M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 369
Appears in International Conference on Pattern Recognition, 1996

An Appearance-based Representation of Action

Aaron F. Bobick and James W. Davis
(bobick — jdavis@media.mit.edu)

Abstract

A new view-based approach to the representation of action is presented. The work is motivated by the observation that a human observer can easily and instantly recognize action in extremely low resolution imagery, even imagery in which individual frames provide no information about three-dimensional structure of the scene. Our underlying representations are view-based descriptions of the coarse image motion associated with viewing given actions from particular directions. Using these descriptions, we propose an appearance-based action-recognition strategy comprised of two stages: first a motion energy image (MEI) is computed that grossly describes the spatial distribution of motion energy for a given view of a given action. The input MEI is matched against stored models which span the range of views of known actions. Second, any models that plausibly match the input are tested for a coarse, categorical agreement between a stored motion model of the action and a parameterization of the input motion. Using a "sitting" action as an example, and using a manually placed stick model, we develop a representation and verification technique that collapses the temporal variations of the motion parameters into a single, low-order vector. Finally we show the type of patch-based motion model we intend to employ in a data driven action segmentation and recognition system.

Categories: Human motion understanding; Action recognition; Motion representation

1 Introduction

The recent shift in computer vision from static images to video sequences has focused research on the understanding of *action* or behavior. In particular, the lure of wireless interfaces (e.g. [10]) and interactive environments [7] has heightened interest in understanding human actions. Recently a number of approaches have appeared attempting the full three-dimensional reconstruction of the human form from image sequences, with the presumption that such information would be useful and perhaps even necessary to understand the action taking place (e.g. [19]). This paper presents an alternative to the three-dimensional reconstruction proposal. We develop a view-based approach to the representation of action that is designed to support the direct recognition of the motion itself.

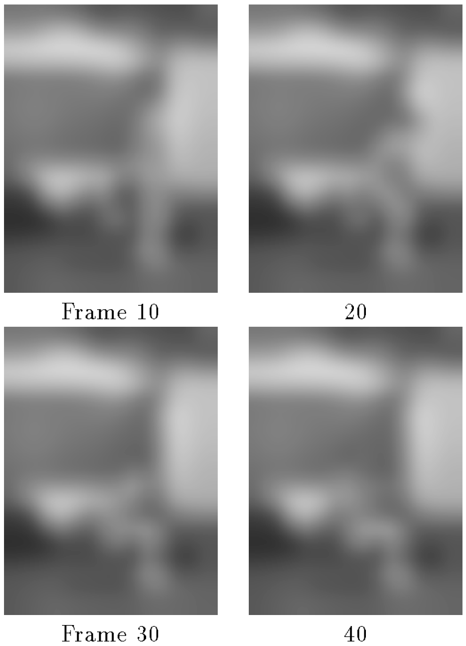

Figure 1: Selected frames from video of someone performing an action. Almost no structure is present in each frame, nor are there any features from which to compute a structural description (as would be in a moving light display). Yet people can trivially recognize the action as someone sitting.

1.1 An observation

The motivation for the work presented in this paper can be demonstrated in a single video-sequence. Unfortunately the media upon which these words are printed precludes the reader from experiencing the impact of viewing the video. A poor substitute is a collection of selected frames of the video, shown in Figure 1.

The video is a tremendously blurred sequence — in this case an up-sampling from images of resolution 15x20 — of a human performing a simple, yet readily recognizable, activity; when shown this video the vast majority of a room full of spectators could identify the action in less than one second from the start of the sequence.[1] What should be quite apparent is that most of the individual frames contain no discernible images of a human being; even if a system knew the image was that of a person, no particular pose could be reasonably assigned.

A more subtle observation is that no good features exist upon which to base a structure-from-motion algorithm[22]. This distinction is important: although individual frames of moving light displays also contain insufficient information from which to recover pose, they do contain features that allow for the structural recovery of the limbs [13] without *a priori* knowledge of the semantic assignments (e.g. "light 1 is the left hip"). One cannot prove that the blurred sequence of Figure 1 cannot be analyzed for three-dimensional structure before the assignment of body parts to image regions. However the lack of any image detail makes such a possibility remote; it would seem that an initial alignment of model to image is necessary.

1.2 Model-based, view-based recognition of action

Given that motion recognition is possible in the absence of features from which to compute three-dimensional structure, how might it be accomplished? To us, the most straightforward answer is that the motion pattern itself is recognized. Much as a two-dimensional, static pattern, say the schematic drawing of a face — a circle, two dots and an arc — is instantly "recognized" as a face, it should be possible to recognize a two-dimensional motion pattern as an instance of a motion field which is consistent with how some known movement appears when viewed from a particular direction. Such a capability requires a view-based, model-based technique. The model, however, is of the motion, not of the body.

