M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 394
Appears in the British Machine Vision Conference, Edinburgh, Scotland — September 1996

Computers Seeing Action

Aaron F. Bobick
MIT Media Laboratory
20 Ames Street
Cambridge, MA 02139 USA
bobick@media.mit.edu
http://www.media.mit.edu/~bobick

Abstract

As research in computer vision has shifted from only processing single, static images to the manipulation of video sequences, the concept of *action recognition* has become important. Fundamental to understanding action is reasoning about time, in either an implicit or explicit framework. In this paper I describe several specific examples of incorporating time into representations of action and how those representations are used to recognize actions. The approaches differ on whether variation over time is considered a continuous mapping, a state-based trajectory, or a qualitative, semantically labeled sequence. For two of the domains — whole body actions and hand gestures — I describe the approaches in detail while two others — constrained semantic domains (e.g. watching someone cooking) and labeling dynamic events (e.g. American football) — are briefly mentioned.

1 Seeing Action

Understanding video sequences is different than conventional image understanding in that one is interested in what is *happening* in a scene, as opposed to what *is in* the scene. One might believe that attempting to describe what is happening in hundreds of images is not a viable research of goal given the difficulty of understanding just one picture.

However, video understanding can be regarded as a way of providing *more* constraint in the interpretation of imagery. We require that the image interpretation be plausible over time: extracted structure must obey the temporal constraints of the domain. For example, if we are annotating an American football play, we might be interested in tracking the quarterback. Unfortunately, current (even near future) technology cannot see or track the quarterback in every frame. However, assuming he never disappears from the field of play, we can "track" him as he enters an amorphous blob and re-emerges six frames later. The program cannot see him during this time, but it knows he's there.

Understanding time can be either explicit, as in the above example, or implicit, captured in the representation of action. One example that we will expand upon later is our work in gesture recognition [3, 22]. In this work gesture is represented either deterministically by an explicit sequence of states through which the hand must move, or probabilistically by a hidden Markov model. In both cases the requirement that the interpretation be consistent with the temporal constraints of the domain is guaranteed by matching the input data to learned representations of action which are sensitive to time.

From our perspective, one of the future directions of computer vision lies in the area of action understanding. In this paper I will detail two different approaches to incorporating time into a representation of action and then causally performing recognition.[1] The first focuses on recognizing whole body motion by using *temporal templates*: a view-based, model-based description of image variation over time. The second technique, applied to hand gesture understanding, develops a state-based model of time captured in a probabilistic framework. Finally, in the conclusion I will refer to additional work where knowledge about time and actions is explicitly expressed in rules that are used by the system to interpret the imagery.

2 Recognizing motion: temporal templates

The lure of wireless interfaces (e.g. [11]) and interactive environments [9] 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. [6, 12, 19, 20]).

Consider, however, an extremely blurred sequence of action; a few frames on one such example is shown in Figure 1. Even with almost no structure present in each frame people can trivially recognize the action as someone sitting. 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. The prior work in this area has addressed either periodic or gross motion detection and recognition [17, 21, 24] or the understanding of facial expressions [23, 1, 10].

In [2, 5] we propose a representation and recognition theory that decomposes motion-based recognition into first describing *where* there is motion (the spatial pattern) and then describing *how* the motion is moving. The basic idea is that we project the temporal pattern of motion into a single, image-based representation — a *temporal template*. This approach is a natural extension of Black and Yacoob's work on facial expression recognition[1].

[1] That is, the temporal segmentation and recognition tasks are performed simultaneously.

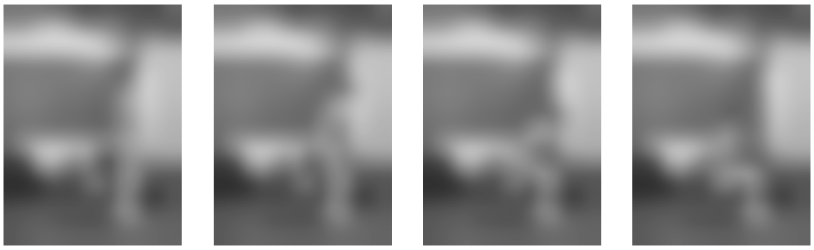

Figure 1: Selected frames from video of someone performing an action. Even with almost no structure present in each frame people can trivially recognize the action as someone sitting.

2.1 Motion images

Consider the example of someone sitting, as shown in Figure 2a. 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. 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 2b we display the MEIs of viewing a sitting action across 90°. In [2] 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.

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 3. 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.

2.2 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. As 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 [13] 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.

