ACTION RECOGNITION USING TEMPORAL TEMPLATES

AARON F. BOBICK AND JAMES W. DAVIS
MIT Media Laboratory
20 Ames St., Cambridge, MA 02139
E-mail: bobick,jdavis@media.mit.edu

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. [13]) and interactive environments [11, 3] 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. [22]). This chapter 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 [4, 6] we described how people can easily recognize action in even extremely blurred image sequences such as shown in Figure 1 and in `lowres_action.mov`[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, 7, 16, 22, 23, 9, 28]. In [4] 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 chapter we continue to develop this approach. We review the construction of a binary *motion-energy* image (MEI) which represents where motion has occurred in an image sequence. We next generate a *motion-*

[1] `http://vismod.www.media.mit.edu/vismod/demos/actions/lowres_action.mov`

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.

history image (MHI) which is a scalar-valued image where intensity is a function of recency of motion. Taken together, the MEI and MHI can be considered as a two component version of a *temporal template*, a vector-valued image where each component of each pixel is some function of the motion at that pixel location. These view-specific templates are matched against the stored models of views of known actions. To evaluate the power of the representation we evaluate the discrimination power on a set of 18 aerobics exercises. 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 approaches to recognizing motion and action has recently grown at a tremendous rate. For an excellent survey on the machine understanding of motion (particularly human motion) see the work of Cédras and Shah [8]. Their review article details methods for extracting motion information (e.g. optic flow) and for performing matching. That survey article also discusses recent work in motion recognition (e.g. lip-reading and gesture recognition). In this chapter we divide the immediately relevant prior work into two general areas: configuration-based tracking and recognition and motion-based recognition.

2.1. CONFIGURATION-BASED TRACKING AND RECOGNITION OF ACTION

2.1.1. *Tracking*

The first and most obvious body of relevant work includes the approaches using structural or appearance-based representations to tracking and understanding human action. Some believe that a 3-D description is necessary and sufficient for understanding action (e.g. [16, 7, 23, 14, 22, 15]), while others choose to analyze the 2-D appearance as a means of interpretation (e.g. [9, 10, 1, 28]). We now take a closer look at these approaches.

The most common method for attaining the 3-D information in the action is to recover the pose of the object at each time instant using a 3-D model of the object. A common method for model fitting in these works is to use a residual measure between the projected model and object contours (e.g. edges of body in the image). This generally requires a strong segmentation of foreground/background and also of the individual body parts to aid the model alignment process. It is difficult to imagine such techniques could be extended to the blurred sequence of Figure 1.

For example, Rehg and Kanade [22] used a 27 degree-of-freedom (DOF) model of a human hand in their system called "Digiteyes". Local image-based trackers are employed to align the projected model lines to the finger edges against a solid background. The work of Goncalves et al. [15] promoted 3-D tracking of the human arm against a uniform background using a two cone arm model and a single camera. Though it may be possible to extend their approach to the whole body as claimed, it seems unlikely that it is appropriate for non-constrained human motion with self-occlusion. Hogg [16] and Rohr [23] used a full-body cylindrical model for tracking walking humans in natural scenes. Rohr incorporates a 1 DOF pose parameter to aid in the model fitting. All the poses in a walking action are indexed by a single number. Here there is only a small subset of poses which can exist. Gavrila and Davis [14] also used a full-body model (22 DOF, tapered superquadrics) for tracking human motion against a complex background. For simplifying the edge detection in cases of self-occlusion, the user is required to wear a tight-fitting body suit with contrasting limb colors.

One advantage of having the recovered model is the ability to estimate and predict the feature locations, for instance edges, in the following frames. Given the past history of the model configurations, prediction is commonly attained using Kalman filtering [23, 22, 15] and velocity constraints [21, 14].

Because of the self-occlusions that frequently occur in articulated objects, some employ multiple cameras and restrict the motion to small regions [22, 14] to help with projective model occlusion constraints. A single camera is used in [16, 15, 23], but the actions tracked in these works had little deviation in the depth of motion. Acquiring the 3-D information from

image sequences is currently a complicated process, many times necessitating human intervention or contrived imaging environments.

2.1.2. *Recognition*

As for action recognition, Campbell and Bobick [7] used a commercially available system to obtain 3-D data of human body limb positions. Their system removes redundancies that exist for particular actions and performs recognition using only the information that varies between actions. This method examines the relevant parts of the body, as opposed to the entire body data. Siskind [25] similarly used known object configurations. The input to his system consisted of line-drawings of a person, table, and ball. The positions, orientations, shapes, and sizes of the objects are known at all times. The approach uses support, contact, and attachment primitives and event logic to determine the actions of dropping, throwing, picking up, and putting down. These two approaches address the problem of recognizing actions when the precise configuration of the person and environment is known while the methods from the previous section concentrate on the recovery of the object pose.

