Recognizing Activities

Ramprasad Polana and Randal Nelson
Department of Computer Science
University of Rochester
Rochester, New York 14627
Email: polana@cs.rochester.edu and nelson@cs.rochester.edu

Abstract

The recognition of repetitive movements characteristic of walking people, galloping horses, or flying birds is a routine function of the human visual system. It has been demonstrated that humans can recognize such activity solely on the basis of motion information. We demonstrate a general computational method for recognizing such movements in real image sequences using what is essentially template matching in a motion feature space coupled with a technique for detecting and normalizing periodic activities. This contrasts with earlier model-based approaches for recognizing such activities.

1 Introduction

The motion recognition ability of the human visual system is remarkable. People are able to distinguish both highly structured motion, such as those produced by walking, running, swimming or flying animals and birds, and more statistical patterns such as those due to blowing snow, flowing water or fluttering leaves. The classic demonstration of pure motion recognition by humans is provided by Moving Light Display experiments [Johansson, 1973], where human subjects were able to distinguish activities such as walking, running or stair climbing, from lights attached to the joints of an actor. This biological use of motion probably reflects the fact that for certain tasks, visual motion provides more effective cues than other modes of visual perception. Motion is a particularly useful cue for certain types of recognition due to the fact that it is relatively easy to extract the motion field independent of illumination and shading of the image.

Motion recognition in general, has received little attention in the literature compared to the volume of work on static object recognition. Most computational motion work in motion in fact, has been concerned with various aspects of the structure-from-motion problem. A specialized area that has seen some attention is the interpretation of moving light displays [Goddard, 1989], [Rashid, 1980]. This work emphasizes rather high-level symbolic models of temporal sequences, an approach made possible by a discrete representation. The results are highly dependent on the ability to solve the correspondence problem and accurately track joint and limb positions.

A few studies have considered highly specific aspects of motion recognition computationally. Anderson et al. [Anderson *et al.*, 1985] describe a method of change detection for surveillance applications based on the spectral energy in a temporal difference image. Finally, there is a body of work based on the analysis of trajectories [Gould and Shah, 1989], [Allmen and Dyer, 1990] and [Tsai *et al.*, 1993]. All these require robust computation of the trajectories or spatiotemporal curves from image sequences before attempting recognition.

We define *activities* to be the motion patterns which are temporally periodic and possess compact spatial structure as opposed to *temporal textures* [Polana and Nelson, 1992] which exhibit statistical regularity but have indeterminate spatial and temporal extent. In this paper, we describe a robust method for recognizing activities, including ones, such as walking, that involve simultaneous translation of the actor. An earlier paper [Polana and Nelson, 1994], described an algorithm to detect periodic activities in an image sequence.

Motion recognition algorithms, both for temporal texture and activity, have potential applications in several areas. One area is automated surveillance. Motion detection via image differencing can be used for intruder detection; however such systems are subject to false alarms, especially in outdoor environments, since the system is triggered by anything that moves, whether it is a person, a dog, or a tree blown by the wind. Motion recognition techniques can be used disambiguate such situations. Another application is in industrial monitoring. Many manufacturing operations involve a long sequence of simple operations each performed repeatedly and at high speed by a specialized mechanism at a particular location. It should be possible to set up one or more fixed cameras that cover the area of interest, and to characterize the allowed motions in each region of the image(s).

2 Detecting Activities

The first step in recognizing an activity is to determine that an activity exists, and localize it in the scene. In an earlier paper we have described a technique for accomplishing this [Polana and Nelson, 1994]. The present work will utilize the information computed in the detection stage for recognition and classification of specific activities.

Activities involve a regularly repeating sequence of motion events. If we consider an image sequence as a spatiotemporal solid with two spatial dimensions x, y and one time dimension t, then repeated activity tends to give rise to periodic or semi-periodic gray level signals along smooth curves in the image solid. We refer to these curves as *reference curves*. If these curves can be identified and samples extracted along them over several cycles, then frequency domain techniques can be used in order to judge the degree of periodicity. We assume the object does not undergo any major rotation and the viewing angle does not change appreciably. A complete discussion of the periodicity detection process and the assumptions made can be found in the previously cited paper.

