

OPPORTUNISTIC FEATURE FUSION BASED SEGMENTATION FOR HUMAN GESTURE ANALYSIS IN VISION NETWORKS

¹*Chen Wu, Hamid Aghajan*

¹chenwu@stanford.edu

Wireless Sensor Networks Lab

Department of Electrical Engineering

Stanford University, Stanford, CA 94305

ABSTRACT

An image segmentation method for human gesture analysis is proposed deriving from the concept of opportunistic vision-based feature fusion. The method for human body part segmentation in image sequences is motivated by a layered and collaborative architecture for gesture analysis in multi-view camera networks. The layered structure aims to accommodate the diversity of gestures while collaboration embodies opportunistic fusion of information from multiple cameras. The proposed segmentation method is based on both motion and color information due to the complementary information represented by the two features. This method uses results from optical flow and background subtraction to initiate markers for a watershed algorithm to find the person's silhouette. Then K-means clustering is used for body part segmentation within the silhouette. Examples are given to illustrate complementary effects of different features and segmentation results. Potentials in feature fusion between multiple cameras are also discussed.

1. INTRODUCTION

The increasing interest in understanding human behaviors and events in a camera context has heightened the need for human 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, smart home care applications [3, 4], and video conferencing [5]. Some approaches to analyzing passive gestures have been investigated in [6, 7].

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. However, in the distributed algorithm, information shared between cameras needs to be low-bandwidth so as not to impose heavy communication load on the network. Therefore, visual data of each single camera is first analyzed locally, and then descriptions of attributes are transferred between cameras to make collaborative decisions.

In our gesture analysis approach, body part segments are obtained in each camera's view based on prior or learned on-the-fly knowledge of features. A human body model is maintained and updated through the parameters of the segments, which are communicated between cameras. Different methods have been proposed to find human body configurations in monocular cameras, some with good performance [8, 9]. However, self-occlusion of human body sometimes poses difficulties when only one view is analyzed, in which case often assumptions and prior motion dynamics need to be defined in order to obtain a reasonable model configuration. Therefore, instead of fitting a human model for a single camera we aim to employ observations from multi-view cameras to obtain a 3D human model description.

In this paper an image segmentation method based on opportunistic feature fusion is proposed in the context of human gesture analysis described above. To motivate this segmentation method, first in Section 2 we set forth a layered and collaborative data analysis framework that systematically exploits available information in the network for analyzing human gestures. The underlying notion is opportunistic fusion of features both within a single camera and between cameras. Image understanding for human body parts and their attributes is a crucial step towards gesture analysis, therefore in Section 3 the opportunistic feature fusion approach is applied to image segmentation for body parts. Details of segmentation with feature fusion in a single camera are presented in Section 3.1.

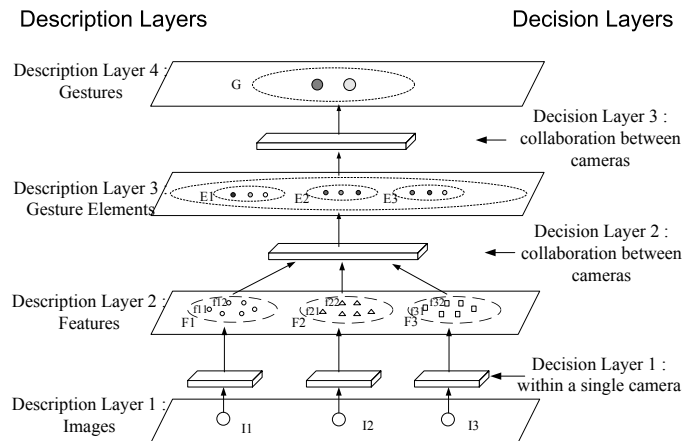


Figure 1: The layered and collaborative architecture of the gesture analysis system. I_i stands for images taken from camera i ; F_i is the feature set for I_i ; E_i is the gesture element set in camera i ; and G is the set of possible gestures.

2. OPPORTUNISTIC APPROACH FOR GESTURE ANALYSIS

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. Second, between multi-view cameras, collaboration is actively pursued in different levels to employ the available pieces of information to reduce decision uncertainty.

2.1. 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 [10]. With the layers going up, the abstraction of information contained in each description layer increases. The layered architecture for gesture analysis aims to accommodate the diversity of gestures as well as achieve efficient recognition in a multi-camera network. We propose a classification for passive 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 layered structure introduces flexibility to decompose manifold gestures into common gesture elements, which can be obtained through image features in a universal way irrespective of the specific gesture. Components of the layers can be adapted to different gesture subsets.

