Detecting Group Activities using Rigidity of Formation [*]

Saad M. Khan
School of Computer Science
University of Central Florida
Orlando, Florida 32816
smkhan@cs.ucf.edu

Mubarak Shah
School of Computer Science
University of Central Florida
Orlando, Florida 32816
shah@cs.ucf.edu

ABSTRACT

Most work in human activity recognition is limited to relatively simple behaviors like sitting down, standing up or other dramatic posture changes. Very little has been achieved in detecting more complicated behaviors especially those characterized by the collective participation of several individuals. In this work we present a novel approach to recognizing the class of activities characterized by their rigidity in formation for example people parades, airplane flight formations or herds of animals. The central idea is to model the entire group as a collective rather than focusing on each individual separately. We model the formation as a 3D polygon with each corner representing a participating entity. Tracks from the entities are treated as tracks of feature points on the 3D polygon. Based on the rank of the track matrix we can determine if the 3D polygon under consideration behaves rigidly or undergoes non-rigid deformation. Our method is invariant to camera motion and does not require an a priori model or a training phase.

Categories and Subject Descriptors

[**Image Processing and Computer Vision**]: Activity Recognition, Scene Analysis

Keywords

Rigid Formations, Structure from Motion, Rank Constraint

1. INTRODUCTION

Modeling and recognition of human activities using video data poses many challenges. However, a successful solution has numerous applications in video surveillance, video retrieval and summarization, video-to-text synthesis, video communications, biometrics, etc. The task is further complicated when the activity is defined by the collective behavior of a group of entities. In such a scenario monitoring the activity of each participant separately might be unnecessary or even misleading for correct activity detection. It is the overall pattern that emerges from local interactions that characterizes a group activity. We propose a novel approach to recognizing group activities like people parades that are characterized by the rigidity in formation. By a formation is meant the 3D polygon emerging from the relative locations of a group of people/objects. A formation can be either rigid (i.e. maintaining structure) or deformable depending on the particular activity. Figure 1 demonstrates our idea of a formation of people. We model each walking person as a corner of a 3D polygon. 2D tracks from the walking people are treated as feature points on the formation. (For the purpose of this paper we assume that hand picked or accurately calculated tracking data is available). We demonstrate how a rank analysis of the tracking data leads to the classification of activities with rigidity in formation. One of the strengths of our method is its inherent property of invariance to changing view and camera motion. Changes in viewpoint affect the apparent motion and therefore complicate the analysis. Typically this problem is addressed by incorporating a view-invariant match function for comparing images [6]. This is not the case with our method of activity classification which is born into the framework of structure from motion [3 4]. Invariance to camera view is implicitly achieved by factoring out the camera and object motion as the relative pose.

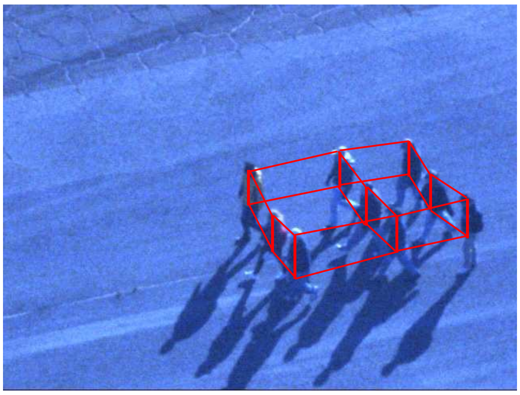


Figure 1: An example of a parade scene. The rigidity of the formation (the red polygon) characterizes the parade activity.

[*] This material is based upon the work funded in part by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

MM'05, November 6–12, 2005, Singapore.

2. RELATED WORK

Though there is plenty of work on single object activities [1, 2, 5, 6], the domain of group activities is relatively unexplored. The typical approach is to work bottom up by focusing on the activity of each object and how it interacts with others in the scene [2]. In contrast to this our methodology is top down. We consider the entire group as a collective (single entity) that performs a particular activity. We are interested in extracting the global patterns of behavior that a group demonstrates.

