M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 386
Appears in: The Workshop on Applications of Computer Vision, December 1996 (WACV'96)

Real-time Recognition of Activity Using Temporal Templates

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

Abstract

A new view-based approach to the representation and recognition of action is presented. The basis of the representation is a *motion-history image* (MHI) — a static image where intensity is a function of the recency of motion in a sequence. We develop a recognition method which uses both binary and scalar-valued versions of the MHI as temporal templates to match against stored instances of actions. The method automatically performs temporal segmentation, is invariant to linear changes in speed, and runs in real-time on a standard platform. The applications we have begin to develop include simple room monitoring and an interactive game.

Frame 5 25 40

Figure 1: Selected frames from video of someone performing an action. Even with almost no structure present in each frame people can trivially recognize the action as someone sitting.

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. [13]). This paper presents an alternative to the three-dimensional reconstruction proposal. We develop a view-based approach to the representation and recognition of action that is designed to support the direct recognition of the motion itself.

In previous work [3] we described how people can easily recognize action in even extremely blurred image sequences such as shown in Figure 1. Such capabilities argue for recognizing action from the motion itself, as opposed to first reconstructing a 3-dimensional model of a person, and then recognizing the action of the model as advocated in [1, 4, 10, 13, 14, 6, 17]. In [3] we proposed a representation and recognition theory that decomposed motion-based recognition into first describing *where* there is motion (the spatial pattern) and then describing *how* the motion is moving. The approach is a natural extension of Black and Yacoob's work on facial expression recognition [2].

In this work we continue to develop this approach. We review the construction of a *motion-energy* image (MEI) which is a binary representation of where motion has occurred in an image sequence. We next introduce *motion-history* images (MHI) which is a scalar-valued image where intensity is a function of recency of motion. These motion-history images will serve as view-specific *temporal templates* which are matched against the stored representations of known actions. Finally we present a recognition method which automatically performs temporal segmentation, is invariant to linear changes in speed, and runs in real-time on a standard platform.

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 two areas: human action recognition 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, 10, 13, 14, 6, 17]. 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.

Alternatively, there is the work on direct motion recognition [12, 15, 16, 2, 8]. These approaches attempt

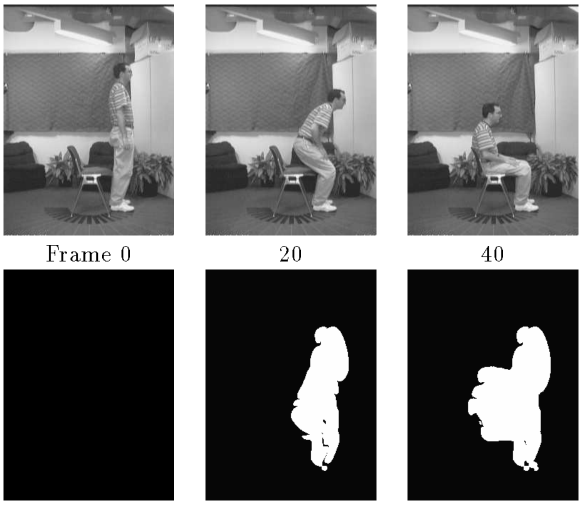

Figure 2: Example of someone sitting. Top row contains key frames; bottom row is cumulative motion images starting from Frame 0.

to characterize the motion itself without any reference to the underlying static images. Of these techniques, the work of Black and Yacoob [2] 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, nor does it use any underlying model of the geometry to interpret the results.

3 Temporal templates

In this section we define a dual component representation of action based upon the observed motion. In [3] we performed this division by using a binary image to represent where motion occurred and a patch model of how the motion moves. Here, we replace the dynamic patch tracking with a static representation of the motion. This new static representation forms the basis of the temporal templates.

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 binary 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).

We refer to these binary cumulative motion images as *motion-energy* images (MEI). Let $I(x,y,t)$ be an image sequence, and let $D(x,y,t)$ be a binary image sequence indicating regions of motion; for many applications image-differencing is adequate to generate D.