In contrast to the 3-D reconstruction and recognition approaches, others attempt to use only the 2-D appearance of the action (e.g. [1, 10, 9, 28]). View-based representations of 2-D statics are used in a multitude of frameworks, where an action is described by a sequence of 2-D instances/poses of the object. Many methods require a normalized image of the object (usually with no background) for representation. For example, Cui et al. [9], Darrell and Pentland [10], and also Wilson and Bobick [26] present results using actions (mostly hand gestures), where the actual grayscale images (with no background) are used in the representation for the action. Though hand appearances remain fairly similar over a wide range of people, with the obvious exception of skin color, actions that include the appearance of the total body are not as visually consistent across different people due to obvious natural variations and different clothing. As opposed to using the actual raw grayscale image, Yamato et al. [28] examines body silhouettes, and Akita [1] employs body contours/edges. Yamato utilizes low-level silhouettes of human actions in a Hidden Markov Model (HMM) framework, where binary silhouettes of background-subtracted images are vector quantized and used as input to the HMMs. In Akita's work [1], the use of edges and some simple 2-D body configuration knowledge (e.g. the arm is a protrusion out from the torso) are used to determine the body parts in a hierarchical manner (first find legs, then head, arms, trunk) based on stability. Individual parts are found by chaining local contour information. These two approaches help alleviate *some* of the variability between people but introduce other problems such as the disappearance of movement that

happens to be within the silhouetted region and also the varying amount of contour/edge information that arises when the background or clothing is high versus low frequency (as in most natural scenes). Also, the problem of examining the entire body, as opposed to only the desired regions, still exists, as it does in much of the 3-D work.

Whether using 2-D or 3-D structural information, many of the approaches discussed so far consider an action to be comprised of a sequence of poses of an object. Underlying all of these techniques is the requirement that there be individual features or properties that can be extracted and tracked from each frame of the image sequence. Hence, motion understanding is really accomplished by recognizing a sequence of static configurations. This understanding generally requires previous recognition and segmentation of the person [21]. We now consider recognition of action within a motion-based framework.

2.2. MOTION-BASED RECOGNITION

Direct motion recognition [21, 24, 20, 2, 27, 25, 12, 4, 6] approaches attempt to characterize the motion itself without reference to the underlying static poses of the body. Two main approaches include the analysis of the body region as a single "blob-like" entity and the tracking of predefined *regions* (e.g. legs, head, mouth) using motion instead of structural features.

Of the "blob-analysis" approaches, the work of Polana and Nelson [21], Shavit and Jepson [24], and also Little and Boyd [20] are most applicable. Polana and Nelson use repetitive motion as a strong cue to recognize cyclic walking motions. They track and recognize people walking in outdoor scenes by gathering a feature vector, over the entire body, of low-level motion characteristics (optical-flow magnitudes) and periodicity measurements. After gathering training samples, recognition is performed using a nearest centroid algorithm. By assuming a fixed height and velocity of each person, they show how their approach is extendible to tracking multiple people in simple cases. Shavit and Jepson also take an approach using the gross overall motion of the person. The body, an animated silhouette figure, is coarsely modeled as an ellipsoid. Optical flow measurements are used to help create a phase portrait for the system, which is then analyzed for the force, rotation, and strain dynamics. Similarly, Little and Boyd recognize people walking by analyzing the motion associated with two ellipsoids fit to the body. One ellipsoid is fit using the motion region silhouette of the person, and the other ellipsoid is fit using motion magnitudes as weighting factors. The relative phase of various measures (e.g. centroid movement, weighted centroid movement, torque) over time for each of the ellipses characterizes the gait of several people.

There is a group of work which focuses on motions associated with facial expressions (e.g. characteristic motion of the mouth, eyes, and eyebrows) using region-based motion properties [27, 2, 12]. The goal of this research is to recognize human facial expressions as a dynamic system, where the motion of interest regions (locations known *a priori*) is relevant. Their approaches characterize the expressions using the underlying motion properties rather than represent the action as a sequence of poses or configurations. For Black and Yacoob [2], and also Yacoob and Davis [27], optical flow measurements are used to help track predefined polygonal patches placed on interest regions (e.g. mouth). 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. Then these qualitative labels were used in a rule-based, temporal model for recognition to determine expressions such as anger or happiness. Recently, Ju, Black, and Yacoob [19] have extended this work with faces to include tracking the legs of a person walking. As opposed to the simple, independent patches used for faces, an articulated three-patch model was needed for tracking the legs. Many problems, such as large motions, occlusions, and shadows, make motion estimation in that situation more challenging than for the facial case. We extended this expression recognition approach by applying a similar framework to the domain of full-body motion [5].

Optical flow, rather than patches, was used by Essa [12] to estimate muscle activation on a detailed, physically-based model of the face. One recognition approach classifies expressions by a similarity measure to the typical patterns of muscle activation. Another recognition method matches motion energy templates derived from the muscle activations. These templates compress the activity sequence into a single entity. In this chapter, we develop similar templates, but our templates incorporate the temporal motion characteristics.

3. Temporal templates

Our goal is to construct a view-specific representation of action, where action is defined as motion over time. For now we assume that either the background is static, or that the motion of the object can be separated from either camera-induced or distractor motion. At the conclusion of this chapter we discuss methods for eliminating incidental motion from the processing.

