

Combining Motion Segmentation with Tracking for Activity Analysis

Jiang Gao, Alexander G. Hauptmann and Howard D. Wactlar
School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213
{jgao, alex, wactlar}@cs.cmu.edu

Abstract

We explore a novel motion feature as the appropriate basis for classifying or describing a number of fine motor human activities. Our approach not only estimates motion directions and magnitudes in different image regions, but also provides accurate segmentation of moving regions. Through a combination of motion segmentation and region tracking techniques, while filtering for temporal consistency, we achieve a balance between accuracy and reliability of motion feature extraction. To identify specific activities, we characterize the dominant directions of relative motions. Experimental results show that this approach to motion feature analysis could be successful in assisting caregivers at a nursing home in assessments of patient's activity levels over time.

1. Introduction

Much recent research has been focused on activity analysis in videos. Past works have been conducted on applying both high-level modeling and feature selection strategies. Several different levels of feature analysis have been proposed, ranging from spatial-temporal histograms to object tracking. While object tracking is more accurate and able to analyze more subtle activities, it is also more fragile, sensitive to noise, and in many cases requires manual initialization.

In this paper, we propose an algorithm for generating a motion feature which is self-initiating, robust to noise, and at the same time sensitive to subtle motions. The algorithm is based on combining motion segmentation and tracking.

We have developed a motion segmentation algorithm for human body motion, based on optical flow and RANSAC [3]. The algorithm is able to detect and label parametric motion patterns of natural human motion with normal clothing and in natural environments.

Based on the previous work, in this paper we propose a strategy for combining motion segmentation with



Figure 1. Activity analysis at a nursing home. The green curves indicate dining activities of the person of interest through out time.

tracking. The combination enables the algorithm to be sensitive to subtle body motions, while remaining self-initiating and robust to tracking errors. Combining motion segmentation and tracking results, we further propose a weighted sequential projection algorithm to detect temporally consistent motions. The algorithm filters out inconsistent or random motions, which are noisy features irrelevant to our task.

The major contribution of this paper is providing a new motion feature by combining the above algorithms. The proposed motion feature is applied to analyze patient activities at a nursing home.

The organization of this paper is as follows: Section 2 reviews existing feature selection strategies in activity analysis. Section 3 briefly reviews our methods for motion segmentation in video. Section 4 discusses our strategy in combining motion segmentation and tracking in video, and proposes a method to detect temporally consistent motions. Section 5 describes the algorithms for recognizing dining room activities at a nursing home. Section 6 gives experimental results. Finally, section 7 concludes the paper and discusses the application issues.

2. Related research

Activity analysis from video has become an active research area in recent years. A successful system combines effective features and high-level modeling in order to recognize activities. In Stauffer and Grimson

[6], a stable, real-time outdoor tracker is proposed to provide low-level features, and high-level classifications are based on analysis of the output from this tracking system. In their real-time outdoor tracker, motion segmentation is based on an adaptive background subtraction method.

In Zelnik-Manor and Irani [12], dynamic events are regarded as long-term temporal objects, and spatio-temporal features at multiple temporal scales are derived and utilized. Specifically, they take the absolute value of normalized space-time gradients at multiple temporal scales as local feature measurements, and use a smoothed histogram of these local features to compute the distance between video sequences. The outcome is a clustering which results in a temporal segmentation of long video sequences into event-consistent sub-sequences.

In Bissacco [2], human gaits are recognized in the space of dynamic systems describing a human body in motion. The features are obtained by tracking a kinematic model of a human body in motion.

Furthermore, in Starner [5], skin color detection and moments of blobs are used as features to recognize sign languages; in Wilson and Bobick [10], wide baseline stereo cameras and flesh tracking are used to compute the three-dimensional position of head and hands. HMM and parametric HMM models are used to recognize sign language and gestures.

There is also much work on high-level modeling of activities, for example, in [9], but the focus of this paper is on how to obtain robust features for activity analysis, so we will not discuss these works in here.

In this paper, our goal is to classify nursing home patients' activity levels and their change over time. We are interested in features that can differentiate and quantify the level (e.g. frequency, duration, magnitude) of these activities. As a result, we are developing a feature that lies between the detailed features needed for human gesture recognition and the more coarse level spatial-temporal features commonly used for long range activity classification.