3 Recognizing Activities

Once an activity has been detected and tracked in a scene, the next step is to recognize it. The tracking and periodicity detection algorithms provide spatial and temporal normalization that can be used to simplify the recognition procedure. In particular, the periodicity detection procedure provides a periodicity measure for each active pixel in a tracked object. By backprojecting this measure, we can locate the pixels in each frame that display periodicity at the dominant frequency. We use this backprojection to refine our initial segmentation, which was based solely on aggregate motion. By fitting a frame to this refined segmentation we compensate for variation in spatial scale and position. Similarly, the fundamental frequency allows us to frame the activity in time, and compensate for variation in temporal scale (i.e. frequency).

The end result of the normalization procedure is a spatio-temporal solid containing the activity of interest in a form that is invariant to spatial scale, spatial translation, and temporal scale. The next step is to compute a descriptor for this solid that can be used to classify the activity it represents. It turns out that a three dimensional template match, with the appropriate motion features in the slots of the template, works well. Essentially, we capitalize on the fact that a periodic activity is characterized by regularly repeating motion events that have fixed spatial and temporal relationships to each other. Specifically, we divide one cycle of the spatio-temporal solid representing the activity into XxYxT cells by partitioning the two spatial dimensions into X, Y divisions respectively and the temporal dimension into T divisions. We then select a local motion statistic and compute the statistic in each cell of the spatiotemporal grid. The feature vector in this case is composed of XYT elements each of which is the value of the statistic in a particular cell.

The normalized spatio-temporal solid, while corrected for temporal scale (frequency) is not corrected for temporal translation (phase). Since the pattern matching phase of the algorithm currently represents only a small fraction of the total computational effort, and the temporal resolution of the pattern is typically small (i.e., less than 10 samples per cycle), we simply try a match at each possible phase and pick the best.

4 Experiments

Figure 1: Sample images from periodic activities: walk, run, swing, jump, ski, exercise and toy frog

We ran experiments on seven different activities, namely, walking on a treadmill viewed from side (**walk**), running on a treadmill viewed from side (**run**), swinging viewed from side (**swing**), skiing on a skiing machine viewed from side (**ski**) exercising on a machine - front view (**exercise**), performing jumping jacks - front view (**jump**), and a toy frog simulating swimming activity viewed from above (**frog**). The image sequences were first recorded on video and then digitized later with suitable temporal sampling so that at least four cycles of the activity were captured in 128 frames. All samples were digitized at a spatial resolution of 128x128 pixels, except those for walk and run which were digitized at a resolution of 64x128 pixels. Pixels were 8 bit gray levels. The swing and exercise activities were shot outdoors and contained background motion. Sample images of these activities are shown in figure 1.

We first digitized eight samples of each activity by the same actor under the same conditions with respect to scene illumination, background, and camera position. We created the reference database taking half (four) of the samples belonging to each activity. The remaining four samples of each activity are used to create the test database. In addition, we digitized four samples of walking by a different person and eight samples of the frog under different lighting conditions and different background and foreground gradients. These samples also differed from the reference database in frequency, speed of motion, and spatial scale. These samples were added to the test database.

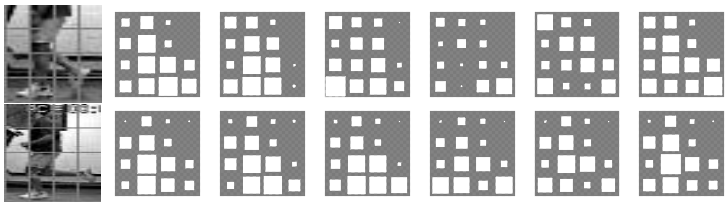

Figure 2: Sample total motion magnitude feature vector for a sample of walk (top) and a sample of run (bottom), one cycle of activity is divided into six time divisions shown horizontally, each frame shows spatial distribution of motion in a4x4 spatial grid (size of each square is proportional to the amount of motion in the neighborhood).