In this section we define a multi-component image representation of action based upon the observed motion. The basic idea is to construct a vector-image which can be matched against stored representations of known

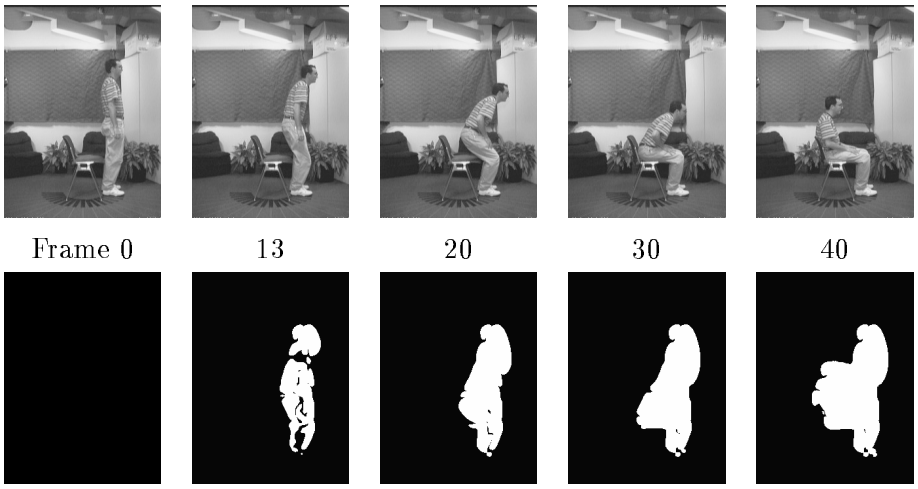

Figure 2. Example of someone sitting. Top row contains key frames; bottom row is cumulative motion images starting from Frame 0.

actions; this image is used as a temporal template.

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 (*where* there is motion) 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 binary 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.

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.

In Figure 3 we display the MEIs of viewing a sitting action across 90°. In [4] 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. In the evaluation section of this chapter we use samplings of every 30° to recognize a large variety of motions.

3.2. MOTION-HISTORY IMAGES

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

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

The result is a scalar-valued image where more recently moving pixels are brighter. Examples of MHIs are presented in Figure 4 and the dynamic construction of an MHI is illustrated in `mhi_generation.mov`[2]. Note that the MEI can be generated by thresholding the MHI above zero.

One possible objection to the approach described here is that there is no consideration of optic flow, the direction of image motion. In response, it is important to note the relation between the construction of the MHI and direction of motion. Consider the waving example in Figure 4 where the arms fan upwards. Because the arms are isolated components — they do not occlude other moving components — the motion-history image implicitly represents the direction of movement: the motion in the arm down

[2] `http://vismod.www.media.mit.edu/vismod/demos/actions/mhi_generation.mov`

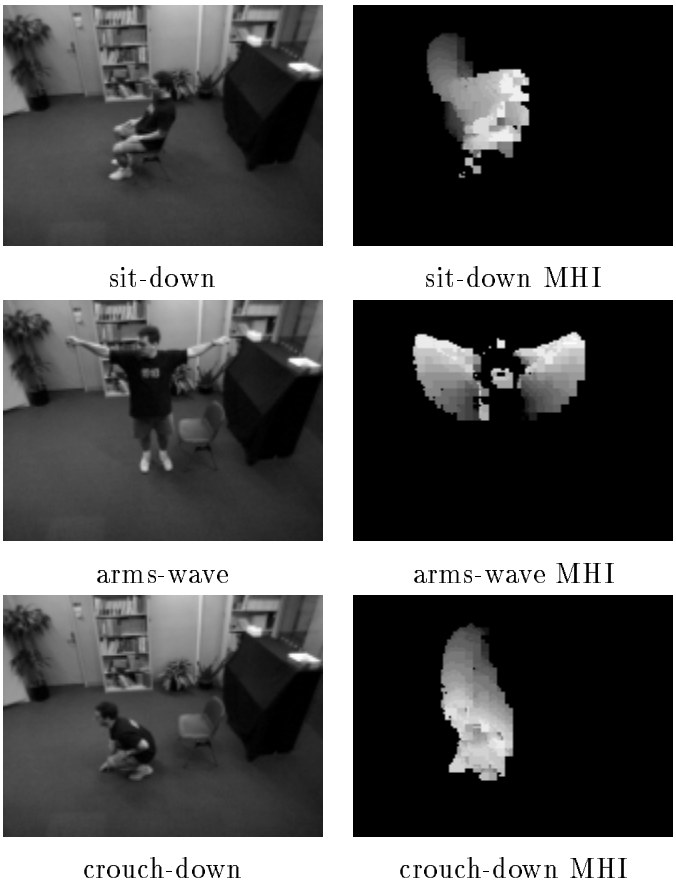

Figure 4. Action moves along with their MHIs used in a real-time system.

position is "older" than the motion when the arms are up. For these types of articulated objects, and for simple movements where there is not significant motion self-occlusion, the direction of motion is well represented using the MHI. As motions become more complicated the optic flow is more difficult to discern, but is typically not lost completely.