The feature developed in this paper is based on a combination of motion segmentation and tracking. *Motion segmentation* aims at finding major motion patterns in image sequences, and segments images into regions corresponding to these motion patterns. We developed an algorithm akin to the layered representation of video by Wang [8] and layer extraction using color segmentation by Altunbasak [1].

We use motion segmentation to provide initialization of moving regions, and *track* these regions through a limited time window. At the same time, we perform a consistent motion filtering, and delete regions with random motions. In this way we provide an effective

motion feature that is both sensitive to subtle motions, self-initiating, and robust to noises at the same time.

3. Motion segmentation

Fig. 2 shows the motion segmentation problem. The key is to find the same moving region in each frame, and the projection transforms mapping this region between frames. We used a layered approach for motion segmentation, but unlike classical layer representation works, our goal is more like that of Tao [7], i.e., finding 2D layers corresponding to different moving objects. The difference from Tao is that we are dealing with non-rigid motion of human body parts.

Our method is to model the apparent motions in video based on optical flow and random sample consensus (RANSAC). We are using affine and homography models to describe motions of body parts.

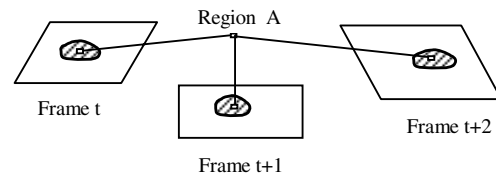


Figure 2. Motion segmentation.

An example result of our motion segmentation algorithm is given in Section 6. In this paper, our focus is on developing an algorithm to combine tracking with motion segmentation, in order to obtain better motion features.

4. Tracking and consistent motion

From motion segmentation, we get several regions in motion and their parametric motion models. We now compute the motion vectors for each pixel in a segmented region, based on its motion model, and average them for the whole region. The averaged motion vector is the motion vector corresponding to the entire region. The advantages of using a region motion vector rather than using the parametric model are:

1. It simplifies a 6 (affine) or 8 (homography) parameter motion model to a translational motion model. The region motion vector description is approximately accurate for the purpose of our activity analysis;
2. Because we use the translational motion model for *tracking* of each region, this will provide a unified motion characterization for each region.

Note that parametric motion models are still needed for motion segmentation. A translational model will not suffice to give good motion segmentation results for non-rigid motion of human body parts.



Figure 3. Estimated region motion vectors. The red arrows are computed motion vectors; the blue masks indicate the segmented moving regions.

4.1. Tracking

We use a strategy of combining motion segmentation with tracking to provide more accurate motion characterization, without sacrificing the ability to self-initialize and robustness to errors.

The advantage of using motion segmentation for motion event detection is that it is self-initiative, i.e., needs no hand labeling of an initial frame telling the system where the motion events will occur. Also, since motion detection is performed at every frame and does not use the results from previous frames, there will be no error accumulation from frame to frame, which usually occurs in object tracking.

On the other hand, a tracking algorithm has the advantage that it can keep track of the motionless objects and subtle motions which can be missed by a motion segmentation algorithm. The problem of applying a tracking algorithm is that it always needs an initial labeling of the regions in order to track them, which is not available in large scale surveillance or the care giving tasks as discussed in this paper. Background subtraction is also not working here. Background subtraction fails in our settings because the motions are complex, the background varies and significant lighting changes occur.

To fully explore the power of both motion segmentation and tracking algorithms in order to benefit the task discussed so far, we propose the following strategy.

First, we assume that motion segmentation results provide reasonable region labeling to initialize the tracking algorithm. This solves the initialization problem for the tracking algorithm. In the system we track all moving regions from outputs of the motion segmentation algorithm within a limited time window. The result of tracking is translational motions of the regions between consecutive frames. The translational motion of region k is estimated by:

$$[t_x, t_y]_k = \min_{d_1, d_2} \text{Err}(k; d_1, d_2), \quad (1)$$

$$\text{Err}(k; d_1, d_2) = \sum_{M_k(i,j,t)=1} |I_t(i, j) - I_{t+1}(i + d_1, j + d_2)|. \quad (2)$$

$I_t(i, j)$ is the image intensity of frame t at pixel (i, j) . M_k is the mask of region k , i.e.,

$$M_k(i, j, t) = 1,$$

iff the pixel at (i, j) in frame t belong to region k .

Due to errors in motion segmentation, for the same moving object in the scene, motion segmentation results of this object from each frame may be slightly different. If both the scene and motion are rigid, we would expect the final tracked region to converge to several moving objects. However, for non-rigid motion of a person in the scene, as discussed in this paper, the scenario is more complex.

