M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 402
Appears in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97)

The Representation and Recognition of Action Using Temporal Templates

James W. Davis and Aaron F. Bobick
MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139
(jdavis | bobick@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 temporal template *— a static vector-image where the vector value at each point is a function of the motion properties at the corresponding spatial location in an image sequence. Using 18 aerobics exercises as a test domain, we explore the representational power of a simple, two component version of the templates: the first value is a binary value indicating the presence of motion, and the second value is a function of the recency of motion in a sequence. We then develop a recognition method which matches these temporal templates against stored instances of views of known actions. The method automatically performs temporal segmentation, is invariant to linear changes in speed, and runs in real-time on a standard platform. We recently incorporated this technique into the* KidsRoom*: an interactive, narrative play-space for children.*

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. [11]) and interactive environments [9, 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. [17]). 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 [4, 5] 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, 6, 12, 17, 18, 8, 21]. 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].

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.

In this work 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-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 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 [7]. 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, 6, 12, 17, 18, 8, 21]. 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 [16, 19, 20, 2, 10, 15, 4]. These approaches attempt to characterize the motion itself without any reference to the underlying static images or a sequence of poses. Of these techniques, the work of Polana and Nelson [16] is the most relevant to the results presented here. The goal of their research is to represent and recognize actions as dynamic systems where it is the spatially distributed properties motion (in their case periodicity) that is matched.

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 paper 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 actions; it 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.

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

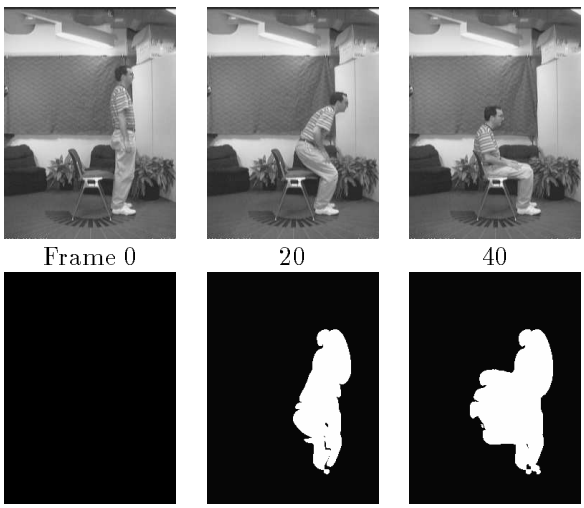

Figure 2: Example of someone sitting. Top row contains key frames; bottom row is cumulative motion images starting from Frame 0.

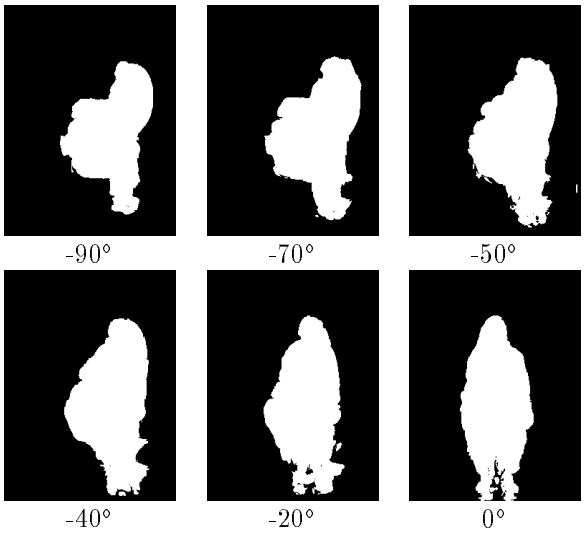

Figure 3: MEIs of sitting action over 90° viewing angle. The smooth change implies only a coarse sampling of viewing direction is necessary to recognize the action from all angles.

to recognize all directions. In the evaluation section of this paper 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

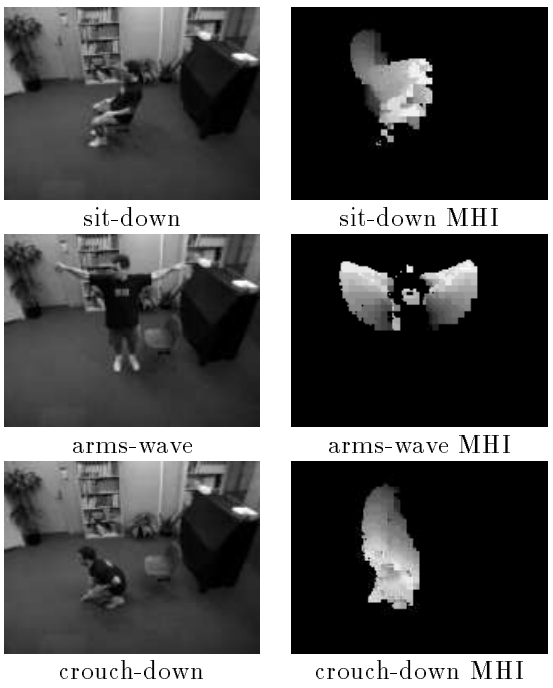

Figure 4: Action moves along with their MHIs used in a real-time system.

and decay operator:

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

The result is a scalar-valued image where more recently moving pixels are brighter. Examples of MHIs are presented in Figure 4. Note that 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 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 [16]). The vector-image template is similar in spirit to the vector-image based on orientation and edges used by Jones and Malik [14] for robust stereo matching.

