

Articulated Motion Modeling for Activity Analysis

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

Abstract

We propose an algorithm for articulated human motion segmentation that estimates parametric motions of body parts and segments images into moving regions accordingly. Our approach combines robust optical flow estimation, RANSAC, and region segmentation using color and Gaussian shape priors. This combination results in an algorithm that can robustly estimate and segment multiple motions, even for moving regions with small support and in low-resolution images. Based on the raw motion segmentation, consistent body motions are detected over time to characterize human activity. The effectiveness of this approach is demonstrated in a real scenario: characterizing dining activities of patients at a nursing home.

1. Introduction

Much recent research has been focused on *activity analysis* in videos. Several different levels of feature analysis have been proposed, ranging from spatio-temporal histograms to object tracking. While *object tracking* is able to analyze more subtle activities, it is also more sensitive to noise, and in many cases needs manual initialization.

For human *activity analysis*, we are developing an effective motion feature that is *self-initializing* and specialized for detecting *human motion*, or motion of body parts. In particular, *articulated motion* is regarded as a good approximation of the motion of human body parts. In this paper, we present a robust algorithm to segment *articulated motions* in video, and apply the algorithm to analysis of dining activity at a nursing home, as shown in Fig. 1(a).

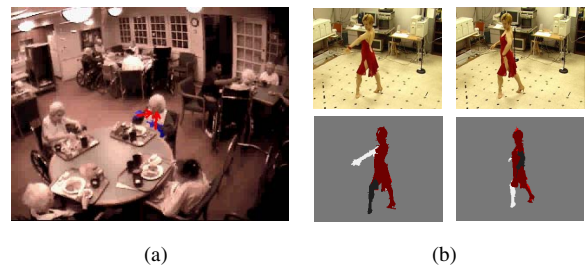


Fig.1. (a) Activity analysis at a nursing home. Arrows indicate directions of movement. (b) Articulated motion segmentation of a dancer in our lab. Both sets of results are produced by our algorithm.

1.1 Previous work

1.1.1 Motion segmentation

The goal of *motion segmentation* is to find the major motions in an image sequence, and segment the images into moving regions accordingly. *Layered representation* of video is a popular approach for motion segmentation (e.g. Wang and Adelson (1994)). In this approach, images are represented by several “layers”, together with their motion models. The *layered representation* needs both estimation of motion models and segmentation of the image into regions. The EM algorithm is a popular, iterative solution method for this task.

In Tao (2000), a general framework is developed for tracking objects through multiple frames using a dynamic layer representation. Pixels belong to objects are represented by layers, and the apparent 2D motion of each layer is estimated by a parametric model. Our goal is similar, in that we also want to find 2D layers corresponding to different moving objects. The difference is that we are dealing with non-rigid motion of articulated human body parts.

We estimate articulated motions in images using a combination of optical flow and RANSAC. RANSAC is able to robustly detect and label natural human motions, even with small support regions for each body part.

Our method has three steps: 1. Estimate *articulated motion* models; 2. Form layer assignments based on these regions; and 3. Optimize the layer representation using an EM-like iteration. In Section 4, output of this algorithm is used to analyze activities at a nursing home.

1.1.2 Activity analysis

Features at different levels have been proposed for human *activity analysis*. In Stauffer and Grimson (2000), a stable, real-time outdoor tracker is proposed, and high-level classifications are based on blobs and trajectories output from this tracking system. In Zelnik-Manor and Irani (2001), dynamic events are regarded as long-term temporal objects, and spatio-temporal features at multiple temporal scales are derived and utilized. In Starner (1998), skin color detection and moments of blobs are used as features to recognize sign languages.

Many approaches to analyzing human body part actions are based on tracking the body as a kinematic linkage. Model-based kinematic tracking of a walking person was pioneered by Hogg (1983), and other influential approaches in this area include Bregler (1998), and Cham and Rehg (1999). These approaches are often brittle, since the human body has many degrees of freedom that cannot be observed well in a 2D image sequence. For this reason, approaches based on body fitting across multiple, simultaneous camera views have been somewhat more successful (Rehg (1995), Gavrilla (1996), and Cheung (2000)). Two good survey papers on the state-of-the-art in vision-based motion capture are Gavrilla (1999) and Moeslund (2001).