The samples in the test database were classified by a nearest centroid classification technique using the samples in the reference database as training set.

We experimented with several different local statistics. In each case the feature vector consisted of the local statistic computed over each of a set of cells constituting a partition of the spatio-temporal solid. We divided each spatial dimension into four divisions and the temporal dimension into six divisions, so that we get a feature vector of length 96. The simplest statistic we experimented with, is the summed normal flow magnitude in each cell. The normal flow direction information is ignored in this case. Sample feature vectors are illustrated in figure 2 using the total motion magnitude statistic for a walk and a run sequence.

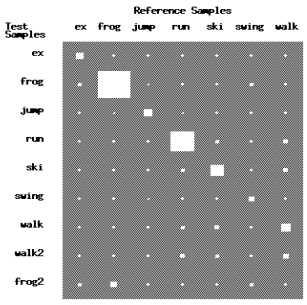

Figure 3: Confusion matrix for the feature vector using total motion magnitude

Somewhat to our surprise, the simplest statistic of total motion magnitude gave better results than the statistics involving direction of motion. The reason for this turned out to be related to the resolution of our images. In order to digitize enough frames to test the technique, we had subsampled the images to 128 x 128 pixels. After filtering for periodicity, significant motion, and direction, it was often the case that few pixels with all these properties were left in any one cell, which made for a large amount of stochastic noise in the signal. Simply put, we didn't have high enough resolution data to appropriately utilize the more specific statistics.

Using the total motion magnitude statistic, the classification resulted in correct classification of every sample in the test database, including the samples using a different actor and different backgrounds, which were not represented in the reference database. The percentage of correct classification does not give a full indication for the quality of classification. Hence, we illustrate the results by the confusion matrix which shows how closely test samples belonging to various classes match the reference samples of the different classes. The confusion matrix using the total motion magnitude statistic is shown in figure 3. A large square indicates a good match. As can be seen from this table, some motions, for instance the swimming frog, do not resemble anything else in the database, while others, for instance running and skiing, are more likely to be confused. The results seem to correspond more or less to human intuition about how similar the motions are.

5 Discussion

The following is a step-by-step description of the periodic activity recognition algorithm:

Input: The input to the algorithm is a digitized 256-level gray-valued image sequence consisting of at least four cycles of a periodic activity.

Output: A known class into which the activity is classified by the algorithm.

Step 1. Compute normal flow magnitude at each pixel between each successive pair of frames using a differential method.

Step 2. Locate and track the activity in the image sequence using periodicity detection algorithm described in section 2.

Step 3. Normalize the activity using pixels exhibiting periodic motion and compute a feature vector.

Step 4. Classify the activity using nearest centroid algorithm.

The method we have described displays several desirable invariances. It is robust to varying image illumination and contrast because the method uses only motion information which is invariant to these. It is also invariant to spatial and temporal translation and scale due to the normalization of the feature vectors, and the multiple temporal matching. It is also fairly robust with respect to small changes in viewing angle. The swing and exercise sequences were taken outdoors where there is a small amount of background motion. This comprises not only moving trees and plants, but also moving people and an occasional crossing of a car. That the activities can be detected even in this case demonstrates that the technique is somewhat tolerant

of background clutter and the occasional disturbance.