The collaborative decision process employs 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 opportunistic feature fusion approach offers the potential to address gesture recognition problems more efficiently and accurately. Specifically, segmentation is obtained from each camera as features, and our current work includes the design of a human body model and the mechanism to incorporate information from multi-view cameras as well as temporal updates. The derived model and gestures may again provide feedback to single cameras for a better directed early vision analysis. The concepts are shown in Fig. 2.

3. IMAGE SEGMENTATION FOR GESTURE ANALYSIS

Image segmentation for human body parts is a crucial step for gesture analysis. Although some gestures such as walking and falling can be detected through global attributes such as motion and silhouettes, detailed analysis requires appearance, position and motion of body parts interesting to the gesture. Image segments are effective representations from image observation to semantic gesture interpretations, as components in gesture elements layer in Fig. 1. One advantage of employing image segments as intermediate analysis outputs is their succinct data representation. This allows for efficient data exchange between cameras since segment descriptions contain small data amounts while carrying key information about the observed object. Another advantage of such a sketchy image representation instead of raw image recording or transmission is privacy protection required by many applications such as smart home care.

The problem setting of image segmentation for gesture

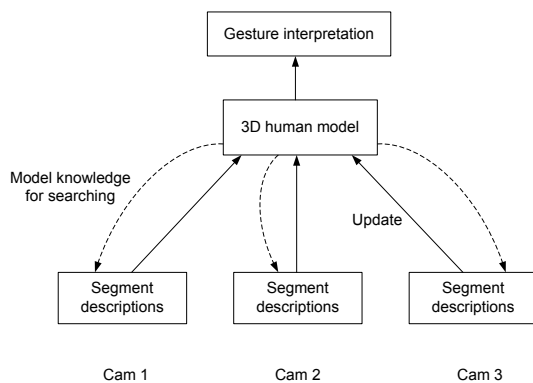


Figure 2: Human body part segmentation in the setting of collaborative gesture analysis.

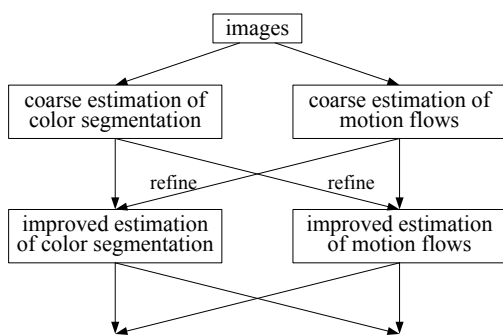


Figure 3: Opportunistic fusion of features. Estimates of color-based segmentation and motion flow are used to refine one another.

analysis is different from that of general image segmentation techniques. First, since the aim is specific, i.e., to segment persons in the image, the other objects are normally treated equally as background. Second, multiple features are available from image sequences and opportunistic fusion of them may greatly increase accuracy and efficiency of the segmentation process. Many generic image segmentation techniques based on a certain image feature have been proposed, and some produce good results at rather high computational complexity levels. However, in many situations it is difficult to obtain sufficient descriptions from images based only on one feature. Different image features complement each other based on attributes of the object of interest. For gestures, motion vectors obtained from optical flow analysis contain much information about body parts. Color is another important feature representing the appearance of the person. In this paper we propose a method for effective combining of these two vision-based features to obtain body part segments. This is conceptually shown in Fig. 3.

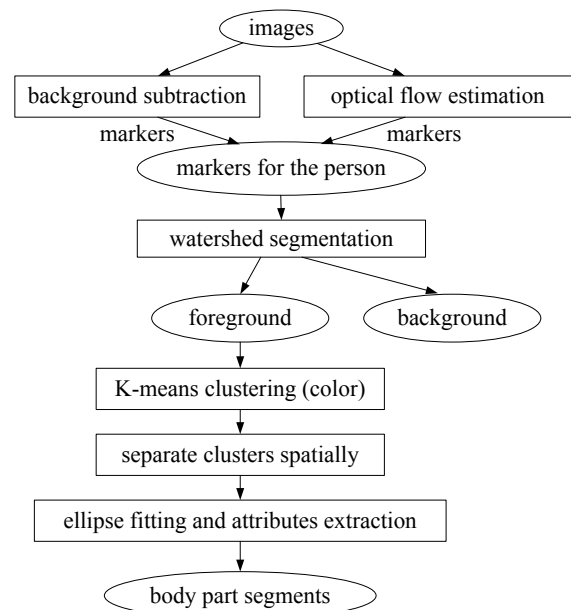


Figure 4: Algorithm flowchart of body part segmentation using color and optical flow information.

