

RECOGNIZING HUMAN ACTIVITIES

Osama Masoud and Nikos Papanikolopoulos

{masoud,npapas}@cs.umn.edu

Department of Computer Science and Engineering
200 Union St. SE, 4-192 EE/CS Bldg., Minneapolis, MN 55455

ABSTRACT

This paper deals with the problem of classification of human activities from video as one way of performing activity monitoring. Our approach uses motion features that are computed very efficiently and subsequently projected into a lower dimension space where matching is performed. Each action is represented as a manifold in this lower dimension space and matching is done by comparing these manifolds. To demonstrate the effectiveness of this approach, it was used on a large data set of similar actions, each performed by many different actors. Classification results are accurate and show that this approach can handle many challenges such as variations in performers' physical attributes, color of clothing, and style of motion. An important result of this paper is that the recovery of three-dimensional properties of a moving person or even two-dimensional tracking of the person's limbs are not necessary steps that must precede action recognition.

1. INTRODUCTION

Recognition of human activity from video streams has many important surveillance applications. One such application is the monitoring suspicious activities. This application is directly related to homeland security and public safety and security at airports, transit, and public places. The approach of proceeding with a computer vision system is attractive due to the availability of high quality inexpensive cameras that makes it feasible to cover a large area. Such a system would be expected to identify suspicious activities like "putting a suitcase down and walking away." Traditionally, operators have to evaluate a large number of video-feeds and as a result some incidents may go by unnoticed. Simple motion detectors suffer from the problem of giving too many false positives. A human, a dog, or a swaying tree will all trigger the alarm. A surveillance system needs to be able to distinguish between a human and other moving objects. Furthermore, it should be able to distinguish a suspicious activity from a regular one.

In the past, our group has done some related work on tracking humans [8] and crowds [9]. In this paper, we con-

centrate on the "close-up" problem of activity classification. Specifically, we present a general method for human activity classification from video. Our method uses motion information directly from the video sequence. The other alternative is to perform tracking in 2-D or in 3-D and then use the tracking information to do action classification. Performing tracking of an articulated body like the human body is a very complex problem due to issues of self-occlusion and the effects of clothing on appearance. Perfect limb tracking is still not a solved problem. Our work is motivated by the need to investigate if it is possible to perform the task without having to perform limb tracking. It is also motivated by psychophysical evidence. In [3], it was demonstrated that our visual capabilities allow us to perceive actions with ease even when presented with an extremely blurred image sequence of an action. These experiments suggest that using motion alone to recognize actions may be favorable to reconstruction-based approaches.

The rest of the paper is organized as follows. Section 2 reviews some of the related work. Section 3 describes the motion features that we use. In Section 4, the learning and recognition algorithms are presented. The data that was used in our experiments is described in Section 5 followed by the experimental results in Section 6. Finally, the conclusion follows in Section 7.

2. RELATED WORK

Work in human activity recognition can be classified into three categories. A good review of work in these categories can be found in [2,4]. The first category are those methods that use 2-D body tracking information (e.g., [11]). The second category methods use 3-D body tracking information. Upon successful 3-D tracking, motion recognition can make use of any or the recovered parameters such as joint coordinates and joint angles. Although there has been a tremendous amount of work in 3-D limb tracking, work done in action recognition that uses 3-D tracking information has been limited to inputs of the form of Moving Light Displays (MLDs) obtained by placing markers on various body joints which are tracked in 3-D [1,5]. The third cate-