The human body possesses many degrees of freedom. At one point a motion segmented region corresponds to a lower arm while at the next point the nearest segmented region may correspond to a whole arm or even the whole body. If the motion segmented regions constantly change and do not correspond to specific objects, there are different strategies to solve the problem.

The first strategy is based on tracking each segmented region for a long period, attempting to infer a consistent segmentation of body parts. Some a priori knowledge of human body structures is needed in this process.

In the second strategy, rather than inferring about body part locations, we can just detect dominant motions within the human body area, and use these as features to give a description of the human activity in the scene.

In this paper, we use the second strategy. The problem is how to define the dominant motions. There are usually several segmented moving regions in one frame, and simply using the magnitude of motions to define which region is conducting dominant motion is not reasonable. Our solution is to leverage this problem by using a temporal consistency constraint.



Figure 4. Tracked and motion segmented regions. The red arrows are motion vectors estimated from motion segmentation. In subsequent frames, these regions are tracked (yellow arrows) to compensate slight motions of these regions.

4.2. Weighted sequential projection

We use a *weighted sequential projection* (WSP) algorithm to find temporally consistent motions. The algorithm is illustrated in Fig. 5.

As discussed above, for each segmented moving region, its motion vector is obtained by tracking or motion segmentation. The WSP algorithm then computes inner product of the motion vectors for this region in consecutive frames.

$$x_t = v_{t-1} \cdot v_t, \quad (3)$$

where v_{t-1} and v_t are motion vectors computed at frames $t-1$ and t , respectively. The result of the correlation depends on magnitudes of two motion vectors and the angle between them. We then use a sigmoid function to normalize the result between $[0,1]$:

$$ncr_t = \frac{1}{1 + e^{-x_t}}, \quad (4)$$

and summarize the sequential projections within several frames:

$$SP_t = \sum_{i=t-m}^{t+m} (c_{i-t} * ncr_i). \quad (5)$$

c_i 's are weighting constants, with

$$\sum_{i=-m}^m c_i = 1, c_i = c_{-i}, c_i > 0, \quad (6)$$

and

$$c_i > c_j, \text{ if } |i| < |j|. \quad (7)$$

We mark v_t as a consistent motion vector, if

$$SP_t > Th, Th \text{ is a threshold.} \quad (8)$$

A result of this algorithm is given in Fig. 6. In Fig. 6, we approximate the region boundary by rectangles, and overlay them on the images. Tracked regions are not overlapping with each other completely, due to motion segmentation error at different frames, and non-rigidity of human body motion.

The WSP algorithm provides a filtering mechanism to find only consistent motions, while filtering out those non-consistent, or "noisy" motions. In this sense, our purpose and motivation is similar to Wixson [11], which shows effectiveness of consistency detection in motion analysis.

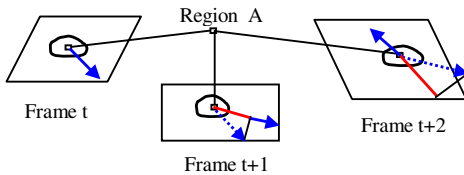


Figure 5. Sequential projection.

By applying the algorithm, at one time, only the regions moving consistently will be kept and added to the candidate dominant motion list, while the other moving regions which are showing contradictory motions or not active enough are simply deleted. While we delete motion

regions from time to time, the motion segmentation algorithm will constantly provide newly segmented regions, and these regions are filtered by the WSP algorithm to keep only consistent motion regions.



Figure 6. Examples of temporally consistent moving regions found using the sequential projection algorithm. Several parallel motion vector arrows indicate the motion direction of several regions overlapping or nearby with each other.

5. Identify activity in a dining room

To identify activities in a dining room, we design an algorithm to find event related motions. There are three steps in this algorithm. First, find individual persons; second, analyze the motion subspace within the region of an individual person, and finally characterize the relative motions.

Let $M(x, y, t)$ be a binary mask indicating all regions of motion in a frame t , i.e.,

$$M(x, y, t) = 1 \quad (9)$$

indicates the pixel at (x, y) in frame t is in a moving region. Then the individual person regions are obtained by accumulating $M(x, y, t)$ over time. Let

$$Cluster_\tau(x, y, t) = \bigcup_{i=0}^{t-1} M(x, y, t-i). \quad (10)$$

τ is the temporal duration. If τ is large enough, $Cluster_\tau(x, y, t)$ will approximately indicate the individual person regions at time t . Typically, we select τ for around 2 minutes. A result is given in Fig. 7.

