An Appearance-Based Representation of Action

Aaron Bobick
James Davis

MIT Media Laboratory
20 Ames St., Cambridge, MA 02139 USA
E-mail: bobick|jdavis@media.mit.edu

Abstract

A new view-based approach to the representation of action is presented. 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.

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. [9]) 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. [16]). 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.

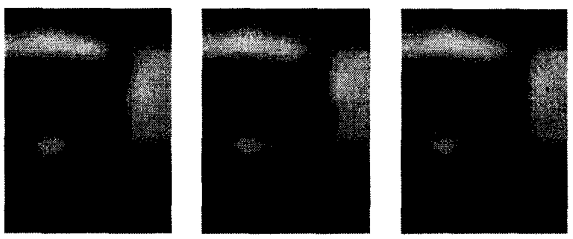

Figure 1: Selected frames from video of someone performing an action. Almost no structure is present in each frame, nor are there any features from which to compute a structural description (as would be in a moving light display). Yet people can trivially recognize the action as someone sitting.

1.1. An observation

The motivation for the work presented in this paper can be demonstrated in a single video-sequence of which a few frames are shown in Figure 1. The video is a tremendously blurred sequence — in this case an upsampling 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. Furthermore, it would difficult to believe that there are features which can be tracked robustly for a

[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."

1015-4651/96 $5.00 © 1996 IEEE
Proceedings of ICPR '96

structure-from-motion routine to first compute a three dimensional structure.

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

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. Using a 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. Finally, we describe some initial experiments using manually placed and tracked "sticks" as the underlying primitive which shows the effectiveness of the representation.

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, 11, 16, 17, 6, 20]. Some of these techniques assume that a three-dimensional reconstruction precedes the recognition of action, while others use only the two-dimensional appearance. 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.

The second area related to this work is that of appearance- or aspect-based recognition (e.g. [14, 13, 8]). The formal description of aspects [14] referred to the visible surfaces of objects undergoing self occlusion. For a range of viewing angles, which surfaces are visible surfaces remains constant and only the shape of their projection changes. Ikeuchi and Hong [13] refer to the shape change within an aspect as a "linear shape change." 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 [15, 18, 19, 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.

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

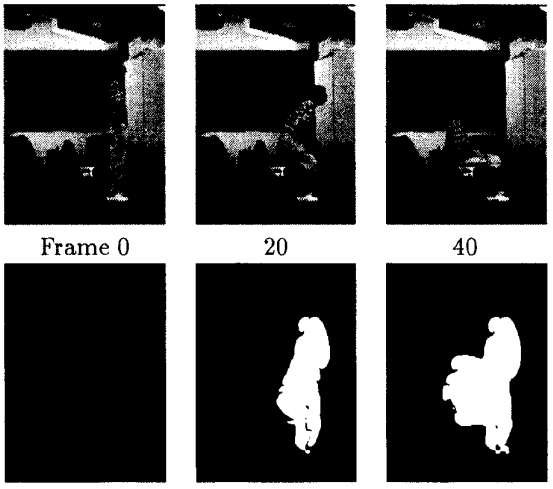

Figure 2: Example of someone sitting. Top row is keys frames; bottom row is cumulative motion images starting from Frame 0.

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 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. As expected the sequence sweeps out a particular region of the 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. 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:

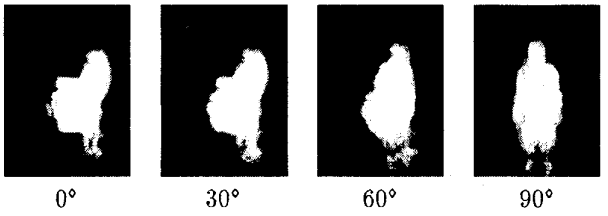

Figure 3: Average MEIs of sitting from 4 viewing angles.

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

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, because motion-differencing measures where pixels have changed (rather than how they moved), we used a binary thresholding of the sum of the squared difference between each frame and the first. Once we have constructed the MEI for each training sequence, we computed an average MEI to represent each viewing angle of the action. The averaging process requires normalizing scale and translation; Figure 3 illustrates the average MEI for the sitting action from several viewing angles.

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 $< m_1, \ldots, m_7 >$ are the Hu moments [12] 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. If there is some latitude in the verification procedure, then one

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

only needs to find the correct action at a near-by viewing direction.

4. Motion modeling

The last component of the representation is the motion description. Our work here seeks to extend the approach of Black and Yacoob [3] by using specific motion parameterizations for different viewing angles of different movements. 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. Finally, we propose a hypothesis and test approach where the MEI matches are used to select the motion patches to be described.