3.3. EXTENDING TEMPORAL TEMPLATES

The MEI and MHI are two components of a vector image designed to encode a variety of motion properties in a spatially indexed manner. Other possible components of the temporal templates include power in directional motion integrated over time (e.g. "in this pixel there has been a large amount of motion in the down direction during the integrating time window") or

the spatially localized periodicity of motion (a pixel by pixel version of Polana and Nelson [21]). The vector-image template is similar in spirit to the vector-image based on orientation and edges used by Jones and Malik [18] for robust stereo matching.

For the results in this chapter we use only the two components derived above (MEI and MHI) for representation and recognition. We are currently considering other components to improve our performance.

4. Action Discrimination

4.1. MATCHING TEMPORAL TEMPLATES

To construct a recognition system, we need to define a matching algorithm for the temporal template. Because we are using an appearance-based approach, we must first define the desired invariants for the matching technique. As we are using a view sensitive approach, it is desirable to have a matching technique that is as invariant as possible to the imaging situation. Therefore we have selected a technique which is rotation (in the image plane), 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 [17] which are known to yield reasonable shape discrimination in a translation- and scale-invariant manner (See appendix). For each view of each action a statistical model of the moments (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. In this section we analyze this distance metric in terms of its separation of different 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 making real-time implementation feasible; 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. TESTING ON AEROBICS DATA: ONE CAMERA

To evaluate the power of the temporal template representation, we recorded video sequences of 18 aerobics exercises performed several times by an expe-

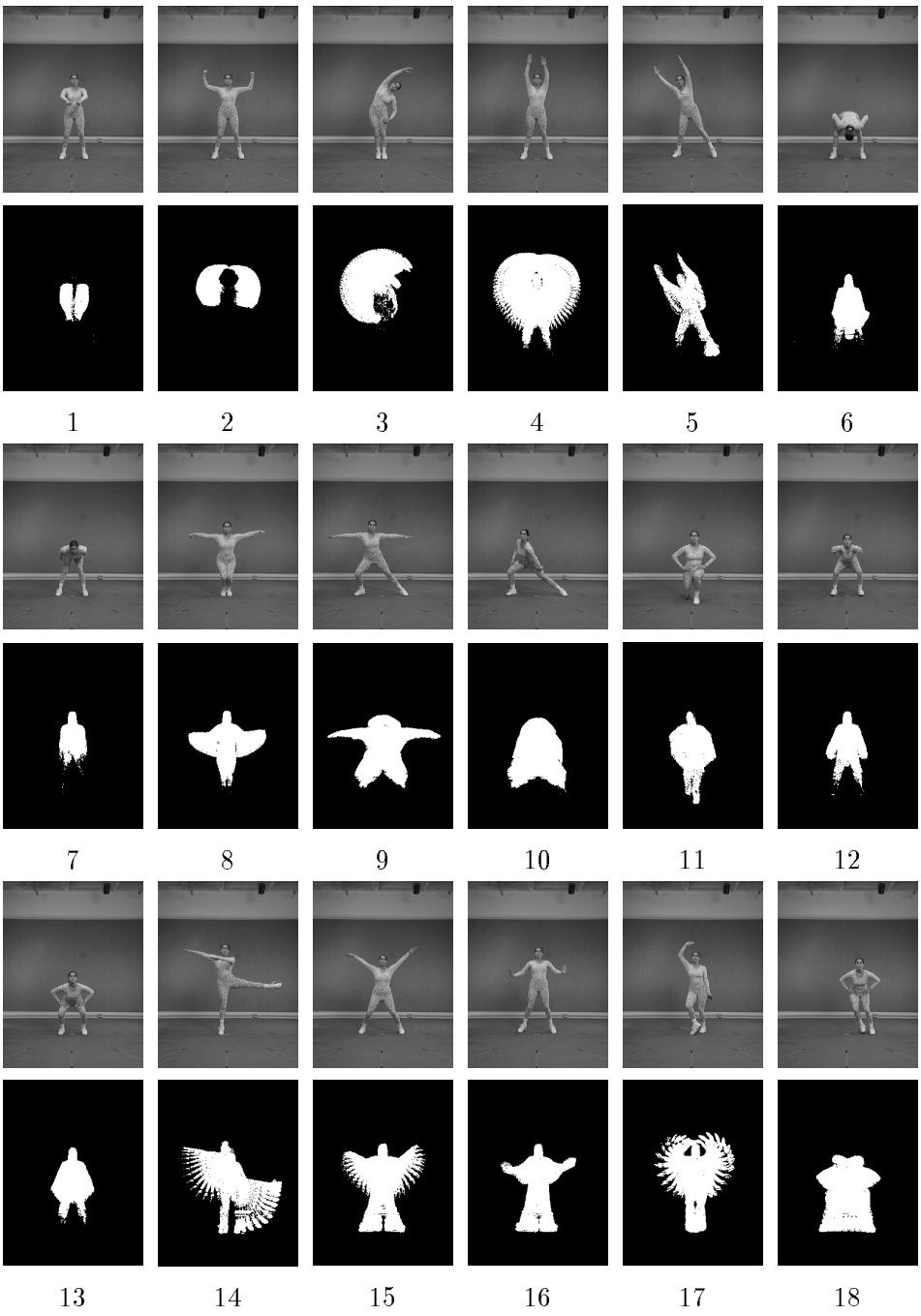

Figure 5. A single key frame and MEI from the frontal view of each of 18 aerobics exercises used to test the representation.