In the remainder of this paper we develop a representation of the motion of an action designed to support such an approach. The basic components of the theory are:

1. A motion model to be recognized is a coarse or categorical description of the motion observed when a known movement is viewed from a given angle.

2. Motion recognition is embedded in a simple hypothesis and test paradigm[12] where a data-driven initial computation is used to index plausible motions which are then verified by a more rigorous match.

3. The spatial distribution of motion integrated over some temporal extent of the motion is employed as the initial filter proposing possible actions and viewing directions.

4. A coarse patch model of motion (similar to [3]) is capable of discriminating motions, once the motion energy distribution is used to pre-filter the hypotheses.

[1] The only instruction was: "You are about to see a particular action happening. Raise your hand as soon as you think you know what action is taking place."

Because we are strongly motivated by the capability of people viewing extremely blurred sequences, all of our examples will be developed on blurred image video, though the extent of the blurring is not as extreme as that of Figure 1.[2]

We begin by considering some prior work on both motion recognition and the view-based techniques in object recognition. Next we develop a feature-based characterization of the motion energy image (MEI) to be used as an initial filter into the set of known movements; the strengths and weaknesses of such a choice are considered. We next explore the appropriate parameterization of the motion appearance models. Our eventual goal is to use motion patch deformations similar to [3] but where the patches selected are different for each known motion. To that end, using the sitting action as an example, we develop a representation and verification technique that collapses the temporal variations of the motion parameters into a single, low-order vector. We describe some initial experiments using manually placed and tracked "sticks" as the underlying primitive which shows the effectiveness of the representation. Finally, we describe how the same parameterization can be applied to tracked motion patches and we how to incorporate the representation in a data driven action segmentation and recognition system.

2 Prior work

The number of papers on and approaches to recognizing motion and action has recently grown at a tremendous rate. For an excellent review on the machine understanding of motion see [5]. We divide the relevant prior work into three areas: action recognition, view-based (usually *aspect*) matching, and motion-based recognition.

The first and most obvious body of relevant work includes all the approaches to understanding action, and in particular human action. Some recent examples include [1, 4, 8, 11, 14, 19, 20, 6, 24]. Some of these techniques assume that a three-dimensional reconstruction precedes the recognition of action (e.g. [11]), while others use only the two-dimensional appearance (e.g. [8]). However, underlying all of these techniques is the requirement that there be individual features or properties that can be extracted from each frame of the image sequence. These approaches accomplish motion understanding by recognizing a sequence of static configurations. It is difficult to imagine that such techniques could be extended to the blurred sequence of Figure 1.

The second area related to this work is that of appearance- or aspect-based recognition (e.g. [17, 16, 9]). The formal description of aspects [17] referred to the visible surfaces of objects undergoing self occlusion. For a range of viewing angles, which surfaces are vis-

[2] All the results presented in this paper use a blur kernel corresponding to the second level of a Gaussian pyramid giving an effective resolution of 80 by 60. These dimensions are for the entire image, so the person is only about 20 by 40 pixels. We wanted to use blurred images of the same resolution as Figure 1 but we had difficulty getting stable motion image estimates.

ible surfaces remains constant and only the shape of their projection changes. Ikeuchi and Hong [16] refer to the shape change within an aspect as a "linear shape change." Eggert, et al. [9] extend the use of aspects to include scale by understanding how the visibility of surfaces is affected by the scale of observation. In general, the term "aspect" recognition has come to include any recognition scheme that partitions the view sphere into distinct models. The motion model we will develop attempts to span as wide an angular range as possible using a single, low order representation of the appearance of the motion. However, when not possible, our model can also accommodate discrete regions or aspects.

Finally there is the work on direct motion recognition [18, 21, 23, 3]. These approaches attempt to characterize the motion itself without any reference to the underlying static images. Of these techniques, the work of Black and Yacoob [3] is the most relevant to the results presented here. The goal of their research is to recognize human facial expressions as a dynamic system, where it is the motion that is relevant. Their approach does not represent motion as a sequence of poses or configurations. Except for using some facial features for alignment of motion models, the system is "featureless" measuring only the motion in particular patches of the face, and using a characterization of that motion to determine the facial expression being produced. Our work can be viewed as an extension of the work by Black and Yacoob to the general problem of action recognition.

3 Spatial distribution of motion

In keeping with the hypothesis-and-test paradigm, our first step is to construct an initial index into a known motion library. To avoid exhaustive search we require a data-driven, bottom up computation that can suggest a small number of plausible motions to test further.

Our approach is to separate the consideration of *where* is there motion from *how* the image is moving. In this section we develop a representation of the spatial distribution of motion which is independent of the type of motion present; this characterization will serve as our initial index. A coarse, compact description of the motion pattern developed in the next section will be used to test the selected hypotheses.

3.1 Motion-energy images

Consider the example of someone sitting, as shown in Figure 2. The top row contains several key frames in a sitting sequence. The bottom row displays cumulative motion images — to be described momentarily — computed from the start frame to the corresponding frame above.[3] As expected the sequence sweeps out a particular region of image; our claim is that the shape of that region can be used to suggest both the action occurring and the viewing condition (angle).