Added Clutter Percentage	*Total Test Samples*	*Successfully Detected*	*Correctly Classified*
25	4	3	3
50	4	3	3
75	4	2	0
100	4	2	0
150	4	1	0
200	4	0	0

Table 1: Classification results with motion clutter (samples are of walk)

To understand how much background clutter can be tolerated by this technique, we have experimented with the walk samples by adding motion clutter produced by blowing leaves This structured motion clutter is added in a controlled fashion so that its mean magnitude represents a varying percentage of the mean magnitude of the signal, and the resulting samples are classified using the total motion magnitude statistic. The results in table 1 show that the recognition scheme can tolerate motion clutter whose magnitude is equal to one half that of the activity, and it displays degraded, but still useful performance for even higher clutter magnitudes.

We have assumed that the actors giving rise to the activity move with constant velocity along linear paths. The case of nonlinearly moving objects can be handled by tracking the object of interest given a coarse estimate of its initial location and velocity, (e.g. with a Kalman filter). This would generate reference curves that are not straight lines. We have already demonstrated the usefulness of the centroid of motion for computing the velocity of linearly moving objects, and providing a rough initial segmentation. It could also be used for tracking the actors moving on more complex trajectories.

The detection scheme also assumes that there is only one activity in the scene except for some background clutter. If there are multiple activities in the scene, this detection technique can still be applied provided the activities can be spatially isolated so that they do not interfere with each other. In this case they can be segmented using the motion information and tracked separately. If a predictive tracker is used, an occasional crossing of different activities can be tolerated as long as the regions can be separated again later. In our experiments, the periodic activity samples consist of at least four cycles of the activity. Four cycles were needed to reliably detect the fundamental frequency given that there is a considerable amount of non-repetitive structure from the background in the case of translating actors.

The complexity of recognition is proportional to the number of pixels involved in the activity. More than half the work is computing the motion vectors at every pixel and then computing the fast Fourier transforms at each of moving pixels. The remaining time is spent computing the feature vector, the time for which depends on the local motion statistic computed. For a 128 image sequence, computation of the feature vector of motion magnitudes takes about 3 seconds. The classification algorithm currently runs on an SGI machine using four processors and it takes maximum 20 seconds to process a 128 frame sequence of 128x128 images.

6 Conclusion

We have described a general technique for periodic activity recognition. This technique uses a periodicity measure to detect the activity and then a feature vector based on motion information to classify the activity into one of several known classes. We have illustrated the technique using real-world examples of activities, and shown that it robustly recognizes complex periodic activities.

Acknowledgements

This work is supported by contracts NSF IRI-9010692 and AFOSR 91-0288.

References

[Allmen and Dyer, 1990] M. Allmen and C.R. Dyer. Cyclic motion detection using spatiotemporal surface and curves. In *Proc. Int. Conf. on Pattern Recognition*, pages 365–370, 1990.

[Anderson *et al.*, 1985] C. H. Anderson, P. J. Burt, and G. S. van der Wal. Change detection and tracking using pyramid transform techniques. In *Proc. SPIE Conference on Intelligent Robots and Computer Vision*, pages 300–305, 1985.

[Goddard, 1989] N.H. Goddard. Representing and recognizing event sequences. In *Proc. AAAI Workshop on Neural Architectures for Computer Vision*, 1989.

[Gould and Shah, 1989] K. Gould and M. Shah. The trajectory primal sketch: A multi-scale scheme for representing motion characterestics. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 79–85, 1989.

[Johansson, 1973] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.

[Polana and Nelson, 1992] R. Polana and R.C. Nelson. Temporal texture recognition. In *Proc. of CVPR*, pages 129–134, 1992.

[Polana and Nelson, 1994] R. Polana and R.C. Nelson. Detecting activities. *Journal of Visual Communication and Image Representation*, 5(2):172–180, 1994.

[Rashid, 1980] R.F. Rashid. *LIGHTS: A System for Interpretation of Moving Light Displays.* PhD thesis, Computer Science Dept, University of Rochester, 1980.

[Tsai *et al.*, 1993] P.-S. Tsai, M. Shah, K. Keiter, and T. Kasparis. Cyclic motion detection. Technical Report CS-TR-93-08, Computer Science Dept, University of Central Florida, 1993.