Kinematic tracking programs are typically initialized by hand, and fail after processing only a few seconds of video. For fully automated processing, a method is needed to bootstrap the detection of body pose configurations. In contrast, we have developed a self-initializing approach to body part detection based on segmented motion regions.

Other approaches to activity analysis are based on pattern analysis of spatio-temporal representations. Niyogi and Adelson (1994) delineate a person's limbs by fitting deformable contours to patterns that emerge from taking spatio-temporal slices of the XYT volume formed from an image sequence. There have been

several papers on using moments to discover body pose and classify activities. Both Rosales and Sclaroff (2000) and Brand (1999) learn to distinguish between human body poses in a monocular image sequence based on moments of binary silhouettes. Bobick and Davis (2001) combine a sequence of silhouettes into a single *motion history image* (MHI), and use invariant Hu moments to classify actions from these temporal representations.

In this paper, our application is to classify nursing home patient activity levels and their variations over time. We are interested in features that can differentiate and quantify the level (e.g. frequency, duration, magnitude) of these activities. It is a difficult application. Background subtraction fails in our setting because the motions are complex, the background varies and significant lighting changes occur.

We are developing a self-initializing feature based on articulated motion segmentation. The feature lies between the detailed features needed for human gesture recognition and more coarse-level spatio-temporal features commonly used for long range activity classification.

The main contributions of this paper include: 1. Use RANSAC to estimate multiple motion models from flow vectors simultaneously, without pre-segmenting the moving objects; 2. Segment regions based on a Gaussian shape prior; 3. Find consistent motions by combining segmentation with tracking and a weighted sequential projection (WSP) algorithm; 4. Apply motion segmentation to a real scenario in activity analysis.

The organization of this paper is as follows: Section 2 presents our method for articulated motion segmentation. Section 3 discusses our strategy of finding consistent motions. Section 4 describes the system for analyzing dining activities and gives experimental results. Section 5 concludes the paper.

2. Articulated motion segmentation

The problem of articulated motion segmentation is illustrated in Fig. 2. Suppose there are several objects in the scene, and we have no a priori knowledge about their appearances (color, texture, etc). However, we know that the objects are moving differently. Is there a way to segment the individual objects while simultaneously estimating their motion?

The example in Fig. 2 illustrates both *translational motion as well as articulated motion* of a linked structure, which is a good approximation of human body part motion. In this example, there are 3 parts conducting 3 different motions at the same time.

Simultaneously estimating these different motion models, without first segmenting the parts, is a challenging problem.

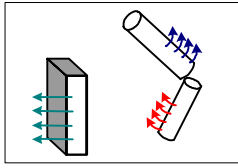


Fig. 2. An example of articulated motions in an image.

We model the motion of each part with a parametric model of 2D image motion. The choice of the parametric model ranges from translational (two parameters), to prismatic (four parameters), to affine (six parameters), and homography (eight parameters). We use affine and homography motion models in our experiments. In the following, we describe our approach based on an affine motion model.

Affine motion is parameterized by six parameters:

$$p_x(x, y) = a_0 + a_1x + a_2y, \quad (1)$$

$$p_y(x, y) = a_3 + a_4x + a_5y. \quad (2)$$

At each pixel (x, y) , $p_x(x, y)$ and $p_y(x, y)$ denote the x and y components of velocity respectively, and the coefficients a_k are the affine motion parameters.

To extract image patch motion, we need to establish a sufficient number of equations to solve for the motion parameters. For this purpose, we propose an adaptive algorithm to estimate optical flow (Section 2.1.2) and then segment the flow vectors corresponding to different object motions using *RANSAC* (Section 2.1.1). After obtaining an initial estimation of parts and their motions, we initialize a layer and motion model for each part (Section 2.2), and then iteratively optimize the motion models and layer supports.

2.1 Simultaneous motion model estimation

Suppose we have already obtained the optical flow estimation. To estimate multiple motion models from flow vectors simultaneously, without pre-segmentation of moving objects, we use an elegant hypothesis and test technique called *RANSAC* (Fischler and Bolles (1981)). As far as we know, our work is the first application of *RANSAC* to articulated motion segmentation as described in this paper.