		Closest Dist	Closest Move	Correct Dist	Median Dist	Rank
Test	1	1.43	4	1.44	2.55	2
	2	3.14	2	3.14	12.00	1
	3	3.08	3	3.08	8.39	1
	4	0.47	4	0.47	2.11	1
	5	6.84	5	6.84	19.24	1
	6	0.32	10	0.61	0.64	7
Test	7	0.97	7	0.97	2.03	1
	8	20.47	8	20.47	35.89	1
	9	1.05	8	1.77	2.37	4
	10	0.14	10	0.14	0.72	1
	11	0.24	11	0.24	1.01	1
	12	0.79	12	0.79	4.42	1
Test	13	0.13	6	0.25	0.51	3
	14	4.01	14	4.01	7.98	1
	15	0.34	15	0.34	1.84	1
	16	1.03	15	1.04	1.59	2
	17	0.65	17	0.65	2.18	1
	18	0.48	10	0.51	0.94	4

TABLE 1. Test results using one camera at 30° off frontal. Each row corresponds to one test move and gives the distance to the nearest move (and its index), the distance to the correct matching move, the median distance, and the ranking of the correct move.

rienced aerobics instructor. The sequence `aerobic_action.mov`[3] provides an example. Seven views of the action — $+90°$ to $-90°$ in 30° increments in the horizontal plane — were recorded. Figure 5 shows the frontal view of one key frame for each of the moves along with the frontal MEI. We take the fact that the MEI makes clear to a human observer the nature of the motion as anecdotal evidence of the strength of this component of the representation. For this experiment the temporal segmentation and selection of the time window over which to integrate were performed manually. Later we will detail a self-segmenting, time-scaling recognition system.

We constructed the temporal template for each view of each move, and then computed the Hu moments on each component. To do a useful Ma-

[3] `http://vismod.www.media.mit.edu/vismod/demos/actions/aerobic_action.mov`

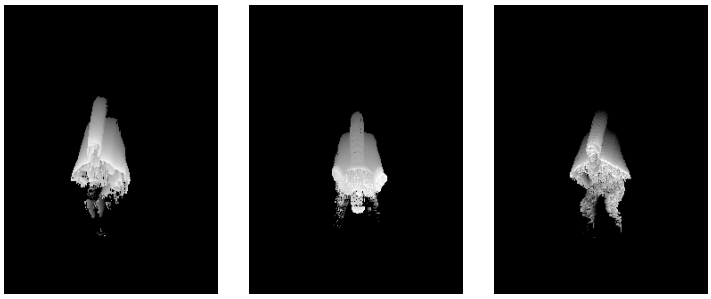

Figure 6. An example of MHIs with similar statistics. (a) Test input of move 13 at 30°. (b) Closest match which is move 6 at 0°. (c) Correct match.

halanobis procedure would require watching several different people performing the same actions; this multi-subject approach is taken in the next section where we develop a recognition procedure.

Instead, we design the experiment to be a measurement of confusion. A new test subject performed each move and the input data was recorded by two cameras viewing the action at approximately 30° to left and 60° to the right of the subject. The temporal template for each of the two views of the test input actions was constructed, and the associated moments computed.

Our first test uses only the left (30°) camera as input and matches against all 7 views of all 18 moves (126 total). We select as a metric a *pooled* independent Mahalanobis distance using a diagonal covariance matrix to accommodate variations in magnitude of the moments. Table 1 displays the results. Indicated are the distance to the move closest to the input (as well as its index), the distance to the correct matching move, the median distance (to give a sense of scale), and the ranking of the correct move in terms of least distance.

The first result to note is that 12 of 18 moves are correctly identified using the single view. This performance is quite good considering the compactness of the representation (a total of 14 moments from two correlated motion images) and the large size of the target set. Second, the typical situation in which the best match is not the correct move, the difference in distances from the input to the closest move versus the correct move is small compared to the median distance. Examples of this include test moves 1, 9, 13, 16, 18. In fact for moves 1, 16, 18 the difference is negligible.

To analyze the confusion difficulties further consider the example shown in Figure 6. Displayed here, left to right, are the input MHI (move 13 at view angle 30°), the closest match MHI (move 6 at view angle 0°), and the "correct" matching MHI. The problem is that an alternative view of a different action projects into a temporal template with similar statistics.

For example, consider sitting and crouching actions when viewed from the front. The observed motions are almost identical, and the coarse temporal template statistics do not distinguish them well.

4.3. COMBINING MULTIPLE VIEWS

A simple mechanism to increase the power of the method is to use more than one camera. Several approaches are possible. For this experiment, we use two cameras and find the minimum sum of Mahalanobis distances between the two input templates and two stored views of an action that have the correct angular difference between them, in this case 90°. The assumption embodied in this approach is that we know the approximate angular relationship between the cameras.

Table 2 provides the same statistics as the first table, but now using two cameras. Notice that the classification now contains only 3 errors. The improvement of the result reflects the fact that for most pairs of this suite of actions, there is some view in which they look distinct. Because we have 90° between the two input views the system can usually correctly identify most actions.

