GROUP ACTIVITY RECOGNITION BASED ON ARMA SHAPE SEQUENCE MODELING

Ying Wang, Kaiqi Huang, and Tieniu Tan

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
(wangying, kqhuang, tnt@nlpr.ia.ac.cn)

ABSTRACT

In this paper, we propose a system identification approach for group activity recognition in traffic surveillance. Statistical shape theory is used to extract features, and then ARMA (Autoregressive and Moving Average) is adopted for feature learning and activity identification. Here only a few points, instead of the complete trajectory of each object are used to describe the dynamic information of group activity. And ARMA is employed to learn activity sequences. The performance of the proposed method is proved by experiments on 570 video sequences, with the average recognition rate of 88% (compared with 81% of HMM). The extracted features are invariant to zoom, pan and tilt, which is also proved in the experiments.

Index Terms— Group Activity, Shape Theory, Landmark, ARMA, Surveillance.

1. INTRODUCTION

Group activity analysis is an extremely challenging task because of the large number of objects and the degree of freedom of their motion. Moreover, multiple occlusions and clutter occur in a crowded surveillance scene from a single camera, and thus low level features are not robust. Correlated research takes trajectories as low level features. To be general, they are not scene invariant. To get a the feature representation method able to work irrespective of the camera location, we adopt a solution base on shape theory [1].

Most of the prior work on activity learning is based on the research of graphical models, such as DBN, HMM and some extensions [2, **?**, 3, 4]. However, the more objects are considered, the more complex structure is constructed, and hence, the performance of these solutions gets worse because of low scalability. So the graphical model methods are feasible only when the number of objects is small. Typically, these methods require much priori knowledge, for example, the complete trajectory data, the number of objects and so on. When multiple occlusions and clutter occur, the information required is difficult to provide and then these methods are insufficient. So we try to develop a method using some disconnected points detected from the sequence. Compared with graphical models, ARMA has lower computation complexity in parameter estimation, and it can be used to characterize Gaussian distribution. Thus we use the ARMA model to learn the nature of the shape sequences.

The proposed method starts with the extraction of moving landmark points in each frame of a suquence. Then a fixed number of landmarks are resampled to represent the configuration of group activity. In addition, the original shape configurations are transformed into a linear tangent space by shape theory. Finally, ARMA model is used to capture the dynamics of activity shape over large training data. The different group activities could be recognized by the model parameters. More details can be seen in section 2. Section 3 illustrates the experiments with analysis. Section 4 concludes the paper. The overall system architecture is illustrated in Figure 1.

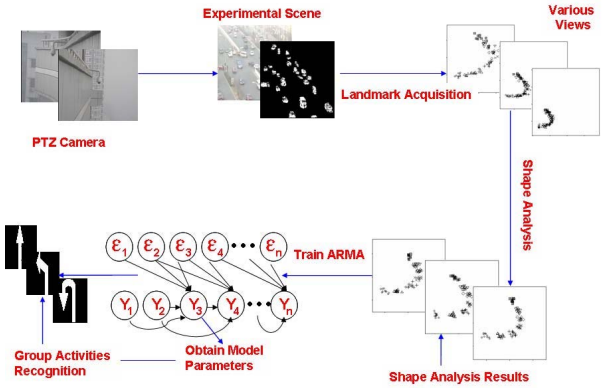

Fig. 1. The flowchart of ARMA-based activity recognition.

2. GROUP ACTIVITY RECOGNITION WITH SHAPE SEQUENCE AND ARMA

2.1. Theory of Shape Analysis

According to Kendall [6], shape is the geometrical information that remains when location, scale and rotational effects are filtered out from an object. In general, there are two kinds of representations of shape, one is continuous, such as contour and outline of object; the other is discrete, such as landmark. Here we use landmarks, a finite number of ordered points, to constitute a shape at each frame. The method begins by trans-

lating the configuration matrix so that its centroid is located at the origin:

$$y_c = Cy_{raw}, \tag{1}$$

where $C = I_k - \frac{1}{k}1_k 1_k^\tau$, 1_k is the $k \times 1$ vector of ones. y_{raw} is $k \times d$ configuration matrix for k points, and each point is denoted by d-dimension vector; Also to remove the size information, we scale y_c by its Euclidean norm:

$$y_{cs} = \frac{y_c}{\|y_c\|} = \frac{Cy_{raw}}{\|Cy_{raw}\|} \tag{2}$$

Given reference pre-shape γ, we can normalize each shape y into the same shape space by the rotation angle θ [6].

$$\theta(y_c, \gamma) = \arg(y_c^* \gamma) \tag{3}$$

$$y = \frac{y_c}{\|y_c\|} e^{j\theta(y_c,\gamma)} = \frac{Cy_{raw}}{\|Cy_{raw}\|} e^{j\theta(y_c,\gamma)} \tag{4}$$