2.1.1 RANSAC

The *RANSAC* algorithm proceeds as follows: First, the motion model parameters are estimated from a minimum set of flow vectors, sampled randomly. For

the affine motion model, a minimal sample contains 3 flow vectors. The equation to estimate the parameters $\vartheta = (a_0, \dots, a_5)^T$ is:

$$\mathbf{v}(\mathbf{x}) = \Psi(\mathbf{x})\vartheta, \quad (3)$$

where

$$\Psi(\mathbf{x}) = \begin{bmatrix} 1 & x & y & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x & y \end{bmatrix}, \quad (4)$$

$\mathbf{x} = (x, y)^T$ is a vector of pixel coordinates in the image plane, and $\mathbf{v}(\mathbf{x}) = (v_x(\mathbf{x}), v_y(\mathbf{x}))^T$ is the flow vector at (x, y) .

We then compare the estimated motion model at each pixel with flow vectors in image plane, and mark those flow vectors supporting this model (with error below a threshold) as “inliers”. A count is made of the number of inliers. In the next iteration, we sample another minimal set of flow vectors, and go through the same process as above. The process repeats until a sufficient number of samples has been explored (Eq.(5)).

After a sufficient number of samples, the estimated motion having the maximum number of inliers is proposed as a body part motion, because it has the most support from the data. All inliers corresponding to this motion model are then removed from the image, and *RANSAC* proceeds to find another motion model for the *remaining* flow vectors. This process continues until there are not enough remaining flow vectors to support a new model.

More recently, *MLESAC* (Torr and Zisserman (1996)) was proposed to use the same procedure as *RANSAC* but replacing “counting” of inliers by a log-likelihood error metric that also considers the distance of each inlier to the model. In this paper, we use the error metric of *MLESAC*.

An important parameter for *RANSAC* algorithm is the number of samples needed before *proposing* a new model. Assume p is the number of flow vectors in each sample, and ε is the fraction of outliers in the data set. The number of samples m is typically chosen according to the following formula, which guarantees that the probability of a “good” sample being seen is greater than P :

$$P = 1 - (1 - (1 - \varepsilon)^p)^m. \quad (5)$$

Fig. 3 shows examples of motion model estimation using *RANSAC*. Even with the low resolution images (320 by 240) and noisy optical flows shown in Fig.3, *RANSAC* still can find multiple major motions.

RANSAC and similar “hypothesis and test” frameworks are well known for good robustness to outliers, while sacrificing accuracy, which means there is a trade off between robustness and variance. As

described at the end of Section 2.2, the final layer representations are optimized by an EM-like iteration, which serves to improve the accuracy of both motion models and layer assignments.