The closest work to ours was presented by Vaswani et al. [7]. Similar to our method they model the scene by the polygonal shape of the configuration of the participants. They learn the mean shape and define a change detection statistic for detecting abnormal behavior. Their method differs from our approach in two respects. Firstly in [7] the authors model a group activity with its projected 2-D shape. As mentioned earlier changing view will affect the apparent motion and the performance of their method. Our method on the other hand models the actual 3D formation of the group activity and is invariant to changing views. Secondly and more importantly their approach is to identify the deviations from a learned pattern that is very specific to the particular conditions. Thus their method is more suited to monitoring and surveillance scenarios where a *deviation* from normal behavior is of interest rather than the actual recognition and categorization of an activity. We on the other hand propose a general framework for the recognition and classification of group activities like parades as demonstrated in this paper. Our approach targets the patterns that are unique to a particular group activity without the need for any prior training.

3. APPROACH

A parade is a common example of an activity that maintains rigidity in formation. Other examples could be synchronized flights, bird flocks or military convoys. We demonstrate our approach using the example of a group of people parading together. (see figure 1 and 3).

A set of P points on the marchers are tracked across F frames with coordinates: $\{(u'_{fp}, v'_{fp}) \mid f = 1 \ldots F, p = 1 \ldots P\}$. The point coordinates are transformed to object-centered coordinates by subtracting their center of mass: $(u_{fp}, v_{fp}) = (u'_{fp} - \bar{u}_f, v'_{fp} - \bar{v}_f)$ for all f and p, where $\bar{u}_f$ and $\bar{v}_f$ are the means of point positions in each frame. The tracking matrix can be constructed as $W^{2F\times P} = \begin{bmatrix} U \\ V \end{bmatrix}$, where matrices U and V are defined as follows:

$$U = \begin{bmatrix} u_{11} & \ldots & u_{1P} \\ \vdots & \vdots & \vdots \\ u_{F1} & \ldots & u_{FP} \end{bmatrix}, V = \begin{bmatrix} v_{11} & \ldots & v_{1P} \\ \vdots & \vdots & \vdots \\ v_{F1} & \ldots & v_{FP} \end{bmatrix}$$

If the tracks belong to a rigid formation (i.e. it can be treated as a rigid object) then it has been shown by Tomasi and Kanade [3] that if the camera is affine (orthographic or weak-perspective) and when there is no noise, then the rank of W is 3 or lower. This constraint arises because the W matrix can be factored into 3D pose matrix $R^{2F\times 3}$ and 3D shape matrix $S^{3\times P}$ [3] i.e. $W = RS$.

When there is noise in the measurement matrix W then the maximum likelihood estimate is obtained by minimizing the squared error:

$$\varepsilon(R, S) = \| W - RS \|^2, \qquad (1)$$

where $\| \cdot \|$ denotes the Frobenius norm.

The global minimum to this non-linear problem is obtained by performing singular value decomposition (SVD) to matrix W. All singular values other than the 3 largest are set to zero and the matrices produced in the SVD step are recomposed to yield the matrices R and S.

When the tracking data W belongs to a formation with non-rigid deformation it can again be factored into 2 matrices [4], but of rank r that is higher than the bounds for the rigid case. Assuming the 3D non-rigid deformation can be approximated by a set of K modes of variation, the 3D shape of a specific object configuration can be expressed as a linear combination of K basis-shapes: $(S_1, S_2, \ldots, S_K)$[4]. The shape at any time instant is given by a linear combination of the basis-shapes:

$$S = \sum_{i=1}^{K} l_i \cdot S_i, \qquad S, Si \in \mathbf{R}^{3\times P}, l_i \in \mathbf{R} \qquad (2)$$

Assuming weak perspective projection at frame t the P points of a shape S are projected onto image points:

$$\begin{bmatrix} u_{t1} & \ldots & u_{tP} \\ v_{t1} & \ldots & v_{tP} \end{bmatrix} = R_t \cdot \sum_{i=1}^{K} l_{i,t} \cdot S_i, \qquad (3)$$

where $R_t = \begin{bmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \end{bmatrix}$, the first two rows of the full 3D rotation matrix. This can be re-written as:

$$\begin{bmatrix} u_{t1} & \ldots & u_{tP} \\ v_{t1} & \ldots & v_{tP} \end{bmatrix} = \begin{bmatrix} l_{1,t}R_t & \ldots & l_{K,t}R_t \end{bmatrix} \cdot \begin{bmatrix} S_1 \\ \vdots \\ S_K \end{bmatrix} \qquad (4)$$

