Activity Modeling and Recognition Using Shape Theory *

Rama Chellappa, Namrata Vaswani, Amit K.Roy Chowdhury
Center for Automation Research
University of Maryland
College Park, MD 20742.

Abstract

Understanding activities arising out of the interactions of a configuration of moving objects is an important problem in video understanding, with applications in surveillance and monitoring, animation, medicine, etc. In this paper, we introduce a novel method for activity modeling based on the observation that that every activity has with it an associated structure characterized by a non-rigid shape and a dynamic model that characterizes the variations in the structure as the activity unfolds. We propose two mathematical models to characterize the non-rigid shape and its dynamics. In our first approach, we propose to model an activity by the polygonal shape formed by joining the locations of these point masses at any time t, and its deformation over time. This uses the statistical shape theory of Kendall. The second approach models the trajectories of each separate class of moving objects in 3D shape space, and thus can identify different kinds of activities. It is based on the factorization theorem for matrices, which has been used before in computer vision for structure estimation. Deviations from the learned normal shape for each activity is used to identify abnormal ones. We demonstrate the applicability of our algorithms using real-life video sequences in an airport surveillance environment. We are able to identify the major activities that take place in that setting and detect abnormal ones.

1 Introduction

Modeling and recognition of human activities using a video sensor network poses many challenges. However, it successful solution has numerous applications in video surveillance, video retrieval and summarization, video-to-text synthesis, video communications, biometrics, etc. Recognizing activities is an extremely complicated task at which even humans are often less than perfect. Most of the early work on activity representation comes from the field of Artificial Intelligence (AI) [1, 2]. The formalisms that have been employed include HMMs, logic programming and stochastic grammars [3, 4, 5, 6, 7, 8, 9, 10, 11]. Many uncertainty-reasoning models have been actively pursued in the AI and image understanding literature, including Belief networks [12], Dempster-Shafer theory [13], and truth maintenance systems (TMS) [14, 15, 16, 17, 18, 19]. Computer vision based activity analysis algorithms have been proposed recently for video surveillance applications. In [20], the authors proposed building a tracking and monitoring system using a "forest of sensors" distributed around the site of interest. In [21], a method for recognizing events involving multiple objects using Bayesian inference has been proposed. In spite of the existence of so many methods, there is near unanimity in the vision community that a lot more needs to be done in order to be able to recognize complex activities from large amounts of video data.

We propose a novel approach to activity modeling using the non-rigid shape of the configuration of moving points, which can be separate point objects or different points on the same object. Our model is based on the observation that every activity has with it an associated structure characterized by a non-rigid shape and a dynamic model that characterizes the variations in the structure as the activity unfolds. We propose two mathematical models for represent the shape and its dynamics.

- The first approach [22] is based on statistical shape theory. The 2D or 3D shapes formed by the relative positions of entities participating in the activity being observed are modeled using Kendalls shape theory. Variations in shape as the activity occurs are characterized using a nonlinear dynamical model. An activity is recognized if it agrees with the learned parameters of the shape and dynamics associated with that activity. Sequential Monte Carlo methods are used to estimate the parameters of the shape model from the tracked points in the video sequence.
- The second method [23], based on subspace analysis, represents multiple activities in a video as a linear combination of 3D basis shapes. The basis shapes modeling each activity are extracted using multi-object non-rigid structure estimation. The structure and dynamical properties of normal activities are learned a-priori and

*Partially supported by a Grant from DARPA/ONR N00014-2-1-0809.

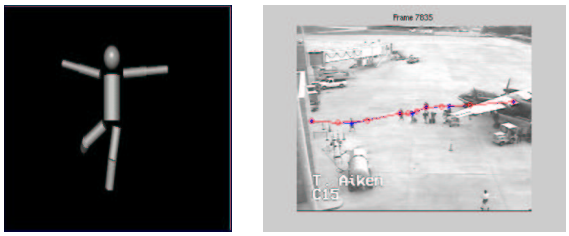
deviation from normal activities or recognition of the modeled activities is done using the learned models. The method, by virtue of its 3D representation, lends itself to easy extension to the situation when multiple cameras are looking at the same scene.