3.1. Feature Fusion in A Single Camera

The two features used in single camera feature fusion are optical flow and color information. Motion is usually a distinctive attribute in human gestures, therefore optical flow fields are used as one feature for segmentation. The advantage of using motion is that in indoor environments motion is a reliable feature to detect human activities most of the time, and the strength, direction and distribution of optical flow fields help distinguish body parts and gestures. But in situations where parts of the body stay still, motion in the image only gives partial information about the person. Color is also generally used for image segmentation. Since the appearance of the person rarely changes abruptly, color can be used to detect and keep track of the person. However, the person can be camouflaged with similar colors in the background. To remedy the weaknesses of both features, motion and color can be used to complement each other (Fig.5). While each of them may reveal part of the person, after combination in proper ways they can yield a better silhouette. The algorithm's flowchart is shown in Fig. 4.

First, optical flow computation and background subtraction are applied to the original image sequence. Optical flow is detected for a selected set of image features. Harris corners are detected as feature points, and then iterative Lucas-Kanade optical flow algorithm is applied in a pyramid on the feature points. The output motion vectors are further filtered to remove background noise with small magnitude and foreground outliers with large magnitude. Background subtraction is direct and the thresholded out-

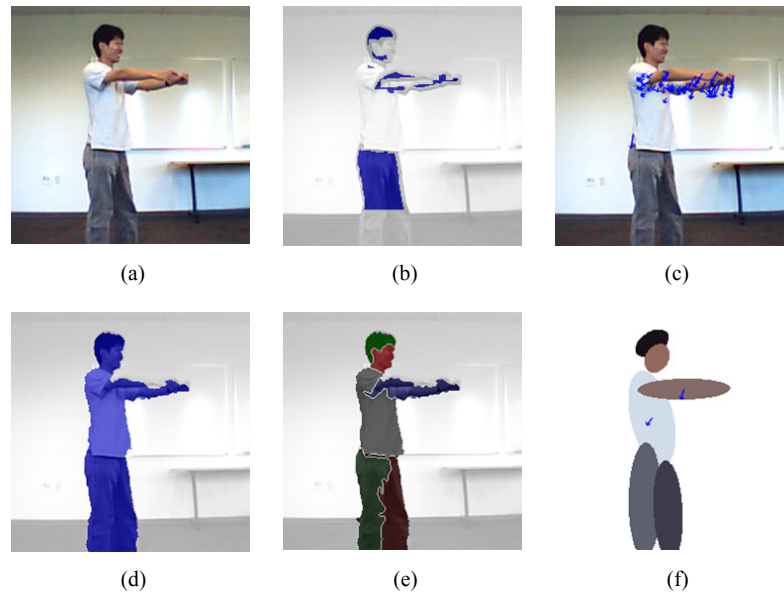


Figure 5: Feature fusion segmentation for body parts. (a) Original image. (b) Background subtraction. (c) Optical flow. (d) Watershed segmentation for foreground, using markers obtained from (b) and (c). (e) Foreground K-means clustering based on color. (f) Ellipsoid representation of body part segments, with average color and motion vector for each segment.

put is also filtered to remove outliers.

Second, the results of optical flow and background subtraction are used to generate markers for the watershed algorithm that separates the person from the background. The markers are preferably to cover as much of the person as possible, but usually motion or color alone would only give part of the person's silhouette. So we combine markers from both methods and then apply the watershed algorithm on the joint markers. There are two advantages of using the watershed algorithm. One is that it yields a single region for the person's silhouette. The other is that it is less sensitive to clusters in the background since the initial markers in the background prevent foreground region from growing too much.

After the person's silhouette is located, it is segmented by K-means clustering based on color. Since the clustering process only considers color without location information, cluster labels for pixels are put back into the image to find spatially connected regions, which will indicate the body parts. Usually five or six largest regions suffice for a meaningful interpretation of body parts. Then ellipsoid fitting is applied to each segment, and average color and average motion vector are calculated as segment descriptors.

