

Collaborative Gesture Analysis in Multi-Camera Networks

Chen Wu

Wireless Sensor Networks Lab
Department of Electrical Engineering
Stanford University, Stanford, CA 94305
chenwu@stanford.edu

Hamid Aghajan

Wireless Sensor Networks Lab
Department of Electrical Engineering
Stanford University, Stanford, CA 94305
aghajan@stanford.edu

Abstract

An architecture for opportunistic discovery of gesture elements for analysis of human gestures in a multi-camera sensor network is presented in this paper. The proposed approach is motivated by the diversity of gestures expressed in passive monitoring applications, and 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. By reducing the uncertainty through different levels of collaboration, the proposed opportunistic approach offers the potential to address gesture recognition problems more efficiently and accurately. Details of the description-layered and processing-layered architecture for collaborative gesture analysis and some illustrating examples are presented.

Categories and Subject Descriptors

I.2.10 [Vision and Scene Understanding]: Modeling and recovery of physical attributes, Perceptual reasoning

General Terms

Algorithms

Keywords

Image sensor networks, Gesture recognition, Opportunistic data fusion, Collaborative decision process

1 Introduction

Image sensor networks have drawn increasing attention in recent years, mainly for two reasons. First, image sensors can act as sources of rich information from the environment, enabling complex interpretations of events. Second, the distributed nature of sensor networks introduces flexibility in system scaling, as well as different perspectives to the same event or subject of interest.

In this paper, we propose an architecture to recognize human gestures in a distributed image sensor network. The architecture is developed based on the concept of using an opportunistic data fusion approach consisting of two aspects. One is fusion of simple features within a single camera, and the other one is fusion based on collaboration between multiple cameras.

Many studies have focused on understanding human gesture based on images and video. Up to date, gesture recognition problems have been extensively studied in Human Computer Interactions (HCI), where gestures are well-defined to

give instructions to machines [5, 3]. However, in many applications, “passive gestures” play an important role. Gesture recognition as well as tracking can bring about intelligence to the environment in security-related contexts such as building surveillance and home security [4, 1]. In clinical environments, automatic gesture recognition can be used to detect emergencies, or for smart home care networks [9]. Other applications include video conferencing, lecture assistance, etc. [6]. Several approaches to analyzing passive gestures have been proposed [2, 8, 11]. However, due to the variety of passive gestures, these methods are usually specific to differentiating a small set of gestures.

2 Collaborative Gesture Analysis

Passive gestures are of particular interest in a wide range of applications. However, the diversity of passive gestures makes them especially difficult to model and recognize. We believe that an appropriate classification of these gestures is essential towards a better understanding of them. Some classifications of gestures have been proposed [5]. Considering the nature of different gestures, we categorize 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.

Gestures falling in different categories may involve quite different elements in the recognition process. Therefore, specific models and methods can be designed for each category. Due to the non-rigidity of the human body and the variety of appearances, the more accurate a model is in describing a gesture, the larger number of constraints it will have. Defined constraints are often far from being comprehensive, so the effectiveness and accuracy of gesture analysis techniques is seriously challenged even in the case of complex models and constraints. Furthermore, accurate models are difficult to obtain. Therefore, the trade-off between the simplicity and effectiveness needs to be balanced.

To address the diversity of passive gestures, we propose an opportunistic approach which has two aspects. First, different low-level features are aggregated adaptively instead of solely relying on a certain feature to make next-level decisions. This approach has potentials to perform better

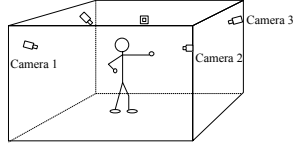


Figure 1. The physical architecture of the system

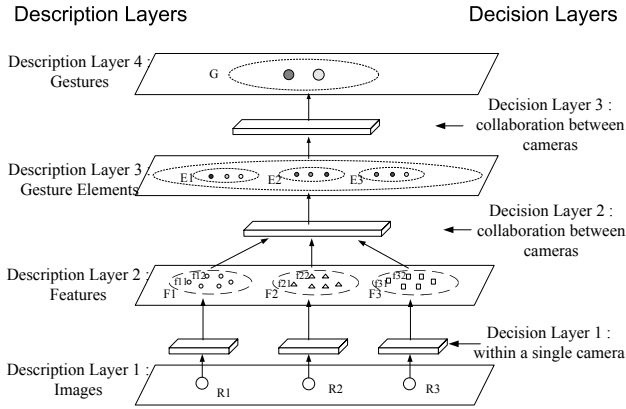


Figure 2. The logical architecture of the system. R_i stands for images taken during a wake-up of camera i ; F_i 's are feature sets for R_i 's; E_i 's are gesture element sets; and G is the set of possible gestures.

since there is no single feature that works well in all circumstances. And information from multiple cameras add to the dimensions of information, which increases the validity of decisions if all information is used properly. Second, our algorithm actively combines information from multi-view images in different levels, i.e., collaborative information processing is performed. Collaborations may happen in several levels, and in some situations they will greatly facilitate decision-making.