To describe the motion pattern we first construct a motion-energy image (MEI) for each training sequence. We have experimented with several methods for determining motion-energy. An obvious approach is to compute optic flow field between each pair of frames using a local, gradient-based technique similar to Lucas and Kanade [2] yielding a vector image $\vec{I}_i(x,y)$ for each sequential pair. The motion energy image is then computed by simple summation:

$$\mathrm{MEI}(x,y) = \sum_i^T \vec{I}_i'(x,y)$$

where $\vec{I}_i'$ is a thresholded version of $\vec{I}_i$ designed to prevent noise in the motion computation from corrupting the process. However, for the blurred images used for this paper we have found that the summation of either the square of image differences, or the square of the images generated by subtracting the first frame from each of the N frames, often provides a more robust motion-distribution signal. This is because motion-differencing is not attempting to determine where pixels have moved but only that they have changed; the results shown in the paper are computed using a binary thresholding of the sum of the squared difference between each frame and the first.

For the training data, the value of T as well as the start of the "action" need to be determined manually. For the initial work presented here we are choosing T to be the length the entire action (about 45 frames). This choice implies that a recognition system using this representation will only be able to recognize sitting after the entire action has been completed. Since we will use the MEI as an index into a library of motion models, we are investigating creating an MEI based only upon initial or incremental phases of a motion. The idea is that an initial detection by MEI could be used to index either the complete or partial motion which would then be verified.

Once we have constructed the MEI for each training sequence we need to compute an average MEI to represent each viewing angle of the action. To do so requires registering MEIs that are of the same action viewed from the same angle but executed by different people. Therefore, to the extent that the MEIs capture the motion of the action, and to the extent that people are roughly the same shape up to a scale factor, the shapes of the MEIs should also be the same, varying only in translation and scale. To register MEIs we compute the standard centroid and moments of each individual MEI, and use these measures to align the motion-energy images. We construct a robust MEI by averaging[4] the aligned images. The robust MEI for sitting from each of 10 viewing conditions ($0°$ to $90°$ in $10°$ increments) is shown in Figure 3. Our task is to use these images as an index for motion recognition.

3.2 MEI feature space

To use the MEI as an index for recognition we need to characterize it. Since the intent is for the MEI to capture the spatial distribution of motion, we select a shape description vector $\vec{m}$ to compare the input MEI of a sequence to the model MEI. Since the MEIs are blob-like in appearance we employ a set of moments-based descriptions. The first seven parameters $< m_1, \ldots, m_7 >$

[3] These motion energy images are computed on the blurred versions of the imagery.

[4] Results using a median are similar.

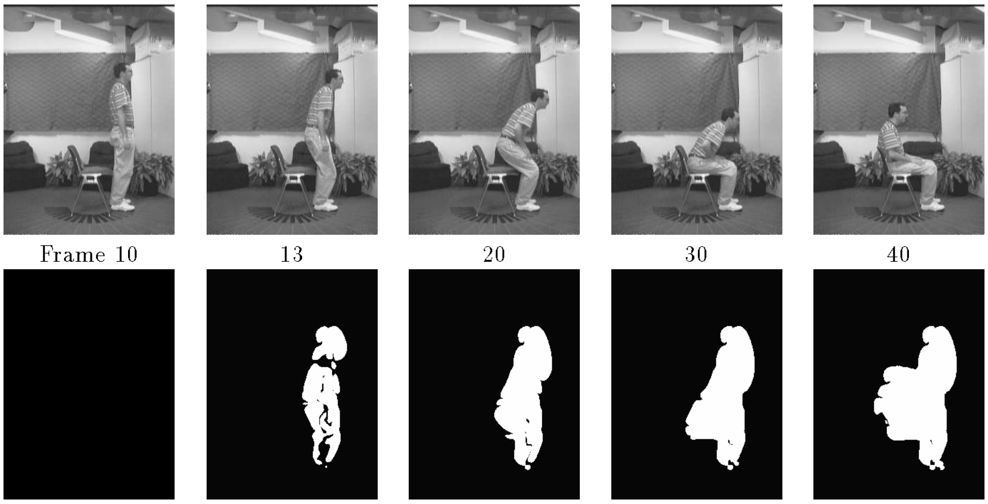

Figure 2: Example of someone sitting. Top row is keys frames; bottom row is cumulative motion images starting from Frame 0. Motion images computed on the blurred imagery.