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 shows the manually placed sticks in three 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:

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

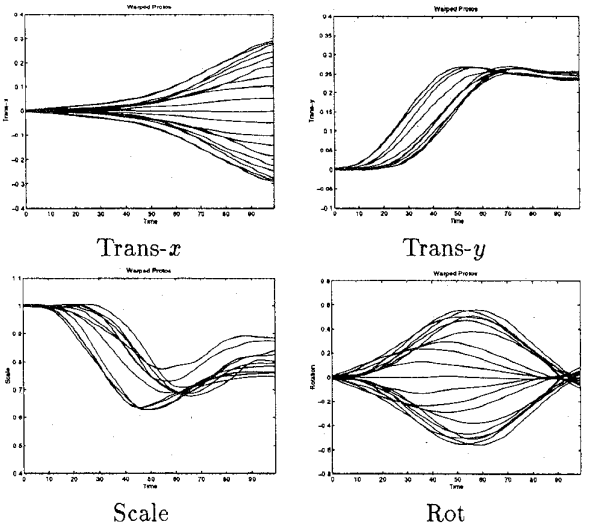

Figure 5: Four motion parameters for one sitting stick, for each of training viewing angles.

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 Figure 5 show the average[2] traces for α every 10° from 0° to 180° for each of the parameters of the torso stick.[3] 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.

4.2. Variation within training data

To determine whether a test motion is consistent with a known movement we need to characterize the vari-

[2] 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.

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

ability 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 $\vec{C}^m(\alpha)$.

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 aerobic exercise 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.[4]

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. Equivalent to the stick deformation would be a parameterization that models optic flow as a four parameter deformation; a 6- or 8-parameter planar model is possible as well.

[4] 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.

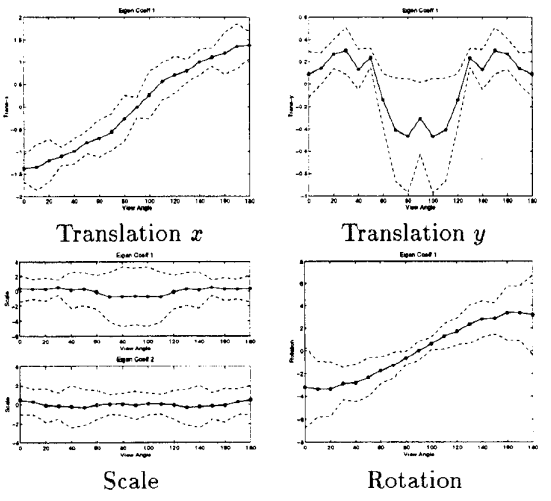

Figure 6: Mean and 3σ acceptance region for one stick of the motion model for sitting as a function of view angle α.

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 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 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 tracking 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 tracking model from which the coefficients are derived. The basic algorithm would be to (1) Use the input MEI to select a target MEI, a tracking model, and coefficient vector; (2) Compute a scale and translation transformation between the MEIs; (3) Use the transformation to align the selected tracking model with the input sequence; (4) Track the model and extract the motion parameters; and (5) Test for acceptance of the action given the view angle associated with the selected target MEI.

5. 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.

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 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. Hierarchical model-based motion estimation. In *European Conference on Computer Vision*, pages 237–252, 1992.

[3] Black, M. and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motion using local parametric models of image motion. In *ICCV*, 1995.

[4] Campbell, L. and A. Bobick. Recognition of human body motion using phase space constraints. In *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. In *ICCV*, 1995.

[7] Darrell, T., P. Maes, B. Blumberg, and A. Pentland. A novel environment for situated vision and behavior. In *IEEE Wkshp. for Visual Behaviors (CVPR-94)*, 1994.

[8] Eggert, D., K. Bowyer, C. Dyer, H. Christensen, and D. Goldgof. The scale space aspect graph. *IEEE Trans. PAMI*, 15(11), 1993.

[9] Freeman, W. Orientation histogram for hand gesture recognition. In *Int'l Workshop on Automatic Face- and Gesture-Recognition*, 1995.

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

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

[12] Hu, M. Visual pattern recognition by moment invariants. *IRE Trans. Information Theory*, IT-8(2), 1962.

[13] Ikeuchi, K. and K. S. Hong. Determining linear shape change: Toward automatic generation of object recognition programs. *CVGIP, Image Understanding*, 53(2), 1991.

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

[15] Polana, R. and R. Nelson. Low level recognition of human motion. In *IEEE Workshop on Non-rigid and Articulated Motion*, 1994.

[16] Rehg, J. and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, 1995.

[17] Rohr, K. Towards model-based recognition of human movements in image sequences. *CVGIP, Image Understanding*, 59(1), 1994.

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

[19] Yacoob, Y. and L. Davis. Computing spatio-temporal representations of human faces. In *CVPR*, 1994.

[20] Yamato, J., J. Ohya, and K. Ishii. Recognizing human action in time sequential images using hidden markov models. In *CVPR*, 1992.