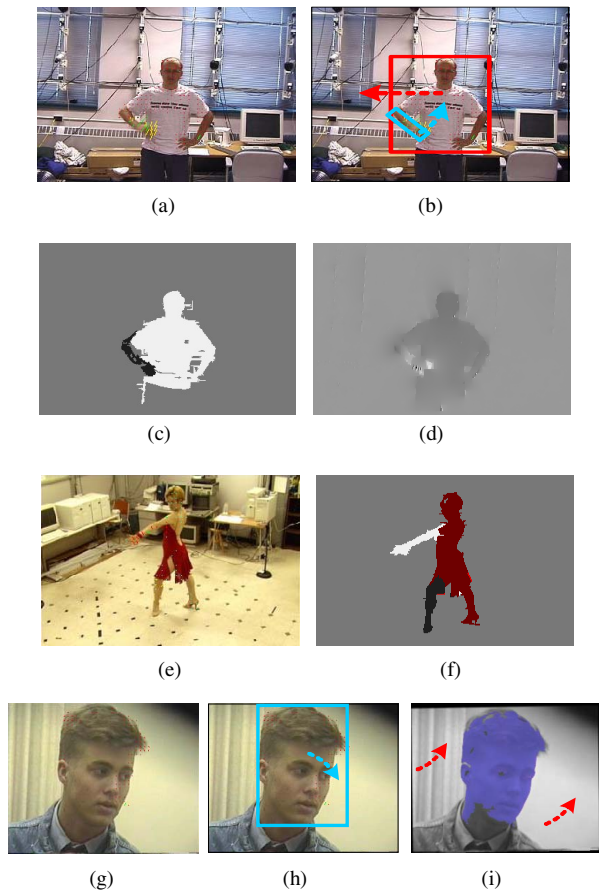


Fig. 3. Applying RANSAC to optical flow vectors. (d) shows a dense flow field, with pixel intensity being proportional to the magnitude of the flow. (a),(e),(g): Flow vectors in different colors are inliers supporting different models, while green represents outliers. (b),(h): Motion models computed from the inliers. (c),(f),(i): Layer assignment based on estimated motion models. In (i) we re-warp the next frame as indicated by the red arrows to form a layer mask (blue) in the current frame.

2.1.2 Robust optical flow estimation

Methods for estimating optical flow from images were reviewed in Barron (1994). Our algorithm is based on an adaptive window matching strategy. For human motion images, multiple image patches within skin or clothing regions may be texture-less, and optical flow estimation in these patches would be erroneous. We have designed an algorithm to select a window size adaptively in order to guarantee that there

is enough texture in the matching windows, and to avoid estimating optical flow at pixels with no surrounding texture.

Let $w(x, y)$ be a matching window centered around pixel (x, y) , with width $(2 * w_H(x, y) + 1)$ and height $(2 * w_V(x, y) + 1)$, respectively. A Sobel edge filter is applied first to compute horizontal (vertical) gradients $S_{H(V)}(x, y)$ in the whole image. Total texture change (gradients) within $w(x, y)$ is computed by

$$G(x, y)_{H(V)} = \sum_{m=x-w_H(x,y)}^{x+w_H(x,y)} \sum_{n=y-w_V(x,y)}^{y+w_V(x,y)} S_{H(V)}(m, n). \quad (6)$$

Based on this measure, the window size is adapted as follows:

$$\text{While } G(x, y)_{H(V)} < h_G \text{ and } w(x, y)_{H(V)} < h_w, \quad (7)$$

$$w(x, y)_{V(H)} = w(x, y)_{V(H)} + 1;$$

$$\text{If } w(x, y)_{H(V)} > h_w \text{ and } G(x, y)_{H(V)} < h_G, \quad (8)$$

skip pixel (x, y) .

Eq. (8) means we only estimate optical flows at pixels with enough texture surrounding them. h_G is the minimum gradient total required in the window, and h_w is the maximum window size. An example is shown in Fig. 4.



Fig.4. Adapting size of matching windows based on total horizontal (a) and vertical (b) gradients in the window.

2.2 Layer assignment

To initialize accurate boundaries for layer assignment, we use a *region-based* approach. We first over-segment the images using the color segmentation algorithm described in Comaniciu and Meer (1997). Fig.5 shows an example.



Fig. 5. Region over-segmentation based on color.

Due to similar colors of skin and clothing across body parts, over-segmentation based only on color can connect body parts together, which means these body parts cannot be assigned to different layers, even

though they conduct different motions. To deal with this problem, we add a constraint that each segmented region assumes a roughly elliptical shape, as represented by a segmentation prior based on a 2D Gaussian distribution (Fig. 6(a)), i.e.

$$\alpha_k(\mathbf{x}) = \frac{1}{2\pi \cdot \sqrt{|\mathbf{C}_k|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{m}_k)^T \mathbf{C}_k^{-1}(\mathbf{x} - \mathbf{m}_k)\right\}, \quad (9)$$

where \mathbf{x} is the coordinate vector of a pixel, and \mathbf{m}_k and \mathbf{C}_k are center vector and covariance matrix, respectively. Considering that human body parts can usually be modeled by ellipses, a Gaussian shape prior is a reasonable choice.