Formally, the shape space is the orbit shape and it is nonlinear. So methods based on shape space are complex with bad performance. Therefore the Procrustes tangent coordinates, the linearization of shape space, is adopted [6] :

$$y_T = [I_k - \mu\mu^*]y = [I_k - \mu\mu^*]\frac{y_c}{\|y_c\|} e^{j\theta(y_c,\gamma)} \tag{5}$$

where μ is the Procrustes mean shape:

$$\mu = \arg \inf_{\mu} \Sigma d_F^2(y_i, \mu) \tag{6}$$

2.2. ARMA Model

To learn the variances of a shape sequence, a dynamical model from actual data is constructed. Current literature shows that the ARMA model has good performance in simulating the change of time sequences in space [7, 8]. Give an ARMA as defined by [9]:

$$\begin{cases} x(t+1) &= Ax(t) + Ke(t) \\ y(t) &= Cx(t) + e(t) \end{cases} \tag{9}$$

where $y(t)$ is the time series of tangent projections of shapes, $x(t)$ is the hidden state of model, $e(t)$ is zero mean white Gaussian noise process. A is the state transition matrix, C, the output matrix and K, the Kalman gain matrix of the innovation representation.

According to [10], we can use SVD (Singular Value Decomposition) to obtain the closed-form parameter matrices of ARMA model.

$$[y(1), y(2), ..., y(\tau)] = U\Sigma^{-1}V^\tau \tag{10}$$

then

$$C = U, \qquad A = \Sigma V^\tau D_1 V (V^\tau D_2 V)^{-1}\Sigma^{-1} \tag{11}$$

in which

$$D_1 = \begin{bmatrix} 0 & 0 \\ I_{\tau-1} & 0 \end{bmatrix}, \quad D_2 = \begin{bmatrix} I_{\tau-1} & 0 \\ 0 & 0 \end{bmatrix} \tag{12}$$

Once model parameters are obtained, we use the principal angles and their corresponding principal distances between ARMA models for recognition. According to [11], the principal angle θ_k are recursively defined as:

$$\cos\theta_k = \max_{a,b} \frac{|a^\tau A^\tau B b|}{\|Aa\|_2 \|Bb\|_2} = \frac{|a_k^\tau A^\tau B b_k|}{\|Aa_k\|_2 \|Bb_k\|_2},$$
$$\text{for k=1, 2 . . . q} \tag{13}$$

where a, b are the columns of parameter matrix A, B respectively. Let M_1 and M_2 be ARMA models of order n. The test sequence is identified as a predefined group activity, if the Frobenius norm based distance d_F is small enough :

$$d_F(M_1, M_2)^2 = 2\sum_{k=1}^{2n} \sin^2\theta_k \tag{14}$$

3. EXPERIMENTAL ANALYSIS

In the wide-area surveillance scene, every object moves at will and the group activity is irregular. For traffic scenes, customary crosswalks and traffic rules constrain the group activities. For example, right-and-left carriageways in the traffic scene have distinct directions. In this paper, we take a T-shaped intersection surveillance scene as an example and analyze five kinds of group activities as show in Figure 2.

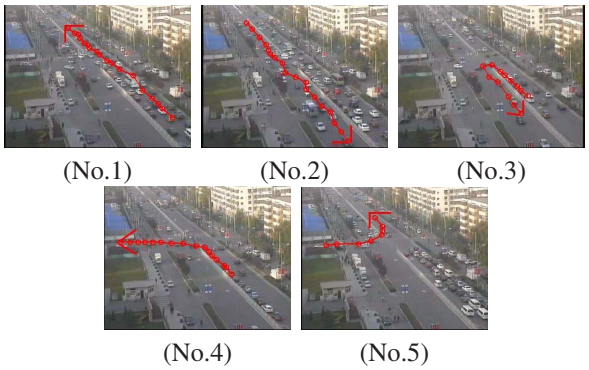

Fig. 2. Group activities: (No.1), straightforward from south to north; (No.2), straightforward from north to south; (No.3), back turn; (No.4), left turn from south to west; (No.5), left turn from west to north.

In our experiments, all images are 320×240 pixels in the video sequence from a surveillance platform which consists of sixteen fixed sensors and three PTZ cameras set around our campus. There are 570 training sequences in our dataset, including 90 straightforward from south to north cases, 90

straightforward case from north to south cases, 150 cases of turning from south to west, 120 cases of turning from west to north and 120 back turn cases. Each sequence is about 60 frames in length. Since shape theory can scale data from various views, three different views (standard view, zoomed view and translated view) are respectively used for the detection of different group activity.

