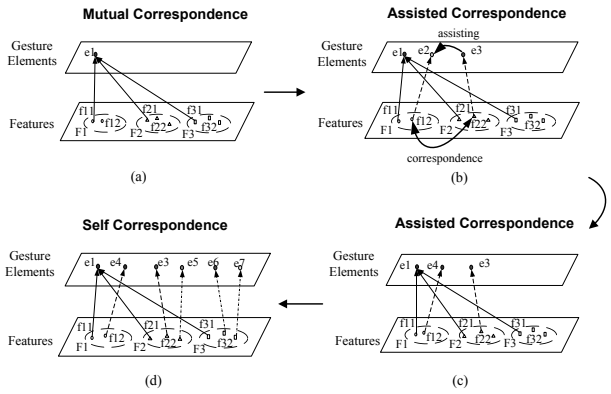

Fig. 6. Illustration for mutual, assisted, and self correspondences. (a) e_1 is decided through features from three cameras; (b) e_2, e_3 are obtained from f_{12}, f_{22} of cameras 1 and 2, respectively. But f_{12}, f_{22} are correlated; (c) e_2 is refined to e_4 according to constraints of the correlation; (d) Other correspondence are made.

mutual and assisted correspondence is shown in Fig. 6. This example is for decision layer 2, between the features layer and the gesture elements layer. A similar process holds true for decision layer 3 [7].

- *Mutual Correspondence.* Mutual correspondence refers to the presence of features from multiple cameras that reach to a common gesture element, as shown in Fig. 6(a). This happens when some gesture elements are more likely to appear simultaneously in multiple views, and their features have great similarities.
- *Assisted Correspondence.* In Fig. 6(b), an intermediate decision is made to (f_{12}, e_2) with a low confidence, and (f_{22}, e_3) with a high confidence. However, the correlation between f_{22} and f_{12} may have such an effect that if (f_{22}, e_3) has a high confidence, then f_{12} is highly probable to link to e_4.
- *Self Correspondence.* For some features in description layer 2, collaboration is not applied either because decisions for them can be obtained within a single camera with high confidence, or that the features are solely in the scope of a single camera.

Examples from a universal motion estimation application are shown in Figs. 7, 8, and 9. Collaboration among multiple cameras is actively pursued in order to identify the motion type. The human body model that the cameras are observing keeps certain consistent attributes between different views, therefore simple features can be used to reveal these consistencies. Collaboration is implemented by mutual and assisted correspondences. First, by mutual correspondence, strong directional trends from the majority of cameras are recognized, which helps to decide whether the motion is horizontal or vertical. Mutual correspondence is also applied to determine whether the vertical motion is sitting down or bending down. For sitting down, motion vectors of the upper body

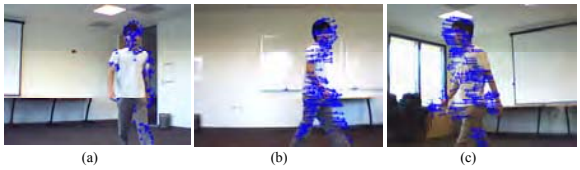

Fig. 7. Horizontal motion estimation. (a)(b)(c) are images from three cameras at the same time instance. Motion flows from a 2-frame optical flow scheme are also shown on top of the images. Horizontal motion is identified to be dominant, and directions are given as (a) forward (b) right (c) left.

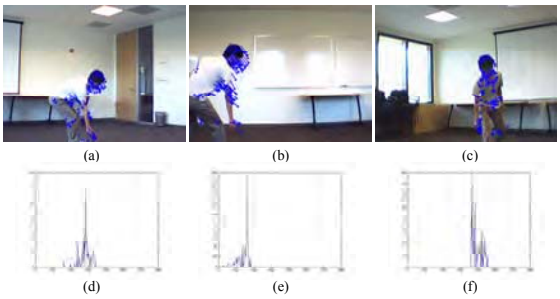

Fig. 8. Activity recognition: bending down for pick-up. (a)(b)(c) are images from three cameras at the same time instance. Vertical motion is identified as dominant. In (d)(e)(f) motion vectors are projected on the vertical axis. The high peaks and asymmetry are activity indicators.

are more uniform. So in Fig. 9 (d) (e) (f), projections of vectors do not have high peaks. Whereas in Fig. 8 (d) (e) (f) for the case of bending down, the projections present dominant local peaks and are highly asymmetric. Second, by assisted correspondence, directions of the motion are identified. Each camera will first make a preliminary estimation of the direction, and then magnitude constraints imposed by observations of all cameras are used to refine the estimates.

4. CONCLUSION

Our interest in passive gestures and a categorization of gesture types are presented. Considering both the flexibility to recognize a variety of gesture types and the distributed nature of multi-camera networks, a layered and collaborative architecture is proposed. The underlying concept of this architecture is an opportunistic fusion of data and decisions. This means that within a single camera, a number of simple features are aggregated based on the model, whereas between the cameras, collaboration is pursued in different levels to employ the available pieces of information in order to achieve better results. Specifically, collaboration is achieved by mutual and assisted correspondences.

The intention of our effort is to define a general approach that can be applied to recognition of the variety of natural gestures. In different applications, the interesting gestures and the set of gesture elements may vary. However, this architecture and the opportunistic approach to fuse information provide sufficient flexibility so that they can generalize.

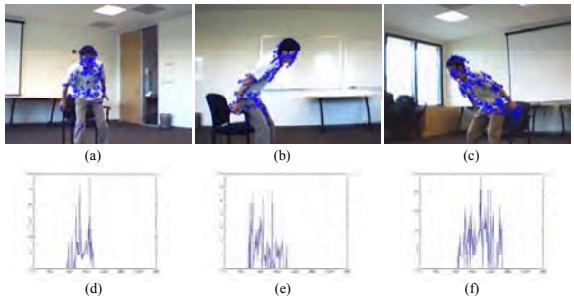

Fig. 9. Activity recognition: sitting down. (a)(b)(c) are images from three cameras at the same time instance. Vertical motion is detected. In (d)(e)(f) motion vectors are projected on the vertical axis. They have different characteristics from those of Fig. 8.

5. ACKNOWLEDGMENTS

Support provided by Micron Foundation is hereby gratefully acknowledged.

6. REFERENCES

[1] B. Kwolek, "Visual system for tracking and interpreting selected human actions.," in *WSCG*, 2003.

[2] G. Ye, J. J. Corso, and G. D. Hager, *Real-Time Vision for Human-Computer Interaction*, chapter 7: Visual Modeling of Dynamic Gestures Using 3D Appearance and Motion Features, pp. 103–120, Springer-Verlag, 2005.

[3] R. Patil, P. E. Rybski, T. Kanade, and M. M. Veloso, "People detection and tracking in high resolution panoramic video mosaic," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Oct. 2004, vol. 1, pp. 1323–1328.

[4] A. M. Tabar, A. Keshavarz, and H. Aghajan, "Smart home care network using sensor fusion and distributed vision-based reasoning," in *Proc. of VSSN*, Oct. 2006.

[5] J. Rittscher, A. Blake, and S. Roberts, "Towards the automatic analysis of complex human body motions," *Image and Vision Computing*, , no. 12, pp. 905–916, 2002.

[6] R. Cucchiara, A. Prati, and R. Vezzani, "Posture classification in a multi-camera indoor environment," in *ICIP05*, 2005, pp. I: 725–728.

[7] C. Wu and H. Aghajan, "Collaborative gesture analysis in multi-camera networks," in *ACM SenSys Workshop on DSC*, Oct. 2006.