2.3 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 [5].

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 3). Except for the head-on view of crouching and sitting which appear quite similar in terms of motion descriptions, the system performs well, rarely misclassifying the actions.

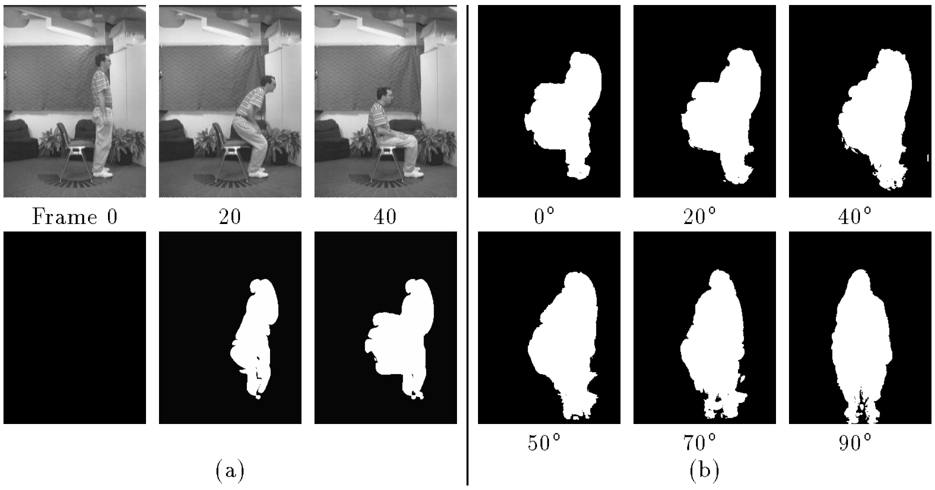

Figure 2: Example of someone sitting. (a) Top row contains key frames; bottom row is cumulative motion images starting from Frame 0. (b) MEIs for each of 6 viewing directions;the smooth change implies only a coarse sampling of viewing direction is necessary to recognize the action from all angles.

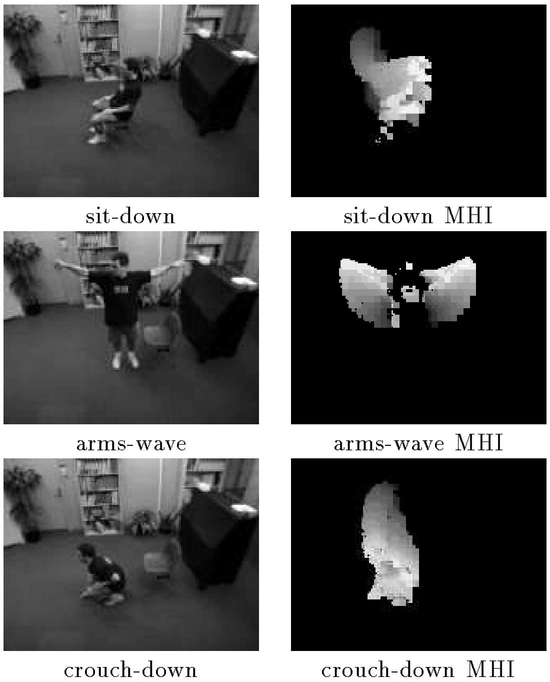

Figure 3: Action moves along with their MHIs used in a real-time system

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. Currently we are developing a multi-camera approach which should increase robustness by requiring both limited consistency across views and a good match from at least one view.

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). 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.

3 Discrete time and temporal states

Another domain which is amenable to view-based techniques is that of *gesture recognition*. First, we note that gestures are embedded within communication. As such, the gesturer typically orients the movements towards the recipient of the gesture. Visual gestures are therefore *viewpoint-dependent* [8, 7]. Second, in the space of motions allowed by the body's degrees of freedom, there is a small subspace that we use in the making of a gesture. Taken together, these observations argue for a view-based approach in which only a small subspace of human motions is represented.

How should a system model human motion to capture the constraints present in the gestures? There may be no single set of features that makes explicit the relationships that hold for a given gesture. In the case of hand gestures, for example, the spatial configuration of the hand may be important (as in a point gesture, when the observer must notice a particular pose of the hand), or alternatively, the gross motion of the hand may be important (as in a friendly wave across the quad). Quek [18] has observed that it is rare for both the pose and the position of the hand to simultaneously change in a meaningful way during a gesture.