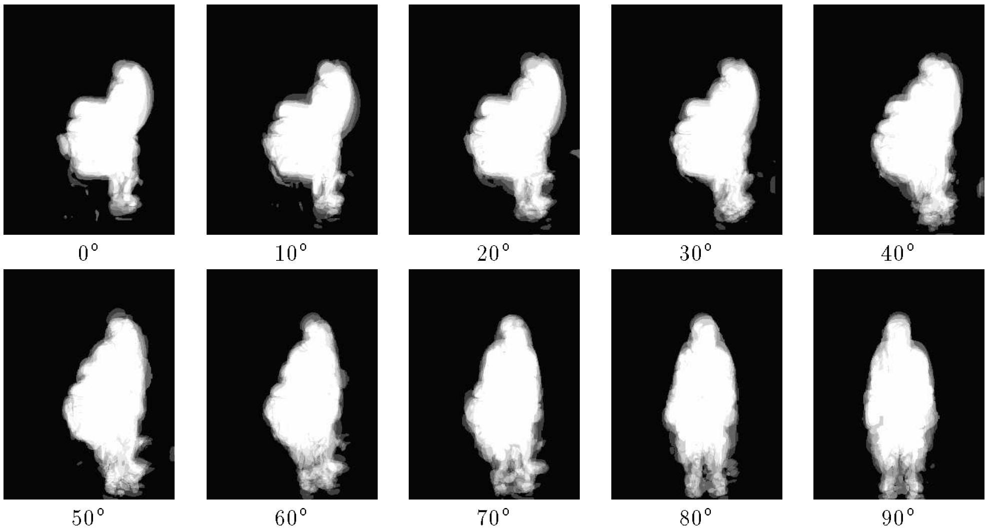

Figure 3: Average MEIs of sitting from 10 viewing angles.

are the Hu moments [15] which are known to yield reasonable shape discrimination in a translation-, scale-, and rotation-invariant manner. Because many of the Hu moments are not sensitive to axis reflection, and because human motion tends to have viewing symmetries, we augment the feature vector to include terms sensitive to orientation and the correlation between the x and y locations: $m_8 = [E(xy) - E(x)E(y)]/[\sigma_x \sigma_y]$. Also, we include a measure of compactness m_9 computed as the ratio of the area of the image to the area of the best fit ellipse whose axes' orientation and relative size are determined by the principal components, and whose overall scale is set to a multiple of the standard deviation of the spatial distribution of the pixels in the MEI.

We should note that this particular choice of shape parameters is ad hoc. Also, we use binary images as opposed to "gray-scale" images for computation of the moments; we have found that doing so makes the computation less sensitive to small variations. However, we make no claim that these parameters are the best method of describing the spatial distribution of the motion energy. For our experiments they were simply adequate; we suspect that greatly differing domains would require different shape descriptors.

To illustrate the effectiveness of the MEI shape descriptions, we performed the following experiment.

First, as described above, we generate average MEIs for all the sitting data. The complete data suite is ten different viewing angles [$0°$ (side view) through $90°$ (frontal view) in $10°$ increments] for 4 people each sitting twice, each time in a different chair. For each angle an average MEI was computed, and shape descriptions were generated.

Next we added to the sitting data the MEIs of 20 different aerobic exercises. We have chosen this domain because 1) there are well defined motions that people know and easily discriminate; 2) the range of motions are broad; and, 3) realistic motions (e.g. sitting, throwing, reaching) are similar to many of the steps. In fact the set includes a simple squat which in many ways is similar to sitting. The complete data set consists of 7 views ($0°$ through $180°$ in $30°$ increments) of each of the 20 movements, with each sequence containing on the order of 100 frames. Therefore, there are 140 MEIs generated by the aerobic data, yielding a total target set of 150.

The experiment consisted of testing sitting examples of a new subject whose data were not included in the robust sitting MEIs; there were 20 examples, 2 repetitions each from $0°$ to $90°$ in $10°$ increments. For each of the 20 inputs, the target MEIs were ranked according to nearness using a metric of independent Mahalanobis distance[5]; a ranking of 1 is the closest element.

The results are as follows: For the 20 input examples, the average rank of the correct sitting example (same angle as input) is 4.0, with a median of 3. This implies that typically the third or fourth best MEI out of 150 would in fact be the correct motion, and that the MEIs would be a good index for a hypothesis and test paradigm were several hypotheses are considered.

Also, if there is some latitude in the verification procedure, then one only needs to find the correct action at a near-by viewing direction. The average ranking of the closest sitting MEI that is within $10°$ of the input move is 1.8, with a median of 1.5, and a worst case of 3. To find a close viewing direction of the correct action, typically only 1 or 2 of the 150 action-angle hypotheses need to be considered.

4 Motion modeling

The last component of the representation is the motion description. The idea is partially motivated by the previously mentioned facial-expression recognition work of Black and Yacoob [3]. Their innovative paper proposed a qualitative description of the motion of predefined patches of the face. The parameterization and location relative to the face of each patch was given *a priori*. The temporal trajectories of the motion parameters were qualitatively described according to positive or negative intervals, and then these qualitative labels were used to recognize such emotions as anger or happiness.