We mention that if the approximate calibration between cameras is not known (and is not to be estimated) one can still logically combine the information by requiring consistency in labeling. That is, we remove the inter-angle constraint, but do require that both views select the same action. The algorithm would be to select the move whose Mahalanobis sum is least, regardless of the angle between the target views. If available, angular order information — e.g. camera 1 is to the left of camera 2 — can be included. When this approach is applied to the aerobics data shown here we still get similar discrimination. This is not surprising because the input views are so distinct.

To analyze the remaining errors, consider Figure 7 which shows the input for move 16. Left to right are the 30° MHIs for the input, the best match (move 15), and the correct match. The test subject performed the move much less precisely than the original aerobics instructor. Because we were not using a Mahalanobis variance across subjects, the current experiment could not accommodate such variation. In addition, the test subject moved her body slowly while wearing low frequency clothing resulting in an MHI that has large gaps in the body region. We attribute this type of failure to our simple (i.e. naive) motion analysis; a more robust motion detection mechanism would reduce the number of such situations.

		Closest Dist	Closest Move	Correct Dist	Median Dist	Rank
Test	1	2.13	1	2.13	6.51	1
	2	12.92	2	12.92	19.58	1
	3	7.17	3	7.17	18.92	1
	4	1.07	4	1.07	7.91	1
	5	16.42	5	16.42	32.73	1
	6	0.88	6	0.88	3.25	1
Test	7	3.02	7	3.02	7.81	1
	8	36.76	8	36.76	49.89	1
	9	5.10	8	6.74	8.93	3
	10	0.68	10	0.68	3.19	1
	11	1.20	11	1.20	3.68	1
	12	2.77	12	2.77	15.12	1
Test	13	0.57	13	0.57	2.17	1
	14	6.07	14	6.07	16.86	1
	15	2.28	15	2.28	8.69	1
	16	1.86	15	2.35	6.72	2
	17	2.67	8	3.24	7.10	3
	18	1.18	18	1.18	4.39	1

TABLE 2. Results using two cameras where the angular interval is known and any matching views must have the same angular distance.

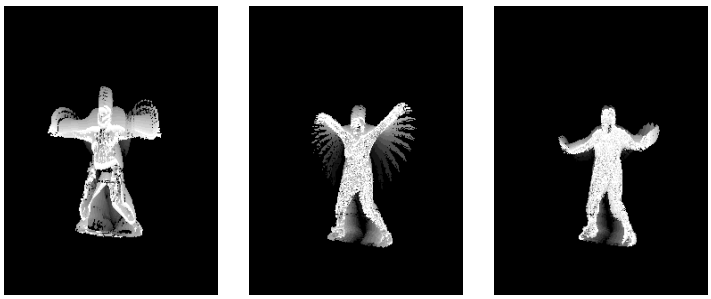

Figure 7. Example of error where failure is cause by both the inadequacy of using image differencing to estimate image motion and the lack of the variance data in the recognition procedure.

5. 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}. 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. We choose $\Delta\tau$ to be $(\tau_{max} - \tau_{min})/(n-1)$ where n is the number of temporal integration windows to be considered.[4] A simple thresholding of MHI values less than $(\tau - \Delta\tau)$ followed by a scaling operation generates $H_{(\tau-\Delta\tau)}$ from H_τ. Iterating we compute all n MHIs at each time step. Binarization of the MHIs yields the corresponding MEIs.

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; the mean and the covariance matrix for each view/action pair is derived from multiple subjects performing the same move. 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.

Our first experimental system recognizes 180° views of the actions *sitting*, *arm waving*, and *crouching* (See Figure 4). The training required four people and sampling the view circle every 45°. The system performs well, rarely misclassifying the actions. 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 sequence `two_cam_demo.mov`[5] provides a demonstration of this two camera technique. An avatar shown in the bottom window changes based upon the recognition state of the system.

The system runs at approximately 9 Hz using 2 CCD cameras connected to a Silicon Graphics 200MHz Indy; the images are digitized at a size of 160x120. For these three moves $\tau_{max=19}$ (approximately 2 seconds), $\tau_{min} = 11$ (approximately 1 second), and we chose $n = 6$. The comparison operation is virtually no cost in terms of computational load, so adding

[4] 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.

[5] `http://vismod.www.media.mit.edu/vismod/demos/actions/two_cam_demo.mov`

more actions does not affect the speed of the algorithm, only the accuracy of the recognition.

6. 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. Initial experiments in both measuring the sensitivity of the representation and in constructing real-time recognition systems have shown the effectiveness of the method.

There are, of course, some difficulties in the current approach. Several of these are easily rectified. As mentioned, a more sophisticated motion detection algorithm would increase robustness. Also, as developed, the method assumes all motion present in the image should be incorporated into the temporal templates. Clearly, this approach would fail when two people are in the field of view. To implement our real-time system we use a tracking bounding box which attempts to isolate the relevant motions.

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 reduce the chance the occlusion is present in all views. For monitoring situations, we have experimented with the use of an overhead camera to select which ground based cameras have a clear view of a subject and where the subject would appear in each image.