gory, to which our work belongs, uses motion features directly without attempting to track body parts. Several methods belong to this category. Yamato *et al.* [12] used Hidden Markov Models (HMMs) to distinguish different tennis strokes. The main advantage of such an approach is that adding a new action can be simply done by training a new HMM. The approach, however, was sensitive to the shape of the person performing the stroke. Use of motion features rather than spatial features may have reduced this sensitivity. Davis and Bobick [3] used what they called *motion-history* images (MHIs). An MHI represents motion recency where locations of more recent motions are brighter than older motions. A single MHI is used to represent an action. A pattern classification technique using seven Hu moments of the image was then used for recognition. They presented results of recognizing aerobic exercises performed by two actors, one for training and one for testing. The choice of an appropriate duration parameter used in the MHI calculation is critical. Temporal segmentation was done by trying all possible parameters. The system was able to successfully classify three different actions: sitting, arm waving, and crouching. Motion information extracted directly from the image sequence was also utilized by Polana and Nelson [10]. In their work, they used normal flow (the component of the flow field which is parallel to the gradient). The feature vector in their case was computed by temporally dividing the action into six divisions and finding the normal flow in each. Furthermore, each division is spatially partitioned into 4 by 4 cells. The summation of the magnitude of the normal flow at each cell was used to make up the feature vector. Recognition was done by finding the most similar vector in the training set using nearest centroid algorithm. They tested their method using six different activities, each performed several times by the same person and one activity performed by a toy frog.

3. FEATURE IMAGE

An Infinite Impulse Response (IIR) filter is used to construct the feature image. In particular, we use the response of the filter as a measure of motion in the image. A slightly different formulation of this measurement has been used by Halevi and Weinshall [6]. The idea is to represent motion by its recency: recent motion will be brighter than older motion. This technique, also called *recursive filtering*, is simple, time-efficient and therefore, suitable for real-time applications. We start with a weighted average at time i , M_i , which is computed as

$$M_i = \alpha \times I_{i-1} + (1 - \alpha) \times M_{i-1}, \quad (1)$$

where I_i , is the image at time i , and α is a scalar in the range 0 to 1. The feature image at time i , F_i , is computed

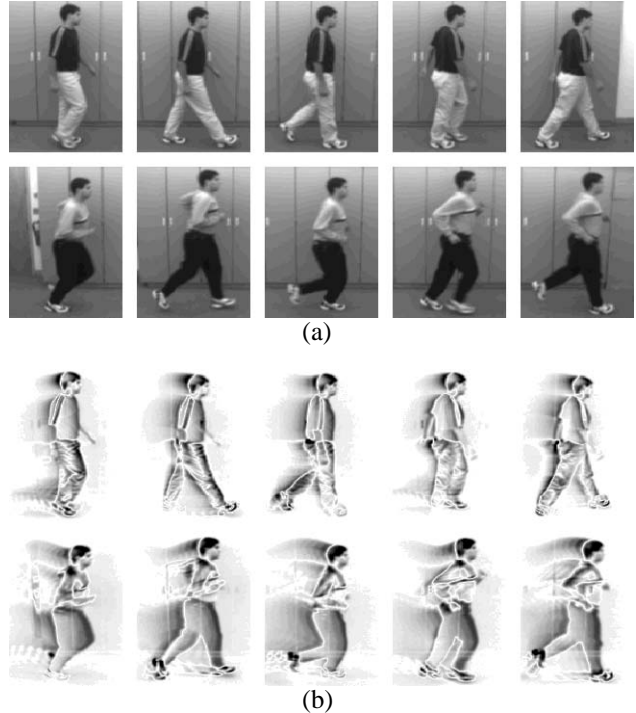


Figure 1 An example of a walking and a running motion sequence: (a) Original images. (b) Filtered images (feature images) with $\alpha = 0.3$.

as follows: $F_i = |M_i - I_i|$. Moving objects result in a fading trail behind them. The speed and direction of motion are implicit in this representation. The spread of the trail indicates the speed while the gradient of the region indicates direction. Figure 1 shows several frames from a motion sequence along with the extracted motion features using this technique. Note that it is the contrast of the gray level of the moving object which controls the magnitude of F , not the actual gray level value. The feature image values are normalized to be in the range $[0, 1]$. They are also thresholded to remove noise and insignificant changes (a threshold of 0.05 was found appropriate). Finally, a low-pass filter is applied to remove additional noise.