2 Justification for Shape-Dynamical Activity Model

The basic idea this paper builds on is that many activities have an associated structure and a dynamical model. Consider, as an example, a dancer or figure skater, who is free to move her hands and feet any way she likes. However, this random movement does not constitute the activity of dancing. For humans to perceive and appreciate the dance, the different parts of the body have to move in a certain synchronized manner. In mathematical terms, this is equivalent to modeling the dance by the structure of the body of the dancer and its dynamics. An analogous example exists in the domain of video surveillance. Consider people getting off a plane and walking to the terminal, where there is no jet-bridge to constrain the path of the passengers. Every person after disembarking, is free to move as he/she likes. However, this does not constitute the activity of people getting off a plane and heading to the terminal. The activity here is comprised of people walking along a path that leads to the terminal. Again, we see that the activity is defined by a structure and the dynamics associated with the structure. Using a shape-dynamical model is a higher level abstraction of the individual trajectories and provides a method of analyzing all the points of interest together, thus modeling their interactions in a very elegant way. In Figure 1, we show one frame for both of the above example activities. Such a structure is present in a large number of activities, e.g. sitting, walking, gymnastics, etc. and should be exploited for any modeling or recognition algorithm. In this paper, we will concentrate on the airport surveillance problem to demonstrate our results. In future, we plan to build a shape-based activity dictionary for modeling different kinds of activities.

3 Activity Recognition Using Statistical Shape Theory

As discussed in [22], we attempt to use Dryden and Mardia's statistical shape theory[24] ideas to model the shape formed by the locations of a group of moving objects and its deformations over time. We consider the example of passengers getting off or boarding a plane in the airport surveillance scenario (see Figure 1(b)). We consider the locations of the passengers in every frame of the video sequence and resample the curve connecting the passenger locations at time t to represent it by a fixed number of points,



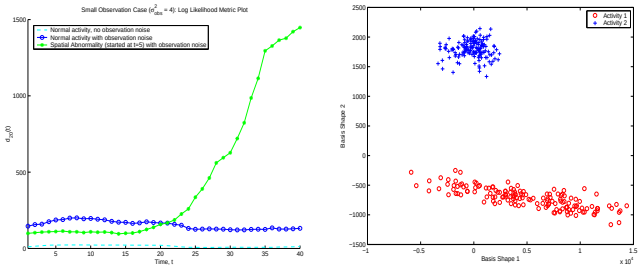
(a) (b)

Figure 1: Two examples of activities, (a) a dancer, (b) people disembarking from an airplane. It is clear that for both of these activities, there is an associated structure and it dynamics, which defines the activity.

k. The complex vector formed by these k points (x and y coordinate forming the real and imaginary parts) is considered. Certain pre-processing steps are carried out in order to normalize the observation vector for scale and location, as described in [24]. The resulting vector is referred to as the shape for the activity at that time instant. Because of the normalization, the space of all shapes is spherical. The mean, computed over all the shapes at different time instants, is learned. Each shape is projected onto the hyperplane, which is tangent to the shape space at the mean, in order to obtain a set of tangent coordinates, represented as v_t. The tangent coordinates lie in a $k-2$ dimensional complex space (one dimension is removed due to location normalization, the other due to the projection), which is equivalent to $2k-4$ dimensional real space. The mapping from the coordinate positions to the tangent hyper-plane is non-linear. The temporal evolution of the tangent projections is used to classify between various activities. In Figure 2(a), we plot the evaluation metric (see [22] for details) for the normal and abnormal activities (an abnormal activity occurs when a passenger deviates appreciably from the regular path followed by others.).

4 Activity Recognition Using Subspace Analysis

Our second approach [23] to activity recognition uses a 3D representation of the shape of the trajectory of each activity. In this case, not only are we able to recognize an abnormality, but are also able to verify each activity within a known class of activities. We hypothesize that each activity can be modeled by a basis shape corresponding to it. From training videos of the various activities, these basis shapes can be learned. The factorization theorem is used to compute the basis shapes and the motion parameters associated with them (see [23] for details). Considering the two activities of passengers deplaning or boarding and the luggage cart arriving or leaving, the plot of the various values of the



(a) (b)

Figure 2: (a) Plots of the evaluation metric, based on Kendall's shape theory, for normal and abnormal activities. (b) Plot of the projections of the various instances of the two activities onto the basis shapes learned using the subspace analysis method.

projections onto the basis shapes, learned from the different training examples, is shown in Figure 2(b), thus depicting the clear demarcation between the two activities. Given a test video sequence, the various activities can be identified, and an abnormal one detected, by computing the projections onto each of the basis shapes.