Our work here seeks to extend that approach by using specific motion patch parameterizations for different viewing angles of different movements. Clearly the relevant motion fields to see a person sitting are not the same as those to see someone performing a push-up. How then do we select which areas of the image to consider for motion analysis, and how do we compare an input sequence with a known motion?

Two possible answers should be considered. The first is that perhaps the raw motion signal could be segmented into distinct patches *based solely on the motion image.* While this may be possible for high resolution images, it would be quite difficult for the blurred images we have presented, and almost certainly impossible for images blurred to the extent of sequence in the introduction.

A second answer, and one that we demonstrate here, is that the stored motion models contain the patch parameterization necessary to qualitatively describe the motion. The basic idea is that a stored motion model consists of an index entry, a method for aligning the patch model with the input data, and a description of the motion fields expected for the given action viewed from the given angle. With these tools it will be possible to recognize a given input sequence as being consistent with a known motion.

In this section we derive a motion description mechanism that collapses motion trajectories — the value of the motion parameters as they vary during the performance of an action — to a single, low-order vector. Then, for each angle of each move we can define an acceptance region — or a probability distribution — on that vector for a given view of a given action. If an input motion falls within that region it can be said to be accepted as an example of the action.

4.1 Sitting sticks

To derive our representation we employ a simplified patch model, namely sticks manually placed and tracked on the imagery. We do this to decouple the nature of the representation from our ability to do patch tracking using optic flow or another motion estimation procedure. Figure 4 show the manually placed sticks in five frames of a sitting action.

A stick is defined by its two endpoints $\{<x_1, y_1>, <x_2, y_2>\}$, and therefore has 4 degrees of freedom. As such we can describe the motion of a stick from one frame to another by four numbers. To help maintain our intuitions we will consider the four variations to be Trans-x, Trans-y, Rotation, and Scale, and we relate them in the usual way to the algebraic transformation:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_1 & -a_2 \\ a_2 & a_1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a_3 \\ a_4 \end{bmatrix}$$

where $[x, y]^T$ are the position of an endpoint in one frame, $[x', y']^T$ are the position of the endpoint in the other, Trans-$x = a_3$, Trans-$y = a_4$, Scale $= a_1$, and Rotation $= a_2$.

For the sitting example, as illustrated, we use three sticks. If we relate the three sticks at each time t back to the original stick configurations at time $t = 0$, we can characterize the motion of the sticks by a 12-dimension, time-dependent vector $\vec{M}(t)$. For each given viewing direction α we would get a different motion appearance so we need to index the motion by α: $\vec{M}_\alpha(t)$.

The four graphs of the left-hand column of Figure 5

[5] That is, a Mahalanobis distance with a diagonal Σ.