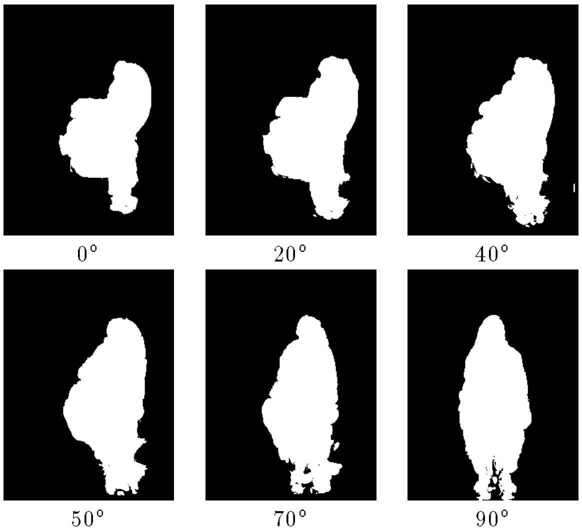

Figure 3: MEIs of sitting action over 90° viewing angle. The smooth change implies only a coarse sampling of viewing direction is necessary to recognize the action from all angles.

Then the MEI $E_\tau(x,y,t)$ is defined

$$E_\tau(x,y,t) = \bigcup_{i=0}^{\tau-1} D(x,y,t-i)$$

We note that the duration τ is critical in defining the temporal extent of an action. Fortunately, in the recognition section we derive a backward-looking (in time) algorithm which can dynamically search over a range of τ.

In Figure 3 we display the MEIs of viewing a sitting action across 90°. In [3] we exploited the smooth variation of motion over angle to compress the entire view circle into a low order representation. Here we simply note that because of the slow variation across angle, we only need to sample the view sphere coarsely to recognize all directions.

3.2 Motion-history images

To represent *how* motion is moving we enhance the MEI to form a *motion-history* image (MHI). In an MHI, pixel intensity is a function of the motion history at that point. For the results presented here we use a simple replacement and decay operator:

$$H_\tau(x,y,t) = \begin{cases} \tau & \text{if } D(x,y,t) = 1 \\ \max\ (0, H(x,y,t-1) - 1) & \\ \quad \text{otherwise} & \end{cases}$$

The result is a scalar-valued image where more recently moving pixels are brighter. Examples of MHIs are presented in Figure 4. Note that unlike MEIs, the MHIs are sensitive to direction of motion. Also note that the MHI can be generated by thresholding the MEI above zero.

4 Recognition of action

4.1 Matching temporal templates

To construct a recognition system, we need to define a matching algorithm for the the MEI and the MHI. Because we are using an appearance-based approach, we must first define the desired invariants for the matching technique. Because we are interested in actions whose orientations (in the image plane) are relatively fixed but which can occur anywhere in the image at arbitrary scale, we have selected a technique which is scale and translation invariant.

We first collect training examples of each action from a variety of viewing angles. Given a set of MEIs and MHIs for each view/action combination, we compute statistical descriptions of the these images using moment-based features. Our current choice are 7 Hu moments [11] which are known to yield reasonable shape discrimination in a translation- and scale-invariant manner. For each view of each action a statistical model (mean and covariance matrix) is generated for both the MEI and MHI. To recognize an input action, a Mahalanobis distance is calculated between the moment description of the input and each of the known actions.

Note that we have no fundamental reason for selecting this method of scale- and translation-invariant template matching. The approach outlined has the advantage of not being computationally taxing; one disadvantage is that the Hu moments are difficult to reason about intuitively. Also, we note that the matching methods for the MEI and MHI need not be the same; in fact, given the distinction we make between where there is motion from how the motion is moving one might expect different matching criteria.

4.2 Real-time segmentation and recognition

The final element of performing recognition is the temporal segmentation and matching. During the training phase we measure the minimum and maximum duration that an action may take, τ_{min} and τ_{max}. However, if the test actions are performed at varying speeds, we need to choose the right τ for the computation of the MEI and the MHI. Our current system uses a backward looking variable time window. Because of the simple nature of the replacement operator we can construct a highly efficient algorithm for approximating a search over a wide range of τ.

The algorithm is as follows: At each time step a new MHI $H_\tau(x, y, t)$ is computed setting $\tau = \tau_{max}$, where τ_{max} is the longest time window we want the system to consider. Because of the recursive nature of the MHI $H_\tau(x, y, t)$ is trivially computable from $H_\tau(x, y, t-1)$. We choose $\Delta\tau$ to be $(\tau_{max} - \tau_{min})/(n-1)$ where n is the number of temporal integration windows to be considered.[1] A simple thresholding and scaling operator allows the computation of $H_{(\tau - \Delta\tau)}$ from H_τ. Iterating we compute all n MHIs at each time step. Thresholding the MHI yields the corresponding MEI.

[1]Ideally $n = \tau_{max} - \tau_{min} + 1$ resulting in a complete search of the time window between τ_{max} and τ_{min}. Only computational limitations argue for a smaller n.