Within each individual region, we define a head-hand model as shown in Fig. 8. At the present stage, this model only describes individual dining persons sitting frontal or half frontal relative to the camera.



Figure 7. Finding individual person. (a) A frame of the video. (b) The individual person regions obtained using our method (using 2 minutes duration).

We use a face detection algorithm (Schneiderman [4]) to find the face of each person in individual regions, and track the detected face over time. The other two components (corresponding to 2 arms/hands) in the head-hand model are detected based on temporally consistent motions found by our algorithm. At each moment, we assign the 2 regions with consistent motions to the two arms/hands (If there are more than two regions in the candidate consistent region list, we select the regions which overlap with most other regions in the list). The motion vectors of these two regions are then mapped to the main axes between the head and the hands, as shown in Fig. 8. We use the projected distance change on these axes to characterize the dining activity of each individual person.

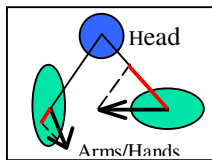


Figure 8. A head-hand model for individual person. The black arrows are motion vectors of 2 major arm/hand components. They are mapped to the head-hand axes to characterize the dining activity of the person. The red lines indicate normalized motion vectors used to characterize dining activities.

6. Experimental results

We show activity curves obtained using our algorithm and also give a result of motion segmentation.

6.1. Motion segmentation result

First we show a result of our RANSAC based motion segmentation. Fig. 9 is an example result for the CMU Motion of Body (MoBo) database. In this example motion of body parts are modeled using six parameter affine models. The masks (blue) represent segmented moving regions resulting from motion segmentation.

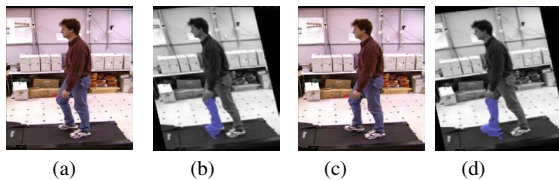


Figure 9. Data from MoBo database. (a), (c): 2 frames in the video. (b), (d): Segmented moving regions (indicated by blue masks). The masks are not overlaid on the current frame, but on the next frame warped back using the estimated motion models, to show accuracy of the models.

6.2. Activity analysis

We show results on 30 minutes of videos for 10 patients at a nursing home. Some of the segmentation and tracking results at this nursing home have been given in Fig. 3-4 and Fig. 6. Fig. 3 shows both the segmented moving regions and their motion vectors; Fig. 4 shows the segmented regions and tracking of these regions; Fig. 6 shows temporally consistent moving regions and their motion vectors after applying sequential projection filtering.

In this section, we compare the activity curves obtained by using combined motion segmentation, tracking and temporal filtering with those estimated using only motion segmentations.

Fig. 10 and Fig. 11 give results for one patient in a video of over 2 minutes duration. In Fig. 10 the motion vectors of two arm/hand components are obtained from the motion segmentation result only, and the curves show magnitudes of the motion vectors mapped to the main axes between the head and hands. In Fig. 11, motion vectors of two arm/hand components are obtained from combining motion segmentation and tracking, and only those temporally consistent motion vectors are mapped on the main axes between the head and hands.

In both figures, the ground truth is labeled using magenta shadings on the time axes. We only care about the motions with directions toward the head, which correspond to positive velocity values on the graphs.

It is obvious to observe that the result from combined motion segmentation and tracking is much more robust to random motions (gives fewer false alarms), and in some cases also can improve the detection rate, partly because tracking is more sensitive to more subtle motions which are more difficult to detect using motion segmentation.

The improvements also depends on patients. In table 1 we give a quantitative result on 10 patients in a dining room, captured on different days. Total length of the videos is 30 minutes.

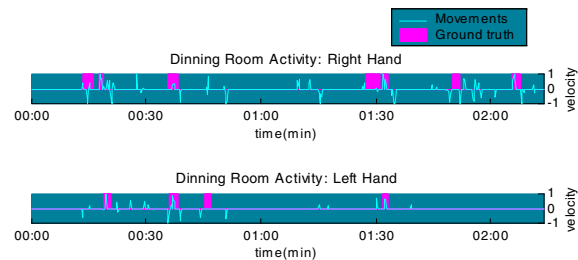


Figure 10. Result based on motion segmentation. The curves show motion of right and left hands mapped to the head-hand axis, with toward-head ground truth labeled.