2.1 Architecture Overview

The physical structure for the multi-camera network is illustrated in Fig. 1. There are no particular requirements for the placement of cameras. As long as a person is in a camera's Field of View (FOV), the camera will actively participate in collaboration for gesture recognition. Instead of recording and analyzing videos, cameras in this system "wake up" at certain time intervals, and sequentially take several frames every time. Employing such duty cycle mode is more compatible with the operation characteristics of sensor networks.

The logical architecture for our gesture analysis approach is illustrated in Fig. 2. This structure is hierarchical, and 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.

We now term specific values of description layers as "elements", and decision flows between description layers as "links". Note the distinction between decision layer 1 and decision layers 2 or 3. Decision layer 1 takes in an image from a camera, and outputs features based on this image, i.e., no collaboration between cameras is involved. However, in decision layers 2 and 3, the decision process takes in low-level elements from multiple cameras and outputs high-level elements. That is, collaboration between cameras happens in decision layers 2 and 3. 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.

2.2 Description Layers

In description layer 1, each element is a sequence of raw images captured by a camera during a wake-up. For example, element R_1 contains images $R_{11}, R_{12}, \dots, R_{1n}$, which are sequentially taken by camera 1 during a short period of time.

In description layer 2, an element f denotes a certain feature that is obtained from a specific element R . As in Fig. 2, F_i is the set of features from R_i . Each feature set F_i may contain several features, $f_{i1}, f_{i2}, \dots, f_{im}$. Those features are descriptive and chosen according to the following criteria: 1. They are representative of the gestures we are interested in; 2. They are simple and computationally efficient. Here we use the opportunistic approach on the features in the following two ways: First, for the multiple features within a single image, one or more are adaptively chosen as primary features, and others as auxiliary ones. Decisions are made relying more on the primary ones. Second, correspondence of features from different views re-enforces credence of each single feature. Hence by collecting several simple features and trying to combine their merits, the algorithm achieves higher reliability that cannot be attained from individual features. Some of the features we are currently using are: global motions parallel and orthogonal to the image plane (from optical flow), and segments with attributes such as color, position, area, aspect ratio, rotation, and local motions.

In description layer 3, each element represents the pair of a gesture element and its confidence. Gesture elements would be designed for gesture models in description layer 4. According to our classification of gestures, we also classify gesture elements into the following categories:

- Posture: body posture at a certain time instance (upright, hands-up, lying down);
- Global motion: motion pattern of the whole body in a time period;
- Local motion: motion pattern of a body part (arms, legs) in a time period;
- Objects: changes in the environment;
- Other people: those the subject is interacting with.

In description layer 4, each element denotes the pair of a gesture and its confidence. For example, in Fig. 2, the darker element may indicate a gesture with higher confidence label.

2.3 Collaborative Decision Layers

There are three decision layers in our model facilitating progression between the description layers using each layer's elements. These are designed based on the underlying idea of allowing collaborations to permeate into as many levels as possible. Collaboration is primarily achieved through analysis of correspondence.

Decision layer 1 uses single elements in description layer 1 without collaboration with other cameras. It consists of simple processes applied to the sequence of images to obtain the features used in layer 2. The flowchart in Fig. 3 illustrates the collection and interactions between the different algorithmic functions used. The lack of collaboration between the cameras in this layer is a result of designing the architecture without image transfer between the network nodes.

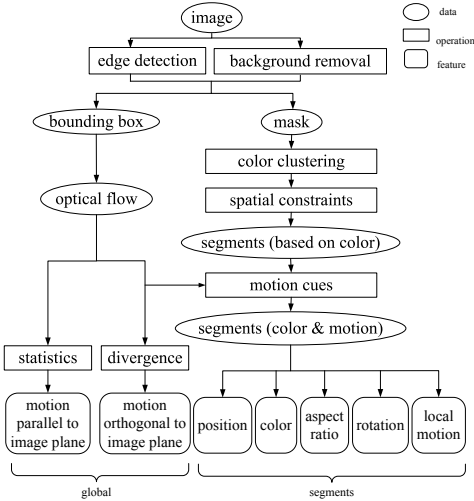


Figure 3. Algorithm flowchart for decision layer 1

In both decision layers 2 and 3, collaboration is actively pursued by the cameras. Although the details are different, these decision layers have common priority rules. Decision layer 2 is used as an example to explain these rules here. It is based on processing priorities as shown in Fig. 4, and collaboration priorities as illustrated in Fig. 5.