In accordance with the Gaussian shape prior (9), we model each pixel \mathbf{x} within a color-segmented region as a Gaussian mixture with mixture weight w_k :

$$p(\mathbf{x}) = \sum_{k=1}^K w_k \cdot \alpha_k(\mathbf{x}), \quad (10)$$

The parameters in Eqs. (9)-(10) are estimated using k -mean clustering and an EM algorithm. Then, we segment pixels into one of the K sub-regions by the probability that pixel \mathbf{x} belongs to the k -th component:

$$\gamma_k(\mathbf{x}) = \frac{w_k \cdot \alpha_k(\mathbf{x})}{p(\mathbf{x})}. \quad (11)$$

We name the overall algorithm *color and shape* based region segmentation. Note that Tao (2000) also used a Gaussian segmentation prior, but there is no similarity with the strategy and algorithm developed here. We are using the prior in the reversed direction. Fig.7 is an example result of our algorithm. *Shape*-based segmentation is conducted only in foreground regions after global stabilization.

After segmentation of the image into regions, we assign the regions to “layers” by letting estimated motion models compete for each *region*. More specifically, we compute:

$$C(i(R_n)) = \sum_{(x,y) \in R_n} [I(x,y)(t) - I(W_{\vartheta_i}(x,y))(t+1)]^2, \quad (12)$$

where $W_{\vartheta_i}(x,y)$ is the warped position of pixel (x,y) under motion model ϑ_i , and $I(x,y)(t)$ is image intensity at (x,y) in frame t . Using a color constancy assumption, the correct motion model for region R_n should have the minimum cost $C(i(R_n))$:

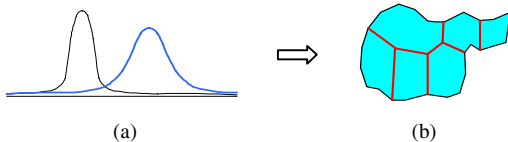


Fig. 6. (a) Gaussian shape prior illustrated in one dimension; (b) A typical region segmentation result based on 2D Gaussian shape priors.

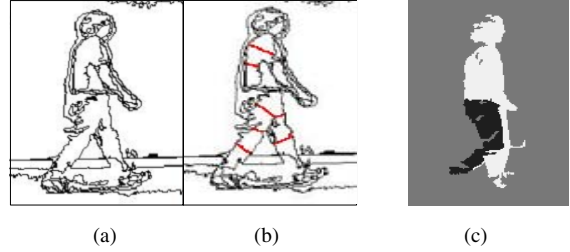


Fig.7. (a) Region segmentation based on color; (b) Region segmentation based on shape and color. We only re-segment human (moving) regions by shape, as illustrated by the red boundaries. (c) By applying shape-based segmentation, legs that are close together but moving differently can be successfully assigned to different layers.

$$i_0(R_n) = \arg \min_i (C(i(R_n))), \quad (13)$$

i.e., we assign motion model ϑ_{i_0} to region R_n .

Now each region corresponds to a motion model, forming an initial *layer representation*. Several EM-like iterations are then applied to re-compute motion models using optical flow within each layer, then reassigning layers based on the new motion models. Accuracy of both motion models and layer assignments are improved in this process.

3. Finding consistent motions

Motion segmentation of optical flow between a pair of frames can only segment the part of the body actually moving at that time instant. An example is shown in Fig. 8. As a result, it is not straightforward to find correspondences of body parts between frames taken at different times.

Different strategies could be used to solve this problem. The first strategy is based on tracking each segmented region for a certain period of time, attempting to infer a consistent segmentation of body parts. Some a priori knowledge of body structure is needed in this process. In the second strategy, we don't attempt to find correspondences of body parts between frames, but rather detect and identify each body part only when it is in motion. In this way, we can just detect dominant motions within specific body part areas, and use these as features to give a description of 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.



Fig.8. (a) At each frame, only some parts of the body (those in motion) can be segmented. (b) In the body area, several different motions may exist simultaneously.

We propose a *weighted sequential projection (WSP)* algorithm to find *temporally consistent* motions. The algorithm is illustrated in Fig. 9.

First, for each moving region obtained by motion segmentation, we compute its *motion vector* as the average motion of pixels within the region:

$$\mathbf{v}_t(R_n) = \frac{1}{N_{R_n}} \sum_{(x,y) \in R_n} \mathbf{v}_t(x,y). \quad (14)$$

where N_{R_n} is the number of pixels in region R_n , and $\mathbf{v}_t(x,y)$ is the motion of pixel (x,y) at frame t , computed by the motion model for region R_n . Fig.11 gives examples of computed motion vectors (red arrows).