In Fig. 5 an example is shown to illustrate the different steps of the algorithm and the complementary roles of color and motion in the segmentation process. Fig. 5(a) is the original image, showing the person doing some exercise. Blue areas in Fig. 5(b) show the result of back-

ground subtraction. The shirt is missing after background subtraction since it has a white color similar to the wall. Blue arrows in Fig. 5(c) are optical flows, from which the arms and part of the shirt can be detected since they are moving, while other parts of the body remain still. Results from Fig. 5(b) and (c) generate markers for the foreground (the person's silhouette), while markers for the background can be simply taken as the bounding box which covers both optical flows and foreground segments from (b). The watershed image segmentation algorithm yields a connected silhouette as shown in Fig. 5(d), which delineates better contours than each of the individual methods in (b) and (c). Because of the region continuity forced by the algorithm, some regions in the silhouette which cannot be recognized as foreground in (b) and (c) are connected to the foreground by the watershed, as long as some of their near neighborhoods are marked as foreground. In Fig. 5(e) K-means clustering on the foreground is applied, and large regions are marked as interesting body parts. In Fig. 5(f), the best-fit ellipse, average color and average motion vector are calculated for each region.

More examples from different scenarios are shown in Fig. 6. The images are segmented using the method described in this section, and the resulting body part segment descriptions can be used as the basis for further gesture analysis.

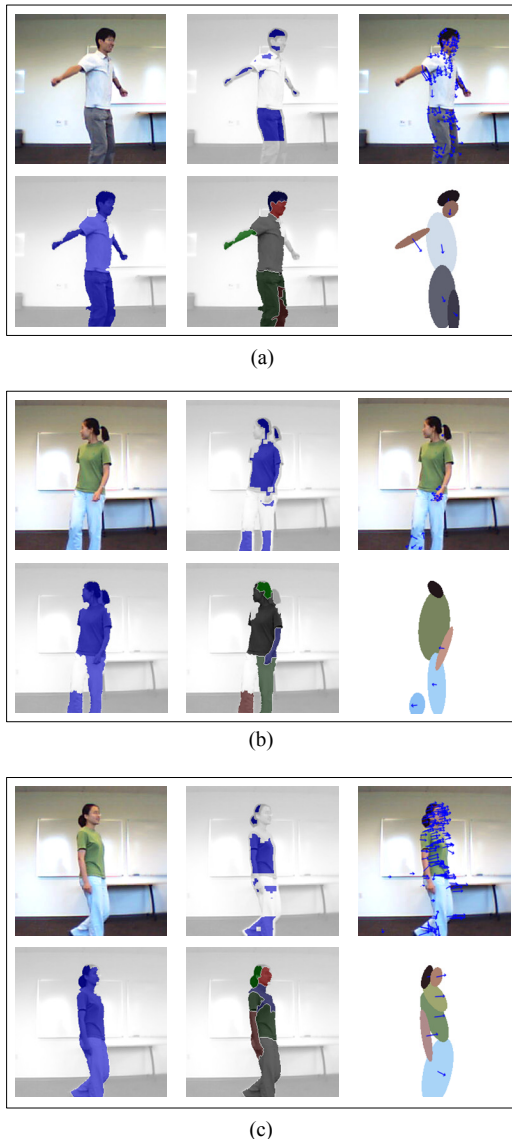


Figure 6: Examples of body part segmentation in gesture analysis. (a) exercise. (b), (c) walking.

4. CONCLUSION

An image segmentation algorithm for human body parts is proposed based on opportunistic fusion of features in image sequences. This algorithm acts as an intermediate step for gesture analysis in camera networks. A layered and collaborative architecture was proposed for algorithm design in camera networks. A segmentation algorithm based on both motion and color information was introduced. The method combines results from these two features to initiate markers for a watershed algorithm to segment the person's silhouette from background. A K-means clustering scheme was proposed for body part seg-

mentation within the silhouette. Examples illustrated complementary effects of different features and segmentation results. Potentials and future work in feature fusion between multiple cameras were also discussed.

5. ACKNOWLEDGMENT

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] A. M. Tabar, A. Keshavarz, and H. Aghajan, "Smart home care network using sensor fusion and distributed vision-based reasoning," in *ACM Multimedia Workshop on VSSN*, Oct. 2006.
- [4] A. Keshavarz, A. M. Tabar, and H. Aghajan, "Distributed vision-based reasoning for smart home care," in *ACM SenSys Workshop on DSC*, Oct. 2006.
- [5] 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.
- [6] 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.
- [7] R. Cucchiara, A. Prati, and R. Vezzani, "Posture classification in a multi-camera indoor environment," in *ICIP05*, 2005, pp. I: 725–728.
- [8] J. Deutscher, A. Blake, and I.D. Reid, "Articulated body motion capture by annealed particle filtering," 2000, pp. II: 126–133.
- [9] H. Sidenbladh and M.J. Black, "Learning the statistics of people in images and video," vol. 54, no. 1-3, pp. 183–209, August 2003.
- [10] C. Wu and H. Aghajan, "Layered and collaborative gesture analysis in multi-camera networks," in *ICASSP*, Apr. 2007.