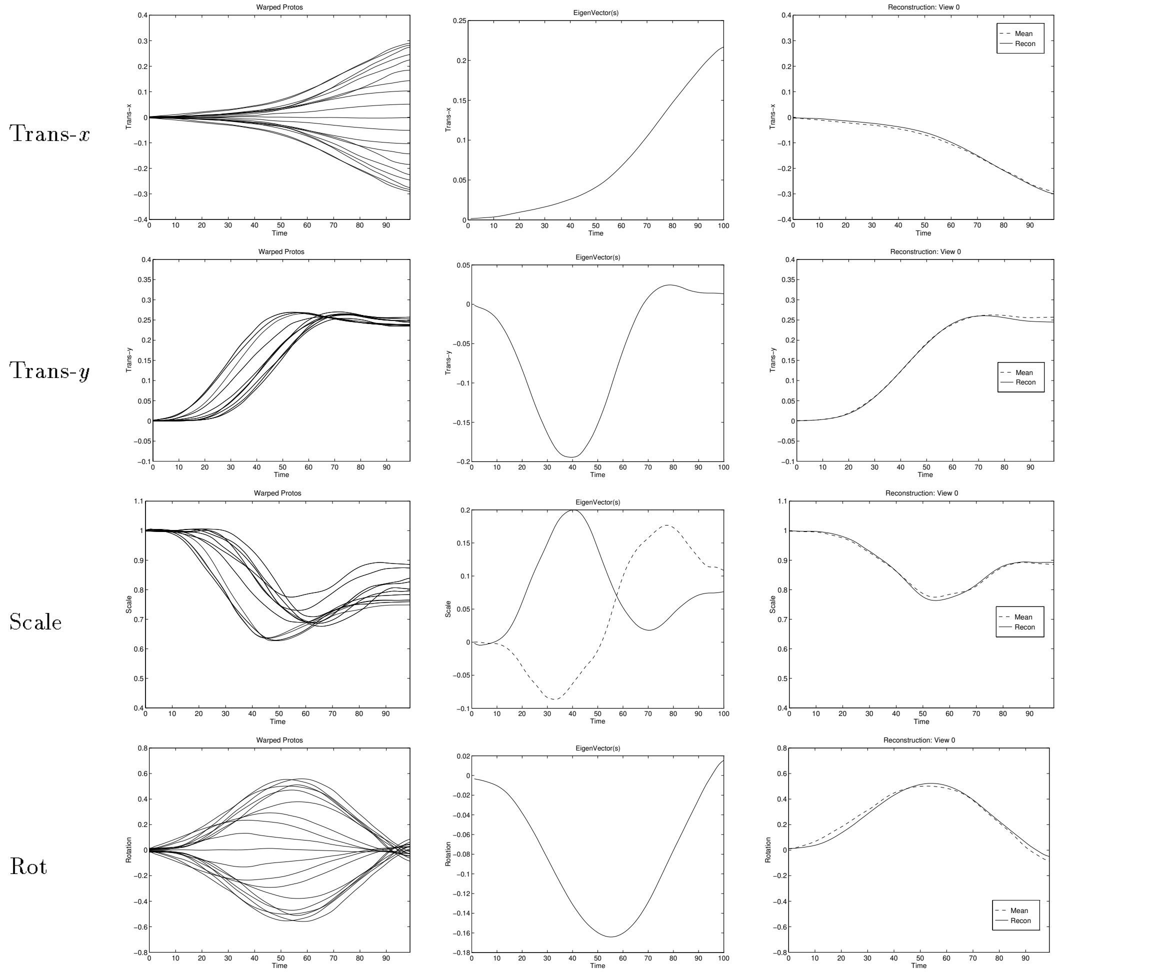

Figure 5: Left: 4 motion parameters for one sitting stick, for each of training viewing angles; Middle: Eigen-functions capturing 90% of the variance. Right: Typical reconstruction of a motion parameter trace.

show the average[6] traces for α every 10° from 0° to 180° for each of the parameters of the torso stick.[7] Note how the curves vary slowly as α changes: since appearance changes slowly with viewpoint so does the parameterization. Highly dependent signals such as these can often be well represented by a principal components decomposition, reducing the data to a linear combination of a small number of basis functions. The second column of the figure shows all the eigen-functions required to capture 90% of the variance of the original instances. Notice that for this stick, only the Scale parameter requires two eigen-functions; the remaining parameters, only a single eigen-function. The right column contains typical reconstructions of the parameters curve using only the eigen-functions shown.

4.2 Variation within training data

To determine whether a test motion is consistent with a known movement we need to characterize the variability of the training data. The above principle component decomposition has the effect of reducing the time-based parameter curve to a low order coefficient vector (in fact a singleton in 3 of the 4 parameters of the stick shown). Therefore we can measure the variation of the coefficients of the training data to determine an acceptance region. Figure 6 displays the mean value of the five coefficients (two for Scale, one for each of the other three) along with a 3σ envelope, where the training data are two repetitions (two different chairs) of four people sitting. The means of the coefficients can be considered as an angle varying vector $\vec{C}^m(\alpha)$.

[6] A dynamic time warping (DTW) of the 12-dimensional signals is done to align the curves of a given angle for each of the four subjects. Because we are using absolute motion, not frame to frame, the amplitude of $\vec{M}_\alpha(t)$ is speed invariant and amenable to scaling by DTW methods.

[7] The angles 100° to 180° are generated by reflecting the original image data.

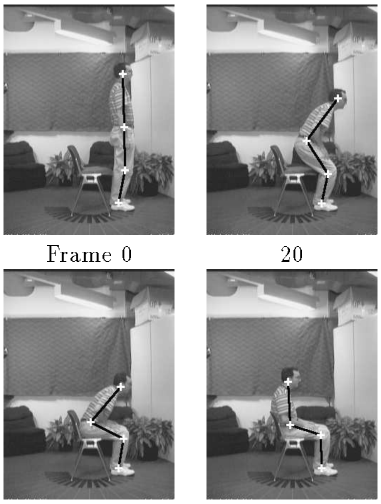

Figure 4: Stick placement for sitting. Sticks are manually placed and tracked on the torso, upper leg, and lower leg.

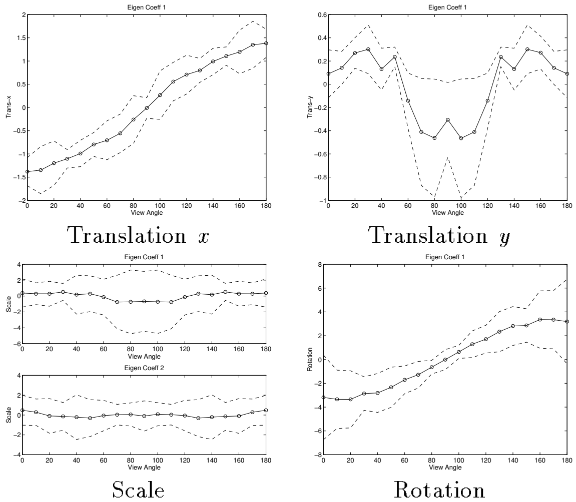

Figure 6: Mean and 3σ acceptance region for one stick of the motion model for sitting as a function of view angle α.