We track each segmented region R_n over a limited time window. Assume $\mathbf{v}_{t-1}(R_n)$ and $\mathbf{v}_t(R_n)$ are its motion vectors (obtained from Eq. (14) or from tracking) at frames $t-1$ and t , respectively. We compute the *sequential projection* of $\mathbf{v}_{t-1}(R_n)$ over $\mathbf{v}_t(R_n)$ as:

$$x_{t-1,t}(R_n) = \mathbf{v}_{t-1}(R_n) \cdot \mathbf{v}_t(R_n) / \|\mathbf{v}_t(R_n)\|, \quad (15)$$

and compute the sum within several frames:

$$P_t(R_n) = \sum_{i=t-m}^{t+m} (c_{i-t} \cdot x_{i-1,i}(R_n)). \quad (16)$$

where $2m+1$ is the width of the window, and the c_i 's are weighting constants satisfying:

$$\sum_{i=-m}^m c_i = 1, c_i = c_{-i}, c_i > 0; c_i > c_j, \text{ if } |i| < |j|.$$

We then mark $\mathbf{v}_t(R_n)$ as a *consistent motion vector*, if $P_t(R_n) > Th$, where Th is a threshold.

Weighted sequential projection provides a filtering mechanism to find only consistent motions. In this sense, our purpose and motivation are similar with Wixson (2000), which shows the effectiveness of consistency detection in motion analysis. However, the

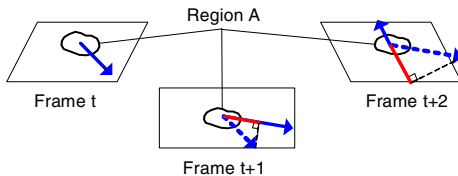


Fig. 9. Finding consistent motions by sequential projection.

strategy and algorithm we developed for this purpose are quite different from Wixson (2000).

4. Experiments

4.1 Articulated motion segmentation

We have provided some pictorial examples of body part segmentation in the previous sections. Here, we measure the accuracy of estimated parametric motion models using our approach.

Fig.10 shows examples from CMU *Motion of Body (MoBo)* database. In this example, motions of body parts are modeled using six-parameter affine models. The masks (blue) represent segmented moving regions (layers). To show accuracy of the motion models, we *reverse-warp* the next frame to the mask in the current frame, using estimated motion models for the masked regions. If the models are not accurate enough, the masks will not overlay well with body parts.

To measure robustness of our approach, we applied articulated motion segmentation to 30 minutes of dining videos captured in a nursing home. Robustness of our motion segmentation algorithm is demonstrated by a high recall rate: near 90 percent of the *relevant motions* (see Section 4.2) are segmented in the low resolution dining videos. (as compared with hand-segmented ground truth). Fig.11 shows some automatically segmented regions with their motion vectors, computed using Eq. (14). In the next section we use these results for dining activity analysis.



(a)



(b)

Fig.10. Two examples from MoBo database. Segmented moving regions are indicated by blue masks. The next frame is re-warped to the mask in the current frame using the estimated motion models, to show accuracy of the models. In (b), two different motion models were estimated, one for each leg.



Fig.11. Articulated motion segmentation in a dining room video. Blue masks indicate segmented moving regions, and red arrows are estimated motion vectors for each region.

4.2 Dining activity analysis

Our goal is to find *eating motions* in dining videos as shown in Fig.11. The estimated activity level of each patient over a long period of time will provide useful assistance to caregivers. Motion detection based on background subtraction fails in our setting, because the motions are complex, the background varies, and significant lighting changes occur.

Our algorithm has three steps: 1. Finding consistent motions in the body area; 2. Detecting faces; and 3. Mapping consistent motion vectors to the *head-hand subspace*. For step 1, our method for finding consistent motions has been described in Section 3. For step 2, we apply the face detection algorithm of Schneiderman and Kanade (2000) to find the face of each person. In step 3, the *head-hand subspace* is defined as shown in Fig.12. Motion vectors of two arms/hands are mapped to the main axes between the head and hands, and projected distance changes on these axes (the red arrows in Fig.12) are used to characterize dining activities of each person. At the present stage, this head-hand model only describes a dining person sitting frontal or half frontal relative to the camera.