4. LEARNING AND RECOGNITION

Our goal is to classify actions into one of several categories. The idea is to compare the feature images with reference feature images of different learned actions and find the best match. There are several issues to consider using this approach. Action duration is not necessarily fixed for the same action. Also, the method should be able to handle small speedups or slowdowns. Even if we assume that actions are performed at the same speed, we cannot

assume temporal alignment and therefore a frame-by-frame matching starting from the first frame should be avoided. The frame-to-frame matching process itself needs to be invariant to the actor's physical attributes such as height, size, color of clothing, etc. Moreover, since an action can be composed of a large number of frames, correlation-based methods for matching may not be appropriate due to their computationally intensive nature.

4.1. Magnitude and Size Normalization

As actions are represented as sequences of feature images, two types of normalization are performed on a feature image:

1. Magnitude normalization: Because of the way feature images are computed, a person wearing clothes similar to the background will produce low magnitude features. To adjust for this, we normalize the feature image by the 2-norm of the vector formed by concatenating all the values in all the feature images corresponding to the action. The values are then multiplied by the square root of the number of frames to provide invariance to action length (in number of frames).
2. Size normalization: The images are resized so that they are all of equal dimensions. Not only does this type of normalization work across different people but also it corrects for changes in scale due to distance from the camera, for instance.

4.2. Principle Component Analysis

Principle component analysis (PCA) has been extensively used in the field of face recognition. The use of PCA in action recognition has been limited, however. Of a particular relevance to this work is the work of Yacoob and Black [11]. In their method, the features used were based on tracking five body parts using the work of Ju *et al.* [7]. Each tracked part provided eight temporal measurements. Thus, in total, 40 temporal curves are used to represent an action. Training data is composed of these curves for every example action. Each training sample is composed by concatenating all 40 curves. The training data is then compressed using a PCA technique. An action can now be represented in terms of coefficients of a few basis vectors. Given a new action, recognition is done by a search process which involves calculating the distance between the coefficients for this action and the coefficients of every example action and choosing the minimum distance. Their method handles temporal variation (temporal shift and temporal duration) by parameterizing this search process using an affine transformation.

Our method differs in that an action is not represented by a single point in eigenspace but rather a manifold whose points correspond to the different feature images the action

goes through. This moves the burden of temporal alignment and duration adjustments from searching in the measurement space to searching in eigenspace. We see two main advantages for doing this:

1. Reduction in search complexity: Because the eigenspace has a much lower dimension than the measurement space, a more exhaustive search can be afforded.
2. Increased robustness: PCA is based on linear mapping. Action measurements are inherently nonlinear and this nonlinearity increases as these measurements are aggregated across the whole action. PCA can provide better discrimination if the action is not considered as one entity but a sequence of entities.

4.3. Recognition

Recognition is done by comparing the manifold of the test action in eigenspace to the reference manifolds. The computed manifold depends on the duration and temporal shift of the action which should not have an effect on the comparison. Our distance measure can handle changes in duration and is invariant to temporal shifts. Given two manifolds **A** and **B**, the distance is defined as the mean minimum distance between every normalized point in **A** and every normalized point in **B**. This distance measure is a variant of the Hausdorff metric (we use the mean of minima rather than the maximum of minima) which still preserves metric properties. Using this distance measure, three different classifiers have been considered:

1. Minimum Distance (MD): The test manifold is classified as belonging to the same action class the nearest manifold belongs to, over all reference manifolds. This requires finding the distance to every reference manifold.
2. Minimum Average Distance (MAD): The mean distance to reference manifolds belonging to each action class is calculated; and the shortest distance decides classification. This also involves finding the distance to every reference manifold.
3. Minimum Distance to Average (MDA) (also called nearest centroid): For each action, the centroid of all reference manifolds belonging to that action is computed. This is also a manifold with a number of points equal to the average number of points in each reference manifold belonging to the action. We do not interpolate to compute this manifold. Instead, the nearest points (temporally) on the reference manifolds are averaged to compute the corresponding point on the centroid manifold. A test manifold is classified as belonging to the action class with the nearest centroid. Testing involves calculating a number of distances equal to the number of action classes.



Figure 2 Several frames from Walk, Run, Skip, and March actions.