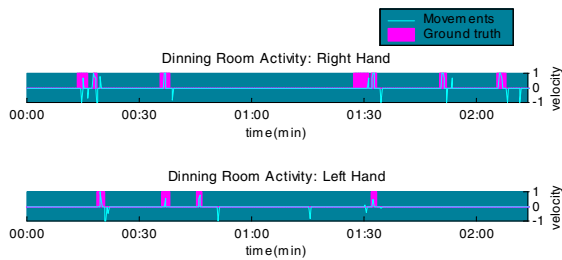


Figure 11. Result by combining motion segmentation and tracking. The curves show motion of right and left hands mapped to the head-hand axis, with toward-head mapped to ground truth labeled.

Table 1. Dining room activities analysis results.

| | Correct Detection | Miss Detections | False Alarms |
|--|-------------------|-----------------|--------------|
| Only Motion Segmentation | 43 | 13 | 39 |
| Motion Seg + Tracking+Temporal Consistency | 50 | 6 | 9 |

7. Conclusions

We developed a new motion feature for describing certain human activities from surveillance videos. Through a combination of motion segmentation and region tracking techniques, while filtering for temporal consistency, we balance the accuracy against the reliability of the motion feature extraction. Motion segmentation allows the algorithm to be self-initializing, and avoids following spurious tracks. Tracking allows the system to be more sensitive to subtle motions, thus improving the accuracy of the overall system for human activity analysis, as well as ensuring consistency in the analysis over time.

Preliminary experimental results show this approach to motion feature analysis can successfully detect activity levels for patients in a dining room. Currently, we are using a threshold with distance between head and hands to further reduce false alarms. Movements far away from the head are not detected as eating motions. We are now working on using a HMM model as an elegant approach to combine both motion and distance measurements between moving regions.

In the long term, through video monitoring in nursing homes, complete and continuous patient activity records can be captured. Using the proposed algorithm, the totality of the automatically analyzed record over time, and the estimates of relative variations within patients may provide useful assistance to caregivers. With activity

analysis methods such as the one proposed in this paper, these records can be transformed into an information asset that empowers geriatric care specialists with greater insights into problems, effectiveness of treatments, and determination of environmental and social influences on patient behavior over time.

8. Acknowledgements

This material is based on work supported by the National Science Foundation (NSF) under Grant No. IIS-0205219.

9. References

- [1] Y. Altunbasak, P.E. Eren and A. M. Tekalp. Region-based parametric motion segmentation using color information. *Journal of Graphical Models and Image Processing*, 60(1):13-23, 1998.
- [2] A. Bissacco, A. Chiuso, Y. Ma, S. Soatto. Recognition of human gaits. *Proc. IEEE conf. on Computer vision and Pattern Recognition*, 2001, (II:52-57).
- [3] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Commun. Assoc. Comp. Mach.*, vol. 24: 381-395, 1981.
- [4] H. Schneiderman, T. Kanade. A statistical method for 3D object detection applied to faces and cars. *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.
- [5] T. Starner, J. Weaver and A. Pentland, Real-time American sign language recognition using desk and wearable computer based video. *IEEE Trans. PAMI*, vol. 20, no.12, 1998.
- [6] C. Stauffer and W.E.L.Grimson. Learning patterns of activities using real-time tracking. *IEEE Trans. PAMI*, 22(8): 747-757, 2000.
- [7] H. Tao, H.S. Sawhney, and R. Kumar, Dynamic layer representation with application to tracking. *Proc. IEEE conf. on Computer vision and Pattern Recognition*, 2000, (II:134-141).
- [8] J. Wang and E. Adelson. Representing moving images with layers. *IEEE Trans. on Image Processing*, 1994.
- [9] Special Section on Video Surveillance. *IEEE Trans. PAMI*, 22(8), 2000.
- [10] A. Wilson and A.F. Bobick, Parametric hidden Markov models for gesture recognition. *IEEE Trans. PAMI*, 21(9), 1999.
- [11] L. Wixson. Detecting salient motion by accumulating directionally-consistent flow. *IEEE Trans. PAMI*, 22(8): 774-780, 2000.
- [12] L. Zelnik-Manor and M. Irani. Event-based analysis of video. *Proc. IEEE conf. on Computer vision and Pattern Recognition*, 2001.