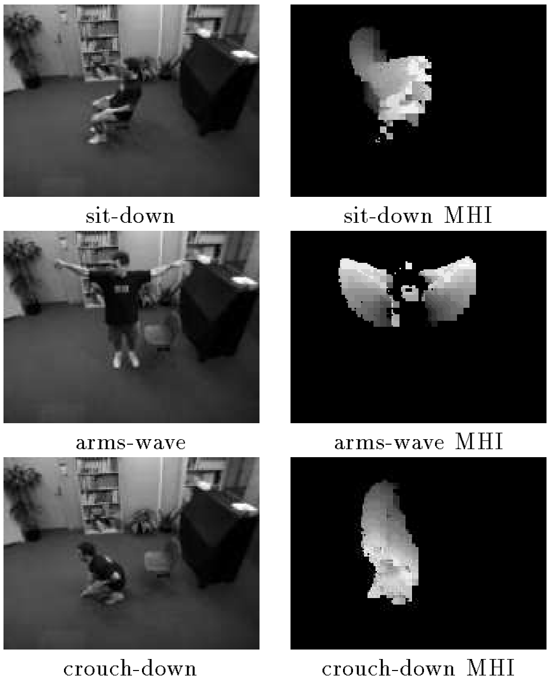

Figure 4: Action moves along with their MHIs used in a real-time system

After computing the various MEIs and MHIs, we compute the Hu moments for each image. We then check the Mahalanobis distance of the MEI parameters against the known view/action pairs. Any action found to be within a threshold distance of the input is tested for agreement of the MHI. If more than one action is matched, we select the action with the smallest distance.

Currently the system recognizes 180° views of the actions *sitting*, *arm waving*, and *crouching* (See Figure 4). Except for the head-on view of crouching and sitting discussed in the next section, the system performs well, rarely misclassifying the actions. However, because we are only using a small number of actions it seems premature to present statistics of recognition rates. The errors which do arise are mainly caused by problems with image differencing and also due to our approximation of the temporal search window $n < (\tau_{max} - \tau_{min} + 1)$.

The system runs at approximately 10 Hz using a color CCD camera connected to a Silicon Graphics Indy. The images are digitized to a size of 160x120, $\tau_{max=19}$ (approximately 2 seconds), $\tau_{min} = 11$ (approximately 1 second), and $n = 6$. The comparison operation is virtually no cost in terms of computational load, so adding more actions does not affect the speed of the algorithm, only the accuracy of the recognition.

5 Extensions, problems, and applications

We have presented a novel representation and recognition technique for identifying actions. The approach is based upon *temporal templates* and their dynamic matching in time.

There are, of course, some difficulties in the approach. Some of these are easily rectified. For example, we currently assume all motion present in the image should be incorporated into the temporal templates. This approach fails miserably when two people are in the field of view. Clearly a bounding window would be required. A worse condition is when one person partially occludes another, making separation difficult, if not impossible. Here multiple cameras is an obvious solution. Since occlusion is view angle specific, multiple cameras reduces the chance the occlusion is present in all views.

Multiple cameras also alleviates the difficulty when certain views are easily confused. For example, consider the actions of sitting and squatting when viewed from the front. The observed motions are almost identical, and the coarse temporal template solution proposed does not well distinguish them. However, if one also has a side view then the action are easily discerned.

A more serious difficulty arises when the motion of part of the body is not specified during an action. Consider, for example, throwing a ball. Whether the legs move is not determined by the action itself, inducing huge variability in the statistical description of the temporal template. To extend this paradigm to such actions requires some mechanism to automatically mask away regions of motion. We have not yet addressed this problem.

In this paper we have introduced a new core technology with many potential applications to monitoring, surveillance, and human-computer interaction. Systems such as ALIVE [7] require good understanding of action. Currently most such systems use sequences of static configurations to identify action and are therefore subject to the same brittleness that static modeling approaches suffer. By using the motion itself we hope to improve the robustness of such interaction. In the same vein we have started developing a "Simon-says" game where the machine watches the player, checking the actions for compliance.

References

[1] Akita, K. Image sequence analysis of real world human motion. *Pattern Recognition*, 17, 1984.

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

[3] Bobick, A. and J. Davis. An appearance-based representation of action. In *ICPR*, August 1996.

[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*, 1995.

[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] Essa, I. and S. Pentland. Facial expression recognition using a dynamic model and motion energy. In *ICCV*, 1995.

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

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

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

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

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

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

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

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

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