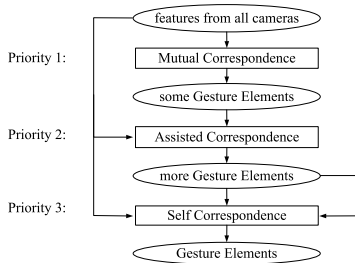


Figure 4. The processing priorities for decision layers 2 and 3

- **Mutual Correspondence.** In Fig. 5(a), three features f_{11}, f_{21}, f_{31} have such a strong correspondence that this correspondence alone can lead to a high confidence gesture element e_1 . This happens when some gesture ele-

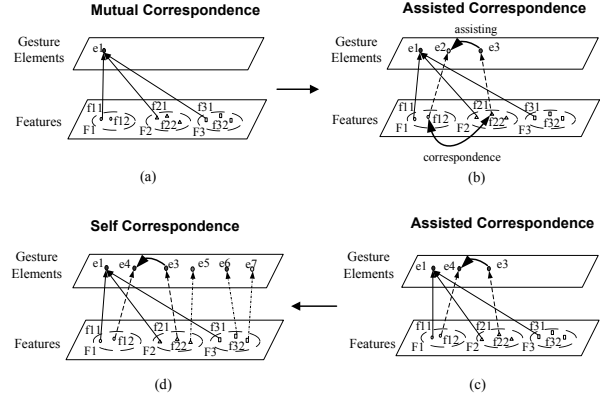


Figure 5. An example for collaboration priorities

ments are more likely to appear simultaneously in multiple views, and features for these gesture elements have great similarities. Hence, mutual correspondence refers to the presence of features in each camera that lead to a common gesture element. This is a rather strong condition, and usually has high confidence results.

- **Assisted Correspondence.** After mutual correspondence has been identified and decisions made accordingly, further decisions will be made on the basis of existing ones. In Fig. 5(b), f_{12} and f_{22} have a correspondence, but they do not apply to priority 1 because they do not have counterparts in F_3 . Then f_{12} links to e_2 based on (f_{11}, e_1) , but with a low confidence. f_{22} links to e_3 based on (f_{21}, e_1) with a high confidence. However, the correspondence 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 . Therefore (f_{22}, e_3) assists the transition of f_{12} . So in Fig. 5(c), we see that f_{12} links to e_4 .
- **Self Correspondence.** Up to this stage, the possible correspondences have all been employed. For the other features in description layer 2 that haven't transitioned to any higher layer elements, decisions will be made within the scope of a single camera, as is shown in Fig. 5(d).

2.4 Posture Analysis

Postures are static descriptions of gestures, and constitute one category of gesture elements. It is useful to analyze postures first, and then combine them with other gesture elements. In layer 2 we use parameters to describe the postures, leaving the semantic assignments to decision layer 3, which uses gesture elements to determine gestures. The posture model is described as (φ, α, β) as in Fig. 6, where $\varphi = (\varphi_1, \varphi_2)$, $\alpha = (\alpha_1, \alpha_2)$, and $\beta = (\beta_1, \beta_2)$ (Now we are only interested in the upper half of the body). We have features from, say, 3 images (I_1, I_2, I_3) from different views, and our aim is to label the blobs of these 3 images collaboratively, and obtain the posture parameters. The feature set of I_i includes $F_i = (m_p, m_o, Seg = \{seg_i\})$, where m_p and m_o are respectively the global motions parallel and orthogonal to the image plane, $seg_i = (pos, c, r, \theta, v)$ is the feature vector

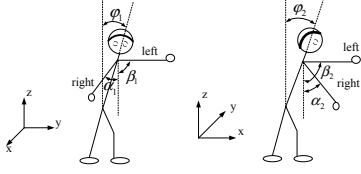


Figure 6. Posture parameters. The person is facing x direction, and we project the posture to zOy and zOx planes, respectively. φ is the angle of the torso; α is for the right arm and β for the left.

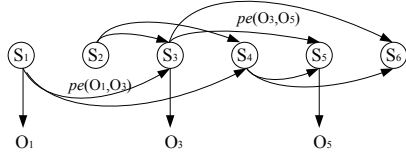


Figure 7. HMM states and transitions. Penalty based on similarity between observations is associated with the arcs.

for the i th segment, pos (position) is the centroid of seg_i , c (color) is the mean color of seg_i , r is the aspect ratio of the ellipse fitted to seg_i , θ is the rotation angle of the ellipse, v is the motion vector of seg_i with respect to the global motion, i.e., it represents relative motion of the segment.