Recently we have presented an approach that represents gesture as a sequence of states in a particular observation space [3]. We then extended that work and developed a technique for learning visual behaviors that 1) incorporates the notion of multiple models — multiple ways of describing a set of sensor data[15]; 2) makes explicit the idea that a given phase of a gesture is constrained to be within some small subspace of possible human motions; and 3) represents time as a probabilistic trajectory through states [22]. The basic idea is that the different models need to approximate the (small) subspace associated with a particular state and membership in a state is determined by how well the state models can represent the current observation. The parsing of the entire gesture is accomplished by finding a likely sequence of states given the memberships and the learned transition probabilities between the states.

The details of the techniques are presented in [3, 22]. The approach is based upon state models that define a *residual* — how well a given model can represent the current sensor input. We then embed this residual-based technique within a Hidden Markov Model framework; the HMMs represent the temporal aspect of the gestures in a probabilistic manner and provide an implicit form of dynamic time warping for the recognition of gesture.

Here we illustrate the technique by way of two examples. Figure 4 — a wave gesture — consists of a single model example but shows the use of the HMM. In this case, the parameters associated with the model of each state are simply a number of the top eigenimages that account for most of the variance of the training images (as indicated by the eigenvalues). The input consists of 32 image sequences of a waving hand, each about 25 frames (60 by 80 pixels, gray-scale) in length.

The recovered Markov model, the mean image at each state, and plots of the memberships and residual for one training sequence are shown in Figure 4. The recovered Markov model allows the symmetry of motion seen in the plot of membership over an observation sequence. Some other observation sequences differ in the extent of the wave motion; in these cases the state representing the hand at its lowest or highest position in the frame is not used.

Our second example describes the position and configuration of a waving, pointing hand (Figure 5). In each frame of the training sequences, a 50 by 50 pixel image of the hand was tracked and clipped from a larger image with a cluttered background. Foreground segmentation was accomplished using the known background. The configuration C of the hand is modeled by the eigenvector decomposition of the 50 by 50 images. The position P of the hand is modeled by the location of the tracked hand within the larger image. The recovered Markov model is similar to that of the waving hand in the previous example except now there are two components of the model of each state. As before, this gesture is recognized if a highly probable parse can be generated by the HMM.

The variance of each feature indicates the importance of the feature in describing the gesture. In this example both the position and configuration of the hand was relevant in describing the gesture. Had the location of the hand varied greatly in the training set, the high variance of the position representation would have indicated that position was not important in describing the gesture. The important point here is that each state defines the important models associated with that phase of the gesture.

4 Reasoning about seeing action

Finally, I mention some current work that makes time explicit. One research effort in our lab is *video annotation*, in particular labeling American football plays. In [14] we developed *closed-world tracking*, a technique that reasons about local contexts at a semantic level (e.g. "grass", "players", "field lines") to build robust templates to track players . We are currently developing context sensitive methods for recognizing the plays themselves. The basic idea is to represent an action as a labeled sequence of events. Borrowing from the object recognition literature, the iterative approach is to use some visual features to reduce the space of possible plays, which in turn constrain the events that need be detected, which further constrain the solution. For more details see:
`http://www-white.media.mit.edu/vismod/demos/football/football.html`

A different focus is taken in [16, 4] where we introduce *SmartCams* — cameraman-less cameras — that respond to

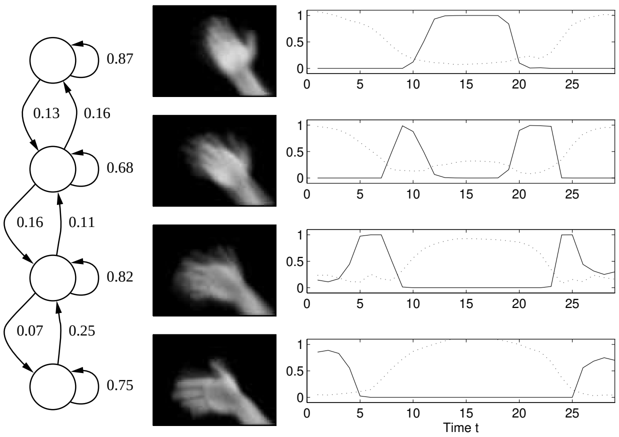

Figure 4: A wave gesture. The recovered Markov model for all training sequences at left shows the symmetry of the gesture. The mean image for each state is shown in the middle. On the right is a plot of membership (solid line) and residual (dotted line) for each state for one training sequence. The exact shape of the plots varies in response to the variance and length of the sequence.