5 Conclusion

We have proposed two methods for representing activity in low-resolution surveillance video using shape theory. The idea of modeling activities using shape theory is based on the premise that that every activity has with it an associated structure characterized by a non-rigid shape and a dynamic model that characterizes the variations in the structure as the activity unfolds. In our first method, we model the dynamic configuration of objects by the shape formed by the locations of these objects at every time instant and their deformations over time. We use Kendall's statistical shape theory to model the activity as a 2D shape, along with its deformations. In the second method, we propose modeling the trajectory of each different class of moving objects. It represents each activity by a 3D basis shape and a set of permissible rotation matrices. The second method can be used to identify between different kinds of activities, in addition to detecting abnormalities.

References

[1] S. Tsuji, A. Morizono, and S. Kuroda, "Understanding a simple cartoon film by a computer vision system," in *Proc. Intl. Jt. Conf. on AI*, 1977, pp. 609–610.

[2] B. Neumann and H.J. Novak, "Event models for recognition and natural language descriptions of events in real-world image sequences," in *Proc. Intl. Jt. Conf. on AI*, 1983, pp. 724–726.

[3] H. Nagel, "From image sequences towards conceptual descriptions," *IVC*, pp. 59–74, 1988.

[4] C. Dousson, P. Gabarit, and M. Ghallab, "Situation recognition: Representation and algorithms," in *Proc. Intl. Jt. Conf. on AI*, 1993, pp. 166–172.

[5] Y. Kuniyoshi and H. Inoue, "Qualitative recognition of ongoing human action sequences," in *Proc. Intl. Jt. Conf. on AI*, 1993, pp. 1600–1609.

[6] H. Buxton and S. Gong, "Visual surveillance in a dynamic and uncertain world," *AI*, pp. 431–459, 1995.

[7] J. Davis and A. Bobick, "The representation and recognition of action using temporal templates," in *CVPR*, 1997, pp. 928–934.

[8] B. F. Bremond and M. Thonnat, "Analysis of human activities described by image sequences," in *Proc. Intl. Florida AI Research Symp.*, 1997.

[9] C. Castel, L. Chaudron, and C. Tessier, "What is going on? a high-level interpretation of a sequence of images," in *ECCV Workshop on Conceptual Descriptions from Images*, 1996.

[10] T. Starner and A. Pentland, "Visual recognition of american sign language using hidden markov models," in *Proc. Intl. Workshop on Face and Gesture Recognition*, 1995.

[11] A. Wilson and A Bobick, "Recognition and interpretation of parametric gesture," in *ICCV*, 1998, pp. 329–336.

[12] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, 1988.

[13] G. Shaffer, *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

[14] J Doyle, "A truth maintenance system," *AI*, pp. 231–272, 1979.

[15] V. Venkateswar and R. Chellappa, "Hierarchical stereo matching using feature groupings," *IJCV*, pp. 245–269, 1995.

[16] D. Ayers and R. Chellappa, "Scenario recognition from video using a hierarchy of dynamic belief networks," in *ICPR*, 2000, pp. 835–838.

[17] S. Intille and A. Bobick, "A framework for recognizing multi-agent action from visual evidence," in *Proc. AAAI*, 1999, pp. 518–525.

[18] P. Remagnini, T. Tan, and K. Baker, "Agent-oriented annotation in model based visual surveillance," in *ICCV*, 1998, pp. 857–862.

[19] T. Huang, D. Koller, J. Malik, G. Ogasawara, B. Rao, S. Russell, and J. Weber, "Automatic symbolic traffic scene analysis using belief networks," in *Proc. AAAI*, 1994, pp. 966–972.

[20] W.E.L. Grimson, L. Lee, R. Romano, and C. Stauffer, "Using adaptive tracking to classify and monitor activities in a site," in *CVPR*, 1998, pp. 22–31.

[21] S. Hongeng and R. Nevatia, "Multi-agent event recognition," in *ICCV*, 2001, pp. II: 84–91.

[22] N. Vaswani, A. RoyChowdhury, and R. Chellappa, "Activity recognition using the dynamics of the configuration of interacting objects," in *CVPR*, 2003.

[23] A. Roy Chowdhury and R. Chellappa, "A factorization approach for event recognition," in *CVPR Event Mining Workshop*, 2003.

[24] I.L. Dryden and K.V. Mardia, *Statistical Shape Analysis*, John Wiley and Sons, 1998.