Therefore W can be factored as follows [4]:

$$W = \underbrace{\begin{bmatrix} l_{1,1}R_1 & \ldots & l_{K,1}R_1 \\ l_{1,2}R_2 & \ldots & l_{K,2}R_2 \\ & \ldots & \\ l_{1,F}R_F & \ldots & l_{K,F}R_2 \end{bmatrix}}_{Q} \cdot \underbrace{\begin{bmatrix} S_1 \\ S_2 \\ \ldots \\ S_K \end{bmatrix}}_{B} \qquad (5)$$

Since Q is a $2F \times 3K$ matrix and B is a $3K \times P$ matrix, in the noise free case W has a rank $r \leq 3K$. It can be observed that this is a generalization of the rigid factorization ($K = 1$ reduces it to rigid factorization). Given the tracks of walking people our goal is to find the smallest value of K that minimizes the re-construction error given in equation 1. In other words, we seek the *true* rank of the matrix W.

Given the rank we can ascertain the rigidity of the formation and hence classify it as a parade activity or not. Rank measurements are highly sensitive to numerical errors and our method provides a robust and reliable way of overcoming this problem. The factorization of the tracking matrix into two matrices; one defining the relative poses of the camera and the object and the other describing the shape itself also yields us the useful property of invariance to camera motion.

All experiments reported here assume weak perspective projection.Weak perspective projection is in practice a good approximation if the perspective effects between the closest and furthest point on the object surface are small. Also we

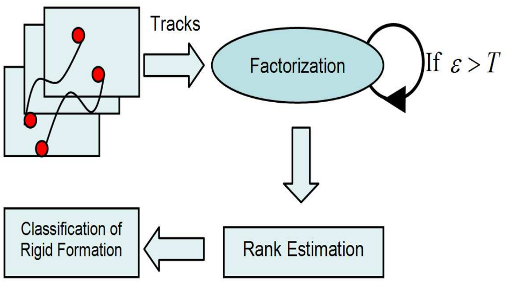


Figure 2: The diagram gives an outline of our proposed method for classification of rigidity in formation.

are assuming that the camera's intrinsic parameters like the focal length remain unchanged. Therefore sequences with zoom in or zoom out might affect results.

4. ALGORITHM

Figure 2 gives an outline of our algorithm. In the first step we obtain tracks and construct the W matrix. Next we factorize W with an increasing number of basis shapes till the reconstruction error falls below the noise threshold. Non-rigid factorization of W as defined by (5) is not a trivial task. We adopt the iterative least squared (ILSQ) method proposed by Torresani et al. [4]. The method initializes Q and B with a rough estimate (usually with rigid factorization of W) and then iterates between finding least squared fit for the elements of matrices Q and B till convergence is reached. Readers interested in the details of this method are directed to [4]. An outline of our algorithm for evaluating the number of basis shapes K that span the tracking matrix W is as follows:

```
1-Initialize $K = 1$
2- Factorize $W = QB$ using ILSQ
3- Evaluate re-construction error:
        $\varepsilon = \| W - QB \|^2 / n$
4- if $\varepsilon > T$ then:
        $K = K + 1$ and go to step 2
   else
        Return $K$
```

Here n is a normalizing variable equal to the number of elements of W, so in effect step 4 calculates the average reconstruction error for a tracked point in each frame. The threshold T is set to allow for some noise tolerance. The value of K returned by the algorithm describes the true rank r of the matrix W such that: $3(K-1) \leq r \leq 3(K)$.

5. RESULTS

We tested our method on tracks obtained from various videos of crowds of people walking. Figure 3 shows 4 (one in each row) of these sequences. Each sequence contains 50 frames and the number of people tracked varied between 4 and 8. We used 50 frames as a reasonable duration over which the rigidity in formation must be maintained. We also tested with different numbers of feature points on each person. Having many feature points did not significantly affect the results so for the purpose of simplicity and consistency we chose only 2 feature points on each participant. It must be noted that feature points should be chosen so that they best represent the world locations of the participants. Points on the head and the torso are therefore preferred. The camera was non-stationary in all of the sequences especially in sequence 2 where there was significant camera motion. Notice the variety of view angles and orientations.