These graphs represent the motion modeling for the sitting action. When testing an input motion, the three sticks need be placed and tracked, and the necessary parameter trajectories recorded. Then, given a hypothesized view angle α_0, the input trajectories are jointly dynamically time warped to match the reconstructed trajectories generated by the eigen-coefficient vector $\vec{C}^m(\alpha_0)$. After warping, the input traces are projected onto the eigen-function basis set to yield the coefficient vector $\vec{C}^{test}$. Finally, the input motion is accepted as an example of sitting motion at angle α if every component of c_i^{test} of $\vec{C}^{test}$ is within the k-σ envelope: $\forall_i, \|c_i^{test} - c_i^m\| < k\,\sigma_i(\alpha_0)$.

To test this approach we performed the following experiment: We first extracted stick parameters for 3 new people, sitting in different chairs, viewed from the 10 viewing angles. We also recorded stick parameters for the 3 aerobics moves that involved full body motion and looked to us to be the most like sitting — the closest example is a simple squat. For a wide range of k, $3.0 \leq k \leq 9.0$, all but one of the sitting examples were accepted, whereas all of the aerobics moves were rejected.[8]

4.3 Patches

While manually instantiated sticks are convenient for deriving our verification method, to actually recognize action we need to automatically recover motion parameters. Our goal is to have a set of polygonal patches whose placement is determined by the hypothesized action and view angle suggested by a matching target MEI. The motion parameters are determined by tracking the patches using a region-based parametric optic flow algorithm. Consistent with the stick approach would be to use a parameterization that models optic flow as a four parameter deformation; a 6- or 8-parameter planar model is possible as well as long as it is stable.

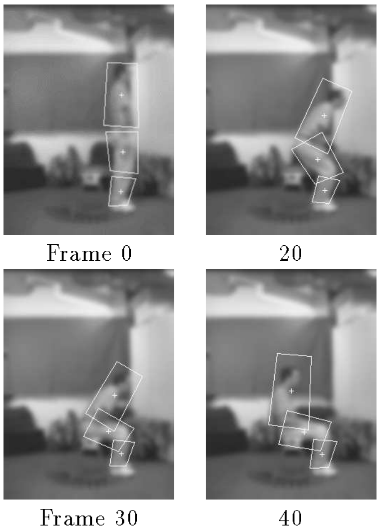

Figure 7: Automatic tracking of patches. After initialization by MEI alignment, the patches are tracked using a parametric optic-flow algorithm; in this case, an affine model is used.

One example of tracked patches is shown in Figure 7. The three polygonal patches are created manually but tracked automatically using an affine model of optic flow[2]. The initial placement and scale of the

[8] The one sitting example rejected was performed by the aerobics instructor performing a sitting action as an exercise. Her perfect posture fell out of the 3σ range — not surprising considering the four training subjects were graduate students.

these patches needs to be adjusted to fit the position and size of the motion in the image. One possibility is to use centroid- and moment- based alignment between the input and target MEIs to define the necessary 3-parameter transformation for the patches. We have not yet achieved a robust enough model placement and tracking algorithm to test the recognition method on patches. Unlike face images of [3], our sitting images can have quite a variety of image textures which makes motion estimation a non-trivial operation.

4.4 Motion aspects

We conclude the section on motion description by noting that it is unlikely that a single patch representation can be robustly tracked over all possible views of an action. Thus, the parameterization of the motion description itself is sensitive to view angle. We refer to the space of view angles over which a parameterization applies as a *motion aspect.*

Fortunately, the hypothesize and test method can be easily modified to use different patch models for different views. Using the MEI as an index, one can not only retrieve the appropriate coefficients $\vec{C}^m(\alpha)$ but also the patch model from which the coefficients are derived. The basic algorithm would be to (1) Use the input MEI to select a target MEI, a patch model, and coefficient vector; (2) Compute a scale and translation transformation between the MEIs; (3) Use the transformation to align the selected patch model with the input sequence; (4) Track the patch model and extract the motion parameters; and (5) Test for acceptance of the action and view angle associated with the selected target MEI.

5 What's next: Segmentation and recognition

Our motion-based recognition technique has been designed to allow for the automatic temporal segmentation of an input video stream. The method relies on looking backward in time, much as is done by [8]. The system will continually construct MEIs backwards, up to a time delay of t_d, and attempt to find a matching MEI in the library. If any matches are found, the corresponding motion appearance model is retrieved and applied backward in time attempting to verify the motion. We have not yet completed implementation of the temporal segmentation mechanism so we cannot report the results. However, the computational complexities of both the MEI computation (mostly image differencing) and the patch parameter estimates (least-squares over small patches) are low, and near real time performance should be possible on standard hardware.