We accumulate masks of moving regions in the dining scene to find the outlines of body areas for individual people. An example result is shown in Fig.12(b). We then find two temporally consistent motions (Section 3) in each body area, as the dominant motions corresponding to a person’s two hands/arms. In this experiment, we assign moving regions to right/left hands based on their relative positions with the head. The motion vectors are then mapped to head-hand axis, and *eating motions* are detected by finding the toward-head motions.

Fig.13 plots results for one patient in a video of over 2 minutes duration. The *movement curves* (Cyan) correspond to projected distance changes on the head-hand axis. The ground truth for toward-head motions is labeled as magenta shadings on the time axes.

Compared with using only motion segmentation, consistent motions give much fewer false alarms for *eating motions*, and in some instances can also improve the recall rate, partly because tracking is sensitive to more subtle motions that are hard to detect using only motion segmentation.

Table 1 shows overall results on 30 minutes of dining video for 10 patients, captured on different days. The result has a recall rate of near 90 percent while maintaining an accuracy of over 80 percent.

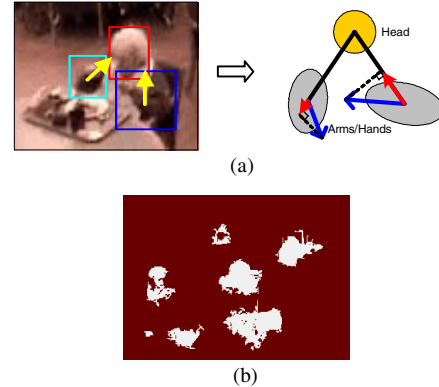


Fig.12. (a)The head-hand subspace for a person. Blue arrows are motion vectors of two arm/hand components. They are mapped to the head-hand axes. Red arrows indicate normalized motion vectors used to characterize eating motions of the person. (b)Individual body areas for the scene in Fig.11, obtained by accumulating masks of moving regions for 2 minutes.

Table 1. Eating motion detection result.

	Correct	Miss	False
Features	Detections	Detections	Alarms
Motion Seg.	43	13	39
Consist. Motion	50	6	9

5. Conclusions

We propose an articulated motion segmentation algorithm, specialized to represent *human motions*, and explore its application to activity analysis in a nursing home dining hall.

Based on an elegant hypothesis and test technique, *RANSAC*, our approach can estimate multiple motions simultaneously from small supporting regions. To achieve layer assignment for human motion with normal clothing, we propose a region over segmentation algorithm combining color segmentation with a *Gaussian shape prior*. The output of motion segmentation is further tracked and filtered to detect *consistent motions*, which approximate *dominant* body part motions.

Dining activity analysis based on this approach achieves a recall rate of near 90 percent for relevant arm motions. To our knowledge, no previous method has been able to achieve a similar performance for this difficult task. The result is especially encouraging, considering *articulated motion segmentation* is a challenging by itself.

We attribute robustness of the algorithm to the combination of the *RANSAC* algorithm, robust optical flow estimation, and region-based layer assignment. This combination enables the algorithm to estimate and segment multiple motions, even with small support for each moving region and in low-resolution images. Our current goal for this line of research is to apply articulated motion segmentation to analyze other daily activities in a variety of natural settings.

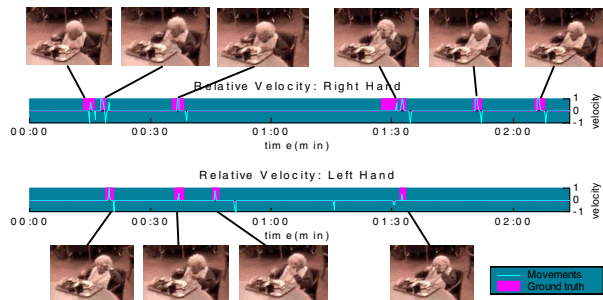


Fig.13. Dining activity analysis based on consistent motions. The movement curves show motions of right/left hands mapped to the head-hand axis, with toward-head ground truth labeled.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) grant IIS-0205219, NSF/RHA grant IIS-0208965, and by DARPA/IAO HumanID under ONR contract N00014-00-1-0915.