3.1. Motion Detection and Low Level Feature Extraction

The most difficult problem in multi-objects tracking is how to match the object correctly, especially in the case of lighting changes, repetitive motions, or long-term scene changes. Therefore, landmark is introduced to represent the shape of group activity without matching each object accurately.

In Figure 3, we show experimental scene and motion detection results from different views obtained by our platform. Fig. 3(a) is the normal case, then we zoom in as shown in Fig. 3(b), translate the camera rightwards in Fig. 3(c). The red parts in (a), (b) and (c) denote the static vehicles waiting for left turn, so they are not detected in video. The corresponding motion detection results are illustrated in Figure 3(d, e, f). Note that detection algorithm is not the focus of this paper, we just use it to obtain required observation data. Detailed discussions can be found in [12]. The normalized data obtained by shape theory are shown in Figure 5. In each figure group, the first row demonstrates the sampled points in three different views: standard view (left), zoomed view (middle) and translated view (right). After the shape theory are used to these original landmarks, their respective shape data are shown in the second row.

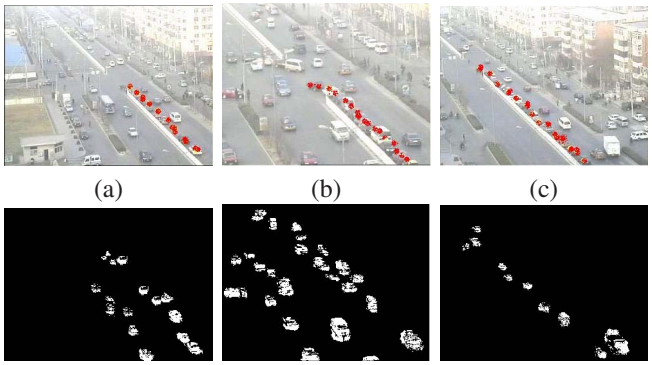

Fig. 3. Experiment scene where (a) is standard view, camera is zoomed and translated in (b) and (c) respectively. (d), (e) and (f) are their corresponding motion detection results.

3.2. Discussion on the Number of Landmark Sample Points

In our experiments, the locations of a group of vehicles are taken as landmarks. We linearly interpolate the group activity shape, and then re-sample the interpolated shape to obtain a predefined number of landmark points. Joining a number of landmarks in a predefined orientation will form a curve, which is used to represent the group activity (Figure 5). The relationship between recognition rate and the different number of landmark points 64, 32, 24, 16 and 8 points are shown in Figure 4. The recognition rate cease to rise when $N > 32$ in the case of straightforward, and the same thing occurs when $N > 24$ in the case of turning back, turning from south to west and turning from west to north. Although these can not represent the full information of these trajectory shapes, the recognition rate is still promising.

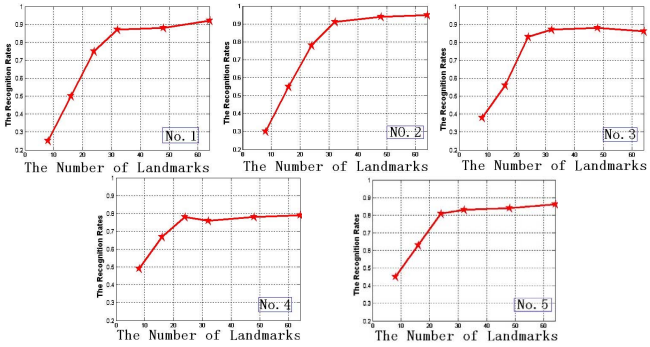

Fig. 4. Results of recognition vs the number of landmarks.

Table 1. The Number of Landmarks in Five Trained Group Activities.

	No. 1	No. 2	No. 3	No. 4	No. 5
Number	32	32	24	24	24

3.3. Comparison of ARMA, AR and HMM

HMM is popular in activity recognition because of its efficient parameter estimation algorithm and simple graphical structure. In discrete and continuous HMMs, the observation output is generally neither uniform nor Gaussian distributed, but in a more complicated form. Moreover, the discrete hidden state in a HMM may not be appropriate for characterizing continuous variations of a group activity. Different from HMMs, the distribution of an ARMA model is asymptotically Gaussian distributed. Concerning computational complexity, the training cost of an HMM is $o(N^2T^2)$, where N is the number of states in the HMM and T is the length of the time series. The cost of the ARMA estimation algorithm is $o(m^3T)$, where $m = max(p, q+1)$. In general, HMMs require more data for training, so that the HMM's computational cost is usually higher than the ARMA model.

AR is also commonly used in modelling the time sequences. However, the AR model only uses zeros to describe the deformation information which is quite insufficient while the

Table 2. Activity Recognition Rates for Three Models.

	No.1	No.2	No.3	No.4	No.5
AR	58%	63%	55%	57%	60%
ARMA	86%	92%	78%	88%	94%
HMM	81%	83%	80%	78%	82%

ARMA model uses both poles and zeros to characterize the shape sequences, providing enough information for activity analysis.