6.1. INCIDENTAL MOTION

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 templates. To extend this paradigm to such actions requires some mechanism to automatically mask away regions of this type of motion. Our current thinking is to process only the motion signal associated with the dominant motions.

Two other examples of motion that must be removed are camera motion and locomotion (if we assume the person is performing some action while locomoting and what we want to see is the underlying action). In both instances the problem can be overcome by using a body centered motion field. The basic idea would be to subtract out any image motion induced by camera movement or locomotion. Of these two phenomena, camera motion elimination is significantly easier because of the over constrained nature of estimating egomotion. Our only insight at this point is that because the temporal template technique does not require accurate flow fields it may

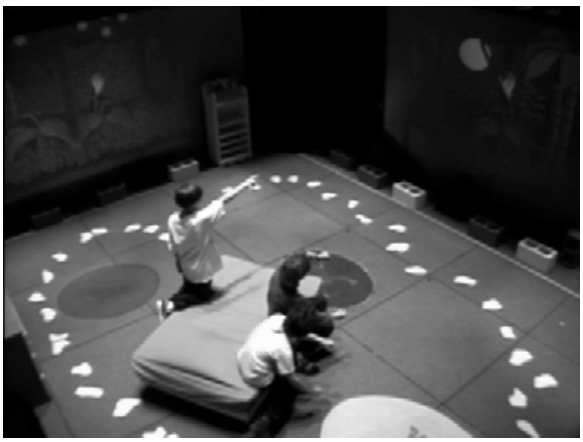

Figure 8. The KIDSROOM interactive play-space. Using a modified version of temporal templates the room responds to the actions of the children. All sensing is performed using vision from 3 cameras.

be necessary only to approximately compensate for these effects and then to threshold the image motion more severely than we have done to date.

6.2. THE KIDSROOM: AN APPLICATION

We conclude by mentioning a recent application we developed in which we employed a version of the temporal template technique described. On October 30th, 1996 we debuted The KidsRoom, an interactive play-space for children [3]. The basic idea is that the room is aware of the children (maximum of 4) and takes them through a story where the responses of the room are affected by what the children do. Computers control the lighting, sound effects, performance of the score, and illustrations projected on the two walls of the room that are actually video screens. The current scenario is an adventurous trip to Monsterland; a snapshot is shown in Figure 8.

In the last scene the monsters appear and teach the children to dance — basically to perform certain actions. Using a modified version of the MEIs[6] the room can compliment the children on well performed moves (e.g. spinning) and then turn control of the situation over to them: the monsters follow the children if the children perform the moves they were taught. The interactive narration coerces the children to room locations where occlusion is not a problem. Of all the vision processes required, the

[6]The MEIs were computed from background subtracted images instead of binary motion images. This change was necessary because of the high variability of incidental body motion. By using the background subtracted images the body was always included.

modified temporal template is one of the more robust. We take the ease of use of the method to be an indication of its potential.

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 *Proc. Int. Conf. Comp. Vis.*, 1995.
3. A. Bobick, J. Davis, S. Intille, F. Baird, L. Campbell, Y. Ivanov, C. Pinhanez, A. Schutte, and A. Wilson. Kidsroom: Action recognition in an interactive story environment. PerCom TR 398, MIT Media Lab, 1996.
4. Bobick, A. and J. Davis. An appearance-based representation of action. In *Proc. Int. Conf. Pat. Rec.*, August 1996.
5. Bobick, A. and J. Davis. An appearance-based representation of action. In *Proc. Int. Conf. Pat. Rec.*, August 1996.
6. Bobick, A. and J. Davis. Real time recognition of activity using temporal templates. In *IEEE Workshop on Applications of Computer Vision*, Sarasota, December 1996.
7. Campbell, L. and A. Bobick. Recognition of human body motion using phase space constraints. In *Proc. Int. Conf. Comp. Vis.*, 1995.
8. Cédras, C., and Shah, M. Motion-based recognition: A survey. *Image and Vision Computing*, 13(2):129–155, March 1995.
9. Cui, Y., D. Swets, and J. Weng. Learning-based hand sign recognition using shoslif-m. In *Proc. Int. Conf. Comp. Vis.*, 1995.
10. Darrell, T. and A. Pentland. Space-time gestures. In *Proc. Comp. Vis. and Pattern Rec.*, 1993.
11. 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.
12. Essa, I. and A. Pentland. Facial expression recognition using a dynamic model and motion energy. In *Proc. Int. Conf. Comp. Vis.*, June 1995.
13. Freeman, W., and M. Roth. Orientation histogram for hand gesture recognition. In *Int'l Workshop on Automatic Face- and Gesture-Recognition*, 1995.
14. Gavrila, D. and L. Davis. Tracking of humans in actions: a 3-d model-based approach. In *ARPA Image Understanding Workshop*, Feb 1996.
15. Goncalves, L., E. DiBernardo, E. Ursella, P. Perona. Monocular tracking of the human arm in 3d. In *Proc. Int. Conf. Comp. Vis.*, June 1995.
16. Hogg, D. Model-based vision: a paradigm to see a walking person. *Image and Vision Computing*, 1(1), 1983.
17. Hu, M. Visual pattern recognition by moment invariants. *IRE Trans. Information Theory*, IT-8(2), 1962.
18. D. Jones and J. Malik. Computational framework for determining stereo correspondence from a set of linear spatial filters. *Image and Vision Computing*, 10(10):699–708, 1992.
19. Ju, S., Black, M., and Y. Yacoob. Cardboard people: a parameterized model of articulated image motion. In *Submitted to the Second International Conference on Automatic Face and Gesture Recognition*, 1996.
20. Little, J., and J. Boyd. Describing motion for recognition. In *International Symposium on Computer Vision*, pages 235–240, November 1995.
21. Polana, R. and R. Nelson. Low level recognition of human motion. In *IEEE Workshop on Non-rigid and Articulated Motion*, 1994.
22. Rehg, J. and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. Int. Conf. Comp. Vis.*, 1995.