(a)

(b)

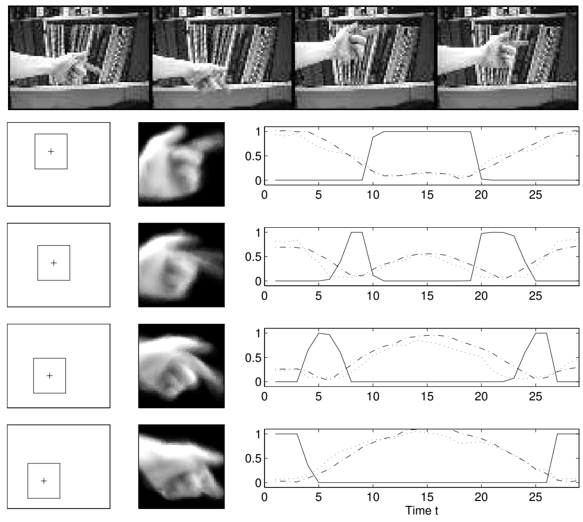

Figure 5: (a) Four representative frames (ordered left to right) are shown from one training sequence. (b) The mean location of the tracked hand in the larger image is shown on the left. The mean image for each state is shown in the middle. On the right is a plot of membership (solid line), configuration residual (dotted line), and the position residual (dash-dotted line) for each state for one training sequence.

a director's requests while filming a cooking show.. Such cameras perform inverse video-annotation: given some symbolic description (*"close-up chef"*) the system needs to generate the correct image. One key element of the system is that it maintains an approximate world model to control the selection of view-based vision routines, and that selection process is controlled by rules that explicitly model time and action. Actions are represented in a frame-based system as a sequence of activities, and each activity is known to have visual correlates. The system uses simple inferencing mechanisms to derive the current visual primitives to exploit given knowledge about the current action; likewise, detection of visual primitives validates that an expected action is occurring. For more details and a demonstration see:
`http://www-white.media.mit.edu/vismod/demos/smartcams/smartcams.html`

5 The future

Computer vision has just begun to consider action. Yet, the vast majority of images recorded are frames of video sequences. For computer vision to begin to understand the images in our environment, or to understand the environment itself, it is clear that understanding action and behavior is fundamental. Computer vision has developed numerous ways of representing a cup (Euclidean solids, superquadrics, spline surfaces, particles); how many ways do we have to represent throwing a baseball? Or even getting a wicket?

Acknowledgment: Thanks to the entire HLV lab and in particular Jim Davis, Andy Wilson, Stephen Intille and Claudio Pinhanez who did most of the labor reported here. The work presented here is supported in part by a research grant from LG Electronics and by ORD contract 94-F133400-000.

References

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

[2] A. Bobick and J. Davis. Real-time recognition of activity using temporal templates. In *Proceedings of WACV 96*, 1996.

[3] A. F. Bobick and A. D. Wilson. A state-based technique for the summarization and recognition of gesture. *Proc. Int. Conf. Comp. Vis.*, 1995.

[4] Aaron Bobick and Claudio Pinhanez. Using approximate models as source of contextual information for vision processing. In *Proc. of the ICCV'95 Workshop on Context-Based Vision*, pages 13–21, Cambridge, Massachusetts, July 1995.

[5] Bobick, A. and J. Davis. Real-time recognition of activity using temporal templates. In *Submitted to WACV*, December 1996.

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

[7] Y. Cui and J. Weng. Learning-based hand sign recognition. In *Proc. of the Intl. Workshop on Automatic Face- and Gesture-Recognition*, Zurich, 1995.

[8] T.J. Darrell and A.P. Pentland. Space-time gestures. *Proc. Comp. Vis. and Pattern Rec.*, pages 335–340, 1993.

[9] 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.

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

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

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

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

[14] S.S. Intille and A.F. Bobick. Closed-world tracking. In *Proc. Int. Conf. Comp. Vis.*, June 1995.

[15] R. W. Picard and T. P. Minka. Vision texture for annotation. *Journal of Multimedia Systems*, 3:3–14, 1995.

[16] Claudio S. Pinhanez and Aaron F. Bobick. Approximate world models: Incorporating qualitative and linguistic information into vision systems. To appear in AAAI'96, 1996.

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

[18] F. Quek. Hand gesture interface for human-machine interaction. In *Proc. of Virtual Reality Systems*, volume Fall, 1993.

[19] Rehg, J. and T. Kanade. Model-based tracking of self-occluding articulated objects. In *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. In *IJCAI Workshop: Looking at People*, 1995.

[22] A. D. Wilson and A. F. Bobick. Learning visual behavior for gesture analysis. In *Proc. IEEE Int'l. Symp. on Comp. Vis.*, Coral Gables, Florida, November 1995.

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

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