As shown in Table 2, HMM gets the promising recognition rates, and the performance of ARMA is comparatively better than HMM, but the results of AR is the worst. ARMA model is able to capture the dynamic shape deformation. Because of the inherent simplicity of AR, is not effective enough to model the dynamics of group activities. The parameter matrix C of ARMA resembles the observation probability matrix of HMM, so it can handle the highly structured shape sequence such as turning back.

4. CONCLUSION

Compared with existing work, the proposed method just samples the group activity curve, regardless of the complete trajectory data of respective object, and thus weakens the constraints of multiple occlusions. Motion detection, instead of tracking is enough for obtaining the required data. Moreover, the low level feature extracted by shape theory is invariant to camera zoom, pan and tilt. ARMA, as a linear parametric model to bridge the low level motion information and high level activity recognition, gets promising experiment results in the proposed method.

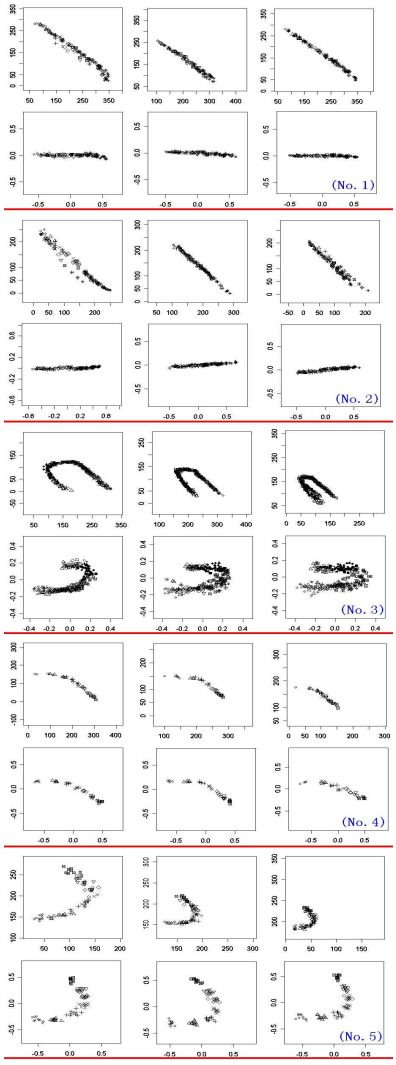

Fig. 5. Shape analysis for group activities in Figure 2.

Acknowledgment

The work reported in this paper was funded by research grants from the National Basic Research Program of China (No. 2004CB318110), the National Natural Science Foundation of China (No. 60605014, No. 60335010 and No. 2004DFA06900) and CASIA Innovation Fund for Young Scientists.

5. REFERENCES

[1] Dean C. Admas, F. James Rohlf, and Dennis E. Slice., "Geometric Morphometrics: Ten Years of Progress Following the Revolution", *Italian Journal of Zoology*, pp.17-25, 2004.

[2] Ying Luo, Tzong-Der Wu, Jenq-Neng Hwang, "Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks", *Computer Vision and Image Understanding*, Vol. 92, pp.196-216, 2003.

[3] Yamato, J., Ohya, J., Ishii, K., "Recognizing Human Action in Time Sequential Images Using a Hidden Markov Model", *CVPR*, 1992.

[4] M. Brand, N.Oliver, A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition", *CVPR*, 1997.

[5] N. Vaswani, A.R Chowdhury and R. Chellappa, "Shape Activity": A Continuous State HMM for Moving/Deforming Shapes with Application to Abnormal Activity Detection", IEEE Transactions on Image Processing, pp. 1603-1616, 2005.

[6] D. Kendall, D. Barden, T. Carne, and H. Le, "Shape and Shape Theory", *John Wiley and Sons*, 1999.

[7] Perrott MH, Cohen RJ, "An efficient approach to ARMA modeling of biological systems with multiple inputs and delays", *IEEE Transactions on Bio-medical Engineering*, vol. 43, 1996.

[8] Gaurav Aggarwal, Amit K. Roy Chowdhury and Rama Chellappa, "A System Identification Approach for Video-based Face Recognition", *ICPR*, 2004.

[9] Peter J, Brockwell and Richard A. Davis, "Introduction to Time Series and Forecasting", *Springer-Verlag New York*, 2002.

[10] P. Overschee, B. Moor, "Subspace Algorithms for the Stochastic Identification Problem", *Automatica*, vol. 29, 1993.

[11] K. D. Cock and D. B. Moor., " Subspace angles and distances between ARMA models", *In Proc. of the Intl. Symp. of Math. Theory of Networks and Systems*, 2000.

[12] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking", *CVPR*, pp. 246–252, 1999.