5. ACTION DATA

To evaluate our recognition method, we recorded video sequences of eight actions, each performed by 29 different people. Several frames from one sample of each action are shown in Figures 2 and 3. The actions are named as follows: Walk, Run, Skip, Line-walk, Hop, March, Side-walk, Side-skip. There are several reasons for our choice of this particular data set. Many of the actions we chose are very similar in the sense that the limbs have similar motion paths. This high degree of similarity among actions makes discrimination more challenging. Another reason is that rather than having the same person perform actions several times, we chose to have different people. This provides a more realistic data since in addition to the fact that people have different physical characteristics, they also perform actions differently both in form and speed. This would be a good test for the versatility of our approach. It can be seen from Figures 2 and 3 that people sizes as well as color of clothing are different. A few samples also had more complex backgrounds. Table 1 shows the variation in action performance speed throughout the data set. The table shows that the actions were performed at significantly varying speeds (more than double the speed in the case of Hop for instance). Another consideration for a more realis-

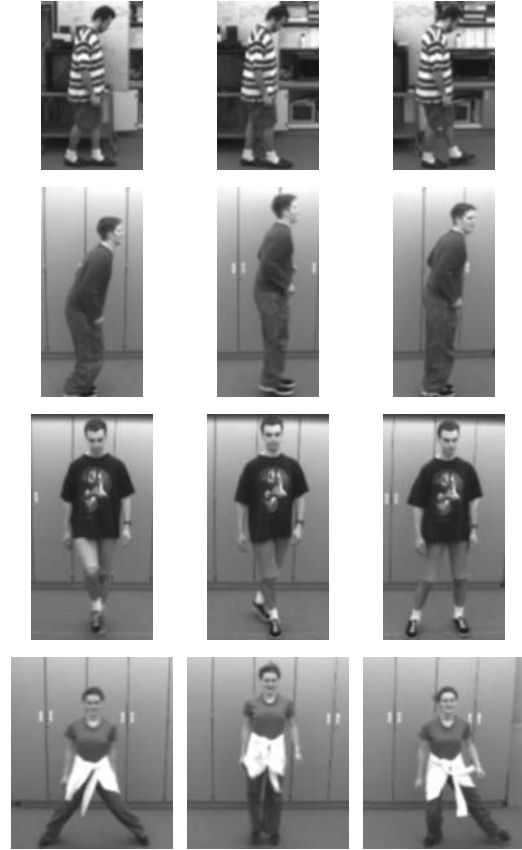


Figure 3 Several frames from Line-Walk, Hop, Side-walk, Side-skip actions.

Action	Minimum Duration (sec.)	Maximum Duration (sec.)
Walk	0.93	1.77
Run	0.70	0.93
Skip	1.10	1.73
March	1.13	1.93
Line-walk	1.47	2.20
Hop	0.70	1.67
Side-walk	1.06	1.80
Side-skip	0.57	0.93

Table 1: Variation in cycle duration for the data set.

tic data set was that we avoided the use of a treadmill. Using a treadmill not only restricts speed but also simplifies the problem since the background is static relative to the actor. To our knowledge, this is one of the largest sets

of action data ever used in terms of the number of subjects performing the actions multiplied by the number of actions.

The video sequences were recorded using a single stationary monochrome CCD camera mounted in such a way that the actions are performed parallel to the image plane. In our approach, we assumed that the height (in the image plane) and location of the person performing the action are known. Recovering location is necessary to ensure that the person is in the center of the feature images. Height is used for scaling the feature images to handle differences in people’s sizes and distance from the camera. To attain the recovery of these parameters, we tracked the subjects as they performed the action. Background subtraction was used to isolate the subject. A simple frame-to-frame correlation was used to precisely locate the subject horizontally in every frame. A small template corresponding to the top third of the subject’s body where little shape variation is expected was used. The height was recovered by calculating the maximum blob height across the sequence. For the general case, a tracking method as in [8] can be used to locate the subject boundaries.