The analysis takes two major constraints into consideration. The first one is the correspondence between the images taken from different views; the second one is the intrinsic location constraints of body parts. We use a Hidden Markov Model (HMM) [10, 7] for each of the body parts. Each image corresponds to a stage, and the states indicate whether or not a body part is in the image. For example, for right arm, we have

$$S_j : \begin{cases} \text{right arm is in } I_{\lceil \frac{j}{2} \rceil} & \text{if } j \text{ is odd,} \\ \text{right arm is not in } I_{\lceil \frac{j}{2} \rceil} & \text{if } j \text{ is even.} \end{cases} \quad (1)$$

So for (I_1, I_2, I_3) the state space is $S = \{S_1, S_2, S_3, S_4, S_5, S_6\}$. When $j = 1, 3, 5$, there are observations $O_j = (pos, r, \theta)$. We assume a Gaussian Mixture Model (GMM) for O , i.e. the pdf is

$$b_j(O) = \sum_{m=1}^M c_{jm} G(O, \mu_{jm}, \Sigma_{jm}), \quad (2)$$

where c_{jm} is the mixture coefficient for the m th mixture at state j , $G()$ is Gaussian, μ_{jm} is the mean, and Σ_{jm} is the covariance matrix. The transitions of the HMM are shown in Fig. 7. Note that when S_i, S_j both have observations, there is an additional cost for the arc representing a dissimilarity penalty $pe_{ij}(c, v)$.

3 Experiments

In our experiments, three synchronized cameras are taking images from different views. Fig. 8 includes two examples of obtaining features from images in decision layer

1. Original images obtained during a certain wake-up are shown in Fig. 8 (1a),(2a),(3a). Fig. 8 (1b),(2b),(3b) are segmentation results based on color and motion, and Fig. 8 (1c),(2c),(3c) show all the features described in Fig. 3. The segments are displayed as ellipses just for illustration, although we don't need the elliptical shapes for later decision processes.

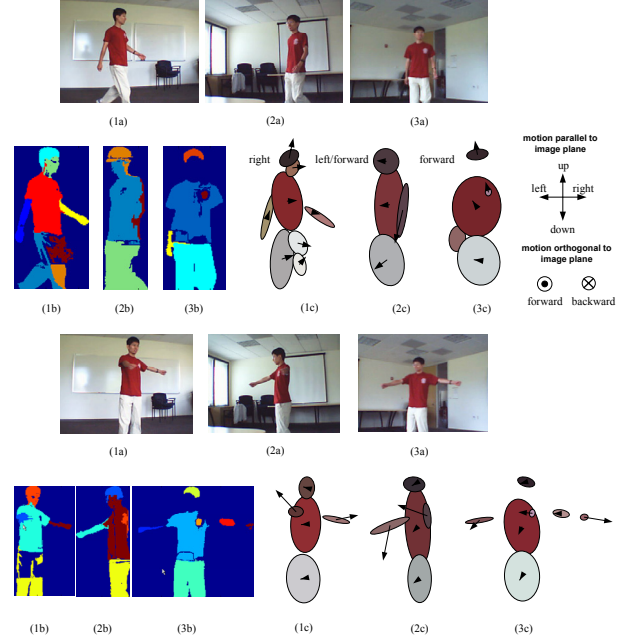


Figure 8. Images and segmentation results. (1a)(2a)(3a) are original images from Camera 1, Camera 2 and Camera 3; (1b)(2b)(3b) are segmentation results based on both color and motion; (1c)(2c)(3c) are features represented graphically.

Since in many cases neither color nor motion would provide sufficient basis for segmentation, they are combined in layer 1 to avoid missing important body parts. As shown in Fig. 9(a), the upper part of the leg is missing from color segmentation, because its color is similar to the background. However, after combining information from optical flow, this part is found as in Fig. 9(b), since it has a strong motion in the direction of the person, who is walking forward.

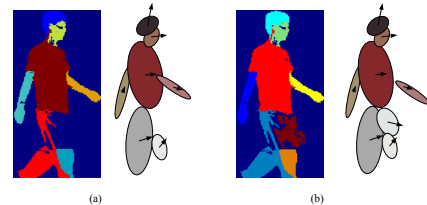


Figure 9. An example of using motion for segmentation. (a) is the result of segmentation based on color information; in (b), motion is used to find missing parts from (a).