References

[1] Barron, J., Fleet, D., and Beauchemin, S., Performance of Optical Flow Techniques, *Int. J. Computer Vision*, 12(1): 42-77, 1994.

[2] Bobick, A.F. and Davis, J.W., Recognition of Human Movement Using Temporal Templates, *IEEE Trans on Pattern Analysis and Machine Intelligence*, 23(3): 257-267, 2001.

[3] Bregler C. and Malik, J., Tracking People with Twists and Exponential Maps, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.8-15, 1998.

[4] Brand, M., Shadow Puppetry, *International Conference on Computer Vision*, pp. 1237-1244, 1999.

[5] Cham, T.J. and Rehg, J.M., A Multiple Hypothesis Approach to Figure Tracking, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.II:239-245, 1999.

[6] Cheung, G.K.M., Kanade, T., Bouguet, J.Y. and Holler, M., A Real Time System for Robust 3D Voxel Reconstruction of Human Motions, *IEEE Conf on Computer Vision and Pattern Recognition*, pp. II:714-720, 2000.

[7] Comaniciu, D. and Meer, P., Robust Analysis of Feature Spaces: Color Image Segmentation, *IEEE Conf. Computer Vision and Pattern Recognition*, 1997.

[8] Fischler, M. and Bolles, R., 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.

[9] Gavrilu, D.M. and Davis, L.S., 3D Model-Based Tracking of Humans in Action: A Multi-View Approach, *IEEE Conf on Computer Vision and Pattern Recognition*, pp.73-80, 1996.

[10] Gavrilu, D.M., The Visual Analysis of Human Movement: A Survey, *Computer Vision and Image Understanding*, 73(1): 82-98, 1999.

[11] Hogg, D., Model-Based Vision: A Program to See a Walking Person, *Image and Vision Computing*, 1(1): 5-20, 1983.

[12] Moeslund, T.B. and Granum, E., A Survey of Computer Vision-Based Human Motion Capture, *Computer Vision and Image Understanding*, 81(3): 231-268, 2001.

[13] Niyogi, S.A. and Adelson, E.H., Analyzing and Recognizing Walking Figures in XYT, *Proc. IEEE Computer Vision and Pattern Recognition*, pp. 469-474, 1994.

[14] Rehg, J.M. and Kanade, T., Model-Based Tracking of Self-Occluding Articulated Objects, *International Conference on Computer Vision*, pp.612-617, 1995.

[15] Rosales, R. and Sclaroff, S., Inferring Body Pose without Tracking Body Parts, *IEEE Conf on Computer Vision and Pattern Recognition*, pp.II:721-727, 2000.

[16] Schneiderman, H. and Kanade T., A Statistical Method for 3D Object Detection Applied to Faces and Cars, *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[17] Starner T., Weaver, J. and Pentland, A., Real-time American Sign Language Recognition Using Desk and Wearable Computer Based Video, *IEEE Trans. PAMI*, 20(12), 1998.

[18] Stauffer, C. and Grimson, W.E.L., Learning Patterns of Activities Using Real-time Tracking, *IEEE Trans. PAMI*, 22(8): 747-757, 2000.

[19] Tao H., Sawhney H.S., and Kumar R., Dynamic Layer Representation with Application to Tracking, *Proc. IEEE conf. Computer vision and Pattern Recognition 2000*, pp. II 134-141, 2000.

[20] Torr P. and Zisserman, A., MLESAC: A New Robust Estimator with Application to Estimating Image Geometry, *Computer Vision and Image Understanding*, 1996.

- [21] Wang J. and Adelson E., Representing Moving Images with Layers, *IEEE Trans. on Image Processing*. 1994.
- [22] Wixson, L., Detecting Salient Motion by Accumulating Directionally Consistent Flow, *IEEE Trans. PAMI*, 22(8): 774-780, 2000.
- [23] Zelnik-Manor, L. and Irani, M., Event-based Analysis of Video, *IEEE Conf. Computer Vision and Pattern Recognition*, 2001.