6 Conclusion

A new view-based approach to the representation of action for recognition is presented. The fundamental idea is to recognize the motion itself, not a sequence of static configurations. The paradigm we consider is hypothesize and test. The hypothesis phase requires a model-free method if exhaustive search is to be avoided. We develop motion-energy images (MEIs) as a method of capturing the spatial distribution of motion, and propose a simple shape description feature-vector as the initial index as to the motion and viewing condition present. Once candidates are proposed, they can be verified using model-based techniques. Using a manually placed and tracked stick model we derive a principal-components method for collapsing the time varying motion parameters to a single, low-order, coefficient vector. The coefficients are a function of view angle and can be used to verify agreement with training data. Our next goals are to apply these techniques to the automatic placement and tracking of patches and to then implement the complete recognition system to automatically segment and recognize actions.

We find our initial results promising, but need to experiment further in a domain where we have many known motions. Our intent is to apply the system to the tasks of recognizing a set of aerobic exercises (trained on professionals but tested on out-of-shape academics) and detecting particular actions in live video (such as simply noticing when anyone sits down anywhere in a room).

References

[1] Akita, K., "Image Sequence Analysis of Real World Human Motion," *Pattern Recognition,* **17**, 1984.

[2] Bergen, J., P. Anadan, K. Hanna, and R. Hingorami [1992] "Hierarchical model-based motion estimation," *Proc. European Conference on Computer Vision,* France, 237–252.

[3] Black, M. and Y. Yacoob, "Tracking and Recognizing Rigid and Non-rigid Facial Motion using Local Parametric Models of Image Motion," *ICCV*, 1995.

[4] Campbell, L. and A. Bobick, "Recognition of Human Body Motion Using Phase Space Constraints," *ICCV,* 1995.

[5] Cedras, C. and M. Shah, "Motion-Based Recognition: A Survey," *Image and Vision Computing,* 1993.

[6] Cui, Y., D. Swets, and J. Weng, "Learning-based Hand Sign Recognition Using SHOSLIF-M," *ICCV,* 1995.

[7] Darrell, T., P. Maes, B. Blumberg, and A. Pentland, "A Novel Environment for Situated Vision and Behavior", *Proc. IEEE Wkshp. for Visual Behaviors (CVPR-94),* IEEE C.S. Press, Los Alamitos, CA, 1994

[8] Darrell, T. and A. Pentland, "Space-Time Gestures," *CVPR,* 1993

[9] Eggert, D., K. Bowyer, C. Dyer, H. Christensen, and D. Goldgof, "The Scale Space Aspect Graph," *IEEE Trans. PAMI*, **15**, 11, 1993.

[10] Freeman, W., "Orientation Histogram for Hand Gesture Recognition," *Int'l Workshop on Automatic Face- and Gesture-Recognition,* Zurich, 1995.

[11] Goncalves, L., E. DiBernardo, E. Ursella, P. Perona, "Monocular tracking of the human arm in 3D," *ICCV*, 1995.

[12] Grimson, W. E., *Object Recognition By Computer: The Role of Geometric Constraints,* MIT Press, 1990.

[13] Hoffman, D. and B. Flinchbaugh, "The Interpretation of Biological Motion," *Biological Cybernetics*, **45**, 1982.

[14] Hogg, D., "Model-based vision: a paradigm to see a walking person," *Image and Vision Computing,* **1**, 1, 1983.

[15] Hu, M., "Visual Pattern Recognition by Moment Invariants," *IRE Trans. Information Theory,* **IT-8**, 2, 1962.

[16] Ikeuchi, K. and K. S. Hong, "Determining Linear Shape Change: Toward Automatic Generation of Object Recognition Programs", *CVGIP, Image Understanding*, **53**, 2, 1991.

[17] Koenderink, and A. van Doorn, " The internal representation of solid shape with respect to vision," *Biological Cybernetics,* **32**, 1979.

[18] Polana, R. and R. Nelson, "Low Level Recognition of Human Motion," *IEEE Workshop on Non-rigid and Articulated Motion,* Austin, 1994.

[19] Rehg, J. and T. Kanade, "Model-based Tracking of Self-Occluding Articulated Objects," *ICCV*, 1995.

[20] Rohr, K, "Towards Model-based Recognition of Human Movements in Image Sequences," *CVGIP, Image Understanding,* **59**, 1, 1994.

[21] Shavit, E. and A. Jepson, "Motion understanding using phase portraits," *IJCAI Workshop: Looking at People,* Chambery, 1995.

[22] Ullman, S., "Analysis of Visual Motion by Biological and Computer Systems," *Computer,* August, 1981.

[23] Yacoob, Y. and L. Davis, "Computing Spatio-temporal Representations of Human Faces," *CVPR*, 1994.

[24] Yamato, J., J. Ohya, and K. Ishii, "Recognizing Human Action in Time Sequential Images using Hidden Markov Models," *CVPR*, 1992.