6. EXPERIMENTAL RESULTS

In our experiments, we used the data for eight of the 29 subjects for training (64 video sequences). This leaves a test data set of 168 video sequences performed by the remaining 21 subjects. The training instances were used to obtain the principle components. The number of selected frames was arbitrarily set to 12. The resolution of feature images was also arbitrarily set to 25 horizontal pixels by 31 vertical pixels.

In our experiments, the choice of m (the number of eigenvectors to be used) was varied from 1 to 50. Using a small m is computationally more efficient but may result in a low recognition rate. As m increases, the recognition rate is expected to improve and approach a certain level. Recognition was done on the 168 test sequences as described in Section 4.3 using all three classifiers (MD, MAD, MDA). Recognition rate was computed as the ratio of number of samples classified correctly to the total number samples. Figure 4 displays the recognition performance for the different classifiers as a function of m . It can be seen that the recognition rate rises rapidly during the first few values of m . At $m = 14$, the rate using MDA reaches over 91.6%. At $m = 50$, the rate is over 92.8% for MDA. MAD performance is slightly lower while MD is about 10% below. One explanation for this behavior is that some clusters are close to each other so that a point, which may be classified correctly using MDA, can be misclassified using MD. Yacoob and Black [11] reported a recognition

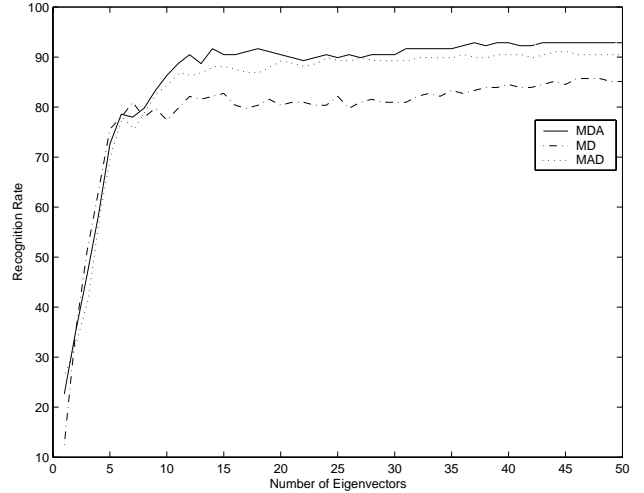


Figure 4 Recognition performance.

	W	R	S	M	LW	H	SW	SS
W	20	0	0	0	1	0	0	0
R	1	20	0	0	0	0	0	0
S	2	0	15	2	0	2	0	0
M	1	0	1	19	0	0	0	0
L	0	0	0	0	21	0	0	0
H	0	0	0	0	0	21	0	0
SW	0	0	0	0	1	0	19	1
SS	0	0	0	0	0	0	0	21

Table 2: Confusion matrix. Letters indicate actions in this order: Walk, Run, Skip, March, Line-walk, Hop, Side-walk, Side-skip.

rate of 82% and had four action classes.

Table 2 shows the confusion matrix for $m = 50$. Most actions had a perfect or near perfect classification except for the Skip action. Although the Skip action was classified correctly about 70% of the time, it was mistaken with Walk, March, and Hop actions numerous times. There were 12 misclassified actions in total. One person had two actions misclassified while the remaining people had at most one misclassification. When the correct action class was allowed to be within the first two choices, the number of misclassified actions becomes five. All these five actions (mostly Skip actions) were either executed erroneously or had a very low color contrast.

To give an indication of the quality of classification, Figure 5 shows a confusion plot which represents the dis-