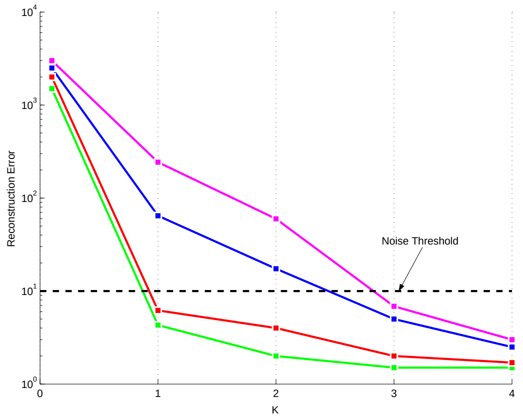


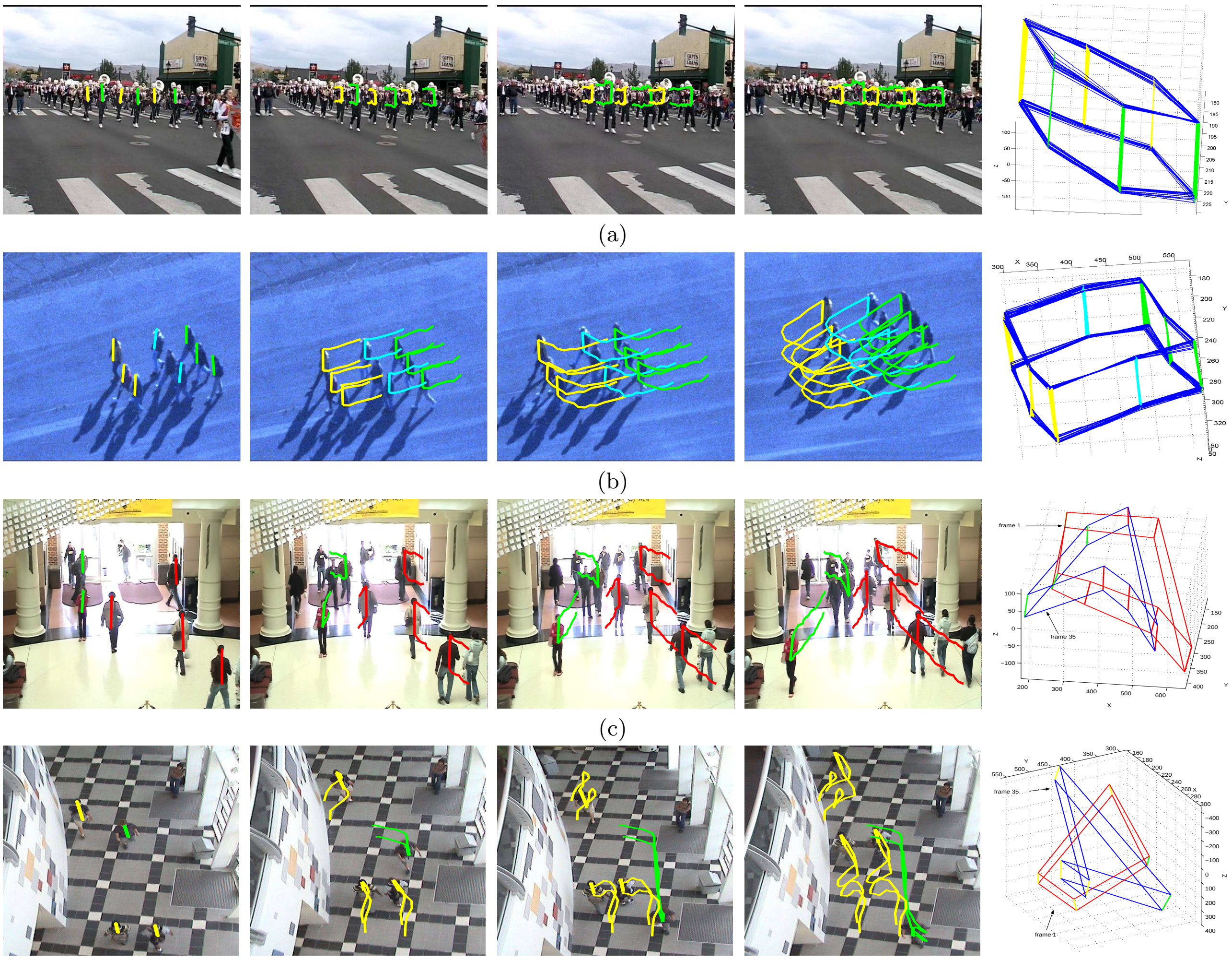
Figure 4: A plot of reconstruction error vs. K the number of basis shapes. Legend is as follows: Seq1-green, Seq2-red, Seq3-blue, Seq4-magenta