23. Rohr, K. Towards model-based recognition of human movements in image sequences. *CVGIP, Image Understanding*, 59(1), 1994.
24. Shavit, E. and A. Jepson. Motion understanding using phase portraits. In *IJCAI Workshop: Looking at People*, 1995.
25. Siskind, J. M. Grounding language in perception. In *SPIE*, September 1993.
26. Wilson, A. and A. Bobick. Learning visual behavior for gesture analysis. In *Proc. IEEE Int'l. Symp. on Comp. Vis.*, Coral Gables, Florida, November 1995.
27. Yacoob, Y. and L. Davis. Computing spatio-temporal representations of human faces. In *Proc. Comp. Vis. and Pattern Rec.*, 1994.
28. Yamato, J., J. Ohya, and K. Ishii. Recognizing human action in time sequential images using hidden markov models. In *Proc. Comp. Vis. and Pattern Rec.*, 1992.

A. General moments

The two-dimensional $(p+q)$th order moments of a density distribution function $\rho(x,y)$ (e.g. image intensity) are defined in terms of Riemann integrals as

$$m_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q \rho(x,y)dxdy\,, \tag{1}$$

for $p,q = 0,1,2,\cdots$.

B. Moments invariant to translation

The central moments μ_{pq} are defined as

$$\mu_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\bar{x})^p (y-\bar{y})^q \rho(x,y)d(x-\bar{x})d(y-\bar{y})\,, \tag{2}$$

where

$$\begin{aligned} \bar{x} &= m_{10}/m_{00}\,, \\ \bar{y} &= m_{01}/m_{00}\,. \end{aligned}$$

It is well known that under the translation of coordinates, the central moments do not change, and are therefore invariants under translation. It is quite easy to express the central moments μ_{pq} in terms of the ordinary moments m_{pq}. For the first four orders, we have

$$\begin{aligned} \mu_{00} &= m_{00} \equiv \mu \\ \mu_{10} &= 0 \\ \mu_{01} &= 0 \\ \mu_{20} &= m_{20} - \mu\bar{x}^2 \\ \mu_{11} &= m_{11} - \mu\bar{x}\bar{y} \end{aligned}$$

$$
\begin{aligned}
\mu_{02} &= m_{02} - \mu\bar{y}^2 \\
\mu_{30} &= m_{30} - 3m_{20}\bar{x} + 2\mu\bar{x}^3 \\
\mu_{21} &= m_{21} - m_{20}\bar{y} - 2m_{11}\bar{x} + 2\mu\bar{x}^2\bar{y} \\
\mu_{12} &= m_{12} - m_{02}\bar{x} - 2m_{11}\bar{y} + 2\mu\bar{x}\bar{y}^2 \\
\mu_{03} &= m_{03} - 3m_{02}\bar{y} + 2\mu\bar{y}^3
\end{aligned}
$$

C. Moments invariant to translation, scale, and orientation

For the second and third order moments, we have the following seven translation, scale, and orientation moment invariants:

$$
\begin{aligned}
\nu_1 &= \mu_{20} + \mu_{02} \\
\nu_2 &= (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2 \\
\nu_3 &= (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \\
\nu_4 &= (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2 \\
\nu_5 &= (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\
&\quad +(3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03}) \\
&\quad \cdot[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \\
\nu_6 &= (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \\
&\quad +4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \\
\nu_7 &= (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\
&\quad -(\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]
\end{aligned}
$$

These moments can used for pattern identification not only independently of position, size, and orientation but also independently of parallel projection.

D. Moments invariant to translation and scale

Under the scale transformation for moment invariants we have

$$\mu'_{pq} = \alpha^{p+q+2}\mu_{pq}\,. \tag{3}$$

By eliminating α between the zeroth order relation,

$$\mu' = \alpha^2\mu \tag{4}$$

and the remaining ones, we have the following absolute translation and scale moment invariants:

$$\frac{\mu'_{pq}}{(\mu')^{(p+q)/2+1}} = \frac{\mu_{pq}}{\mu^{(p+q)/2+1}}\,, \tag{5}$$

for $p+q = 2, 3, \cdots$ and $\mu'_{10} = \mu'_{01} \equiv 0$.