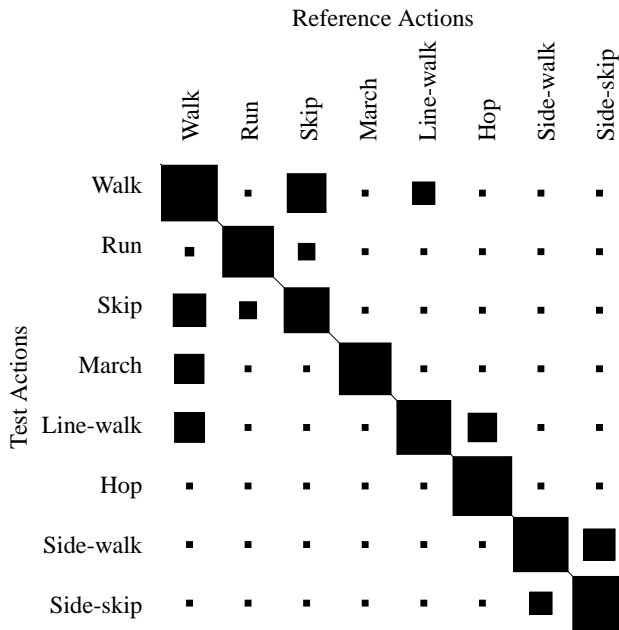


Figure 5 Confusion plot. The area of the squares indicates the distance using the distance measure in Section 4.3. The distances are averaged over all test samples.

subject. The larger the box size, the smaller the distance it represents. The diagonal in the figure stands out and very few other boxes come near the sizes of the boxes at the diagonal. However, it can be seen that there is mutual closeness in matching between Walk and Skip actions (a Walk action is close to a Skip action and vice-versa). This was expected due to the high degree of similarity between these two actions.

7. CONCLUSION

This paper describes a motion recognition approach. The approach is based on low level motion features which can be efficiently computed using an IIR filter. Once computed, motion features at every frame which we call feature images are compressed using PCA to form points in eigenspace. An action sequence is thus mapped to a manifold in eigenspace. A distance measure was defined to test the similarity between two manifolds. Recognition is performed by calculating the distances to some reference manifolds representing the learned actions. Experimental results for a large data set (168 test sequences) were presented and recognition rates of over 92.8% have been achieved. The results demonstrate the promise and efficiency of the proposed approach.

8. ACKNOWLEDGEMENTS

The work described in this paper has been funded in part by the ITS Institute (University of Minnesota) and the

National Science Foundation through grant IIS-0219863.

9. REFERENCES

- [1] L. Campbell and A. Bobick, "Recognition of human body motion using phase space constraints," in *Proc. of International Conference on Computer Vision*, pp. 624-630, Cambridge, 1995.
- [2] C. Cedras and M. Shah, "Motion-based recognition: a survey," *Image and Vision Computing*, vol. 13, no. 2, pp. 129-155, March 1995.
- [3] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," in *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 928-934, 1997.
- [4] D. M. Gavrila, "The visual analysis of human movement: a survey," *CVIU*, vol. 73, no. 1, pp. 82-98, January 1999.
- [5] D. M. Gavrila and L. S. Davis, "3-D model-based tracking of humans in action: a multi-view approach," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 73-80, San Francisco, 1996.
- [6] G. Halevi and D. Weinshall, "Motion of disturbances: detection and tracking of multi-body non-rigid motion," in *Proc. of IEEE Conference Computer Vision and Pattern Recognition*, pp. 897-902, June 1997, Puerto Rico.
- [7] S. Ju, M. Black, and Y. Yacoob, "Cardboard people: A parameterized model of articulated image motion," in *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 38-44, Killington, 1996.
- [8] O. Masoud and N.P. Papanikolopoulos, "A robust real-time multi-level model-based pedestrian tracking system", in *Proc. of ITS America Seventh Annual Meeting*, June 1997.
- [9] B. Maurin, O. Masoud, and N.P. Papanikolopoulos, "Camera surveillance of crowded traffic scenes", in *Proc. of ITS America Twelfth Annual Meeting*, Long Beach, CA, April 2002.
- [10] R. Polana and R. Nelson, "Detection and recognition of periodic, nonrigid motion," *International Journal of Computer Vision*, vol. 23, no. 3, pp. 261-282, 1997.
- [11] Y. Yacoob and M. J. Black, "Parameterized modeling and recognition of activities," *Journal of Computer Vision and Image Understanding*, vol. 73, no. 2, pp. 232-247, 1999.
- [12] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time sequential images using Hidden Markov Model," in *Proc. of IEEE Conference on CVPR*, pp. 379-385, 1992.