For the results in this paper 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 [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 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 experienced aerobics instructor. 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

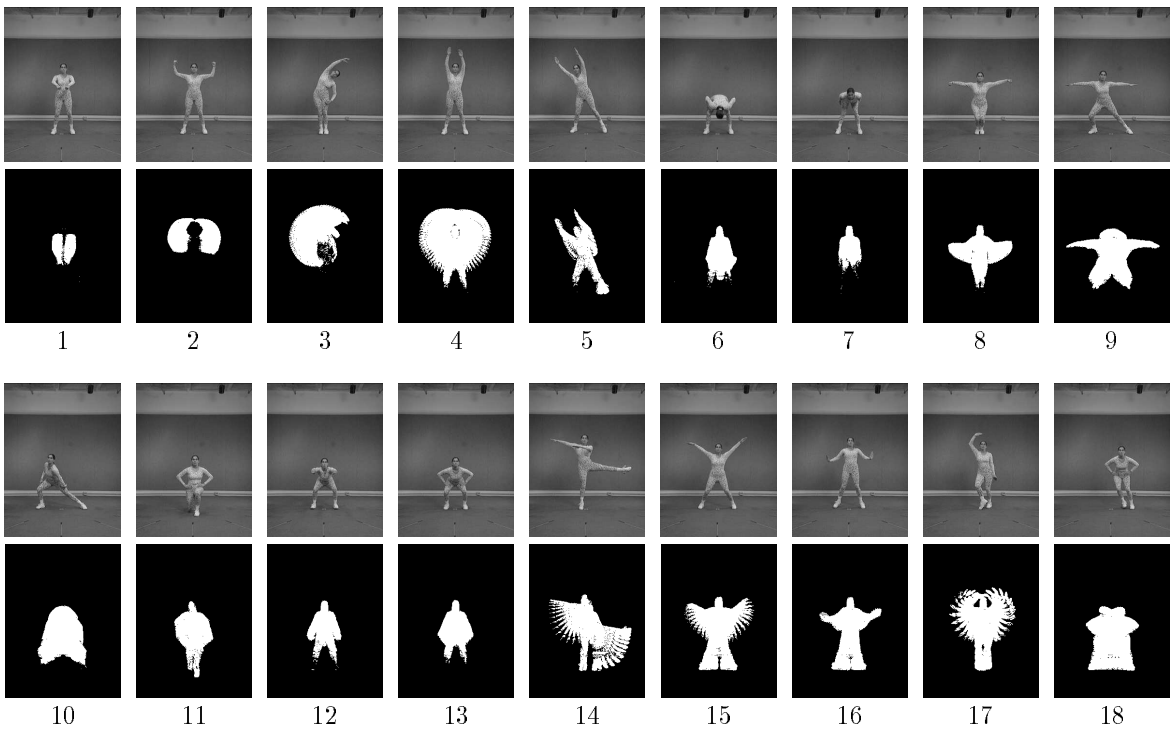

Figure 5: A single key frame and MEI from the frontal view of each of 18 aerobics exercises used to test the representation.

		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.

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 Mahalanobis 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 change 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

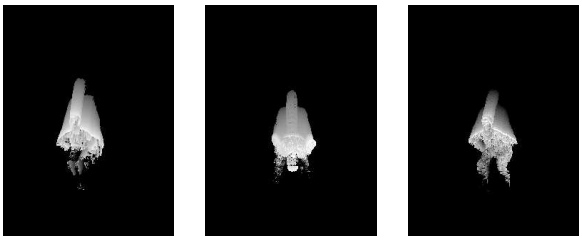

Figure 6: An example of MHIs with similar statistics. (a) Test input of move 13 at 30°. (b) Closest match which is move 6 at 0°. (c) Correct match.

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 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

		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.

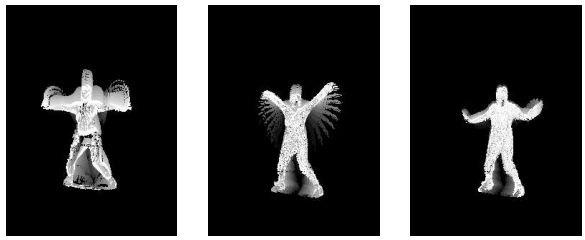

Figure 7: Example of error where failure is cause by both the inadequacy of using image differencing to estimate image motion and the lack of the variance data in the recognition procedure.

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.

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.[1] 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 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 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 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 scenes projected on the two walls of the room that are actually video screens. The current scenario is an adventurous trip to Monsterland (similar in spirit to Maurice Sendak's "Where the Wild Things Are"); 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[2] 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

[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.

[2] 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.

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.

vision processes required, the 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. Real time recognition of activity using temporal templates. In *IEEE Workshop on Applications of Computer Vision*, Sarasota, December 1996.

[6] Campbell, L. and A. Bobick. Recognition of human body motion using phase space constraints. In *Proc. Int. Conf. Comp. Vis.*, 1995.

[7] Cédras, C., and Shah, M. Motion-based recognition: A survey. *Image and Vision Computing*, 13(2):129–155, March 1995.

[8] Cui, Y., D. Swets, and J. Weng. Learning-based hand sign recognition using shoslif-m. In *Proc. Int. Conf. Comp. Vis.*, 1995.

[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 A. Pentland. Facial expression recognition using a dynamic model and motion energy. In *Proc. Int. Conf. Comp. Vis.*, June 1995.

[11] Freeman, W., and M. Roth. 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] 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.

[15] Little, J., and J. Boyd. Describing motion for recognition. In *International Symposium on Computer Vision*, pages 235–240, November 1995.

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

[17] Rehg, J. and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. Int. Conf. Comp. Vis.*, 1995.

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

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

[20] Yacoob, Y. and L. Davis. Computing spatio-temporal representations of human faces. In *Proc. Comp. Vis. and Pattern Rec.*, 1994.

[21] 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.