Figure 4 shows the plot of reconstruction error against increasing values of K. The plot depicts the value of K that is needed to reduce the reconstruction error below the noise threshold T (we use T=10 based on empirical observations). For sequences 1 and 2 (red and green lines), K=1 basis shapes i.e. rigid factorization minimizes the reconstruction error to below the noise threshold. Thus the crowd shows rigidity in formation and is correctly classified as a parade. On the other hand sequences 3 and 4 require K=3 basis shapes to reduce the reconstruction error below noise threshold. This implies that the rank of the track matrices (for seq 3, 4) is higher than 3, which violates the rigidity constraint. Another way to interpret this result is that the formations in sequences 3 and 4 are undergoing non-rigid deformation. This is in contrast to the pattern of rigidity that is characteristic of parade behavior. Therefore such sequences are classified as not containing parade activities.

For the purpose of visualization we also recovered the shape from motion for the crowd formations using K=3 bases shapes to model any deformations [4]. Note that it is not necessary to construct the 3D structures, as demonstrated in this paper only a study of rigidity based on rank is sufficient. The recovered 3D formations are shown as the last plot in each row of figure 3. For sequences 1 and 2 (figure 3(a), 3(b)) the formations at each time frame (total of 50 frames) are overlaid to view any deformations. As expected there is very little variance in the shape indicating the rigidity of the formation. The formation in sequences 3 and 4 (figures 3(c), 3(d)) undergoes non-rigid deformation and for the sake of clarity reconstructions at only two frames (1 and 35) are shown. The reconstructed formations are color coded to match the tracks of the corresponding people.

6. CONCLUSION

We have proposed a novel approach of modeling and classifying group activities based on the structure emerging from



(a)

(b)

(c)

(d)

Figure 3: Each row contains frames from a video sequence of multiple people walking. The frames are overlaid with tracks of the participating people. The first two sequences are parades while the last two sequences are random crowds.

local interactions between participants. We showed how a parade activity can be classified as a rigid formation of walking people. To reach this end we used rank constraint and an analysis of number of basis shapes needed to model the deformation in the parade formation.

In the future we plan to extend our approach to a variety of activities like monitoring flight formations of unmanned air vehicles, and herd behavior in animals etc. Classifying different activities based on the shape of their formations is another interesting direction of research. Perhaps exploiting the periodic nature of parade activities would also be useful.

7. REFERENCES

[1] 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.

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

[3] C. Tomasi and T. Kanade. "Shape and motion from image streams under orthography: a factorization method." Int. J. of Computer Vision, 9(2):137-154,1992.

[4] L. Torresani, D. B.Yang, E. J.Alexander, C. Bregler. "Tracking and Modeling Non-Rigid Objects with Rank Constraints." In CVPR 2001

[5] Ramprasad Polana and Randal C. Nelson. "Detection and recognition of periodic, non-rigid motion." Int. Journal of Computer Vision, 23(3):261-282, 1997.

[6] S. M. Seitz and C. R. Dyer. "View-invariant analysis of cyclic motion." Int. Journal of Computer Vision, 25(3), 1997

[7] Namrata Vaswani, A. RoyChowdhury, Rama Chellappa, "Activity Recognition Using the