In Fig. 10, the process of collaborative labeling is illustrated. This process transitions from description layer 2 to description layer 3, i.e., using features to label gesture elements. In this example gesture elements are defined to

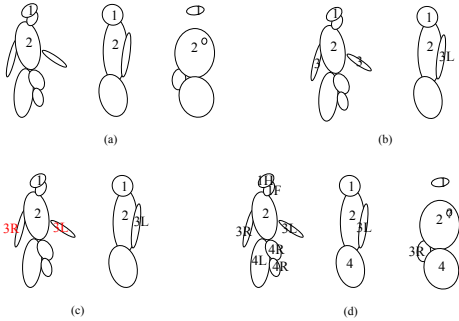


Figure 10. An example of gesture elements. (a) is the labeling result after Mutual Correspondence; (b)(c) show the process of Assisted Correspondence; The final labeling is shown in (d). 1:head (1H:hair, 1F:face), 2:body, 3:arm (3L:left arm, 3R:right arm), 4:leg (4L:left leg, 4R:right leg).

a set including head (1, and 1H for hair, 1F for face), body (2), arm (3, and 3L for left arm, 3R for right arm), leg (4, and 4L for left leg, 4R for right leg). First the algorithm tries to find Mutual Correspondence. Because of their similarity in relative position, color, area, etc. in all the three views, head and body are labeled in the first place. This correspondence also has a high reliability in that head and body are likely to be seen from all viewing angles, while arms and legs are likely to be occluded in some views. It is exactly this occlusion-disparate property of different body parts that makes Mutual Correspondence efficient for labeling. Assisted Correspondence proceeds as follows. In the right image in Fig. 10(b), the left arm is labeled from information of both global motion and occlusion hypothesis. For the left image in Fig. 10(b), we cannot tell which arm is left and which is right. However, from the right image, we know that the left arm has the same motion direction as the global motion, and using this information we look back into the left image again, and the left and right arms can be labeled as shown in Fig. 10(c). Note that depending on the specific application, different gesture sets can be defined, followed by different definitions of gesture elements. But the collaboration priorities are expected to achieve higher efficiency and accuracy in most cases.

4 Conclusion

The interest in passive gestures is presented and a categorization of them is put forward in order to gain a systematic approach to gesture recognition. An architecture for gesture recognition is proposed afterwards. The four description layers and three decision layers of this architecture are explained with some illustrating examples. The underlying concept set forth by the proposed approach to address the problem is an opportunistic fusion of data and decisions. This means that within a single camera, a number of simple features are aggregated adaptively for the model, whereas between multi-view cameras, collaboration is pursued in different levels to employ the available pieces of information to reduce decision uncertainty. Experimental results are shown to illustrate the proposed opportunistic approach.

In different applications, the interesting gesture expressions and accordingly the type and composition of gesture elements would vary. However, this architecture and the opportunistic approach to fuse information will generalize. In fact, the intention of our effort is exactly to define a general approach that can be applied to recognition of the variety of natural gestures. Our future work consists of building up a library of tools for different levels of descriptions and decision processes that enable the network to adaptively choose a subset of available tools for online gesture recognition.

5 References

- [1] G. J. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *IEEE Computer Vision and Pattern Recognition*, pages I: 594–601, 2006.
- [2] R. Cucchiara, A. Prati, and R. Vezzani. Posture classification in a multi-camera indoor environment. In *ICIP05*, pages I: 725–728, 2005.
- [3] Guangqi Ye, Jason J. Corso, and Gregory D. Hager. *Real-Time Vision for Human-Computer Interaction*, chapter 7: Visual Modeling of Dynamic Gestures Using 3D Appearance and Motion Features, pages 103–120. Springer-Verlag, 2005.
- [4] L. D. I. Haritaoglu, D. Harwood. W4: real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830, 2000.
- [5] B. Kwolek. Visual system for tracking and interpreting selected human actions. In *WSCG*, 2003.
- [6] R. Patil, P. E. Rybski, T. Kanade, and M. M. Veloso. People detection and tracking in high resolution panoramic video mosaic. In *Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*, volume 1, pages 1323–1328, October 2004.
- [7] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [8] J. Rittscher, A. Blake, and S. Roberts. Towards the automatic analysis of complex human body motions. *Image and Vision Computing*, (12):905–916, 2002.
- [9] A. M. Tabar, A. Keshavarz, and H. Aghajan. Smart home care network using sensor fusion and distributed vision-based reasoning. In *Proc. of VSSN 2006*, October 2006.
- [10] A. D. Wilson and A. F. Bobick. Parametric hidden markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):884–900, 1999.
- [11] Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding: CVIU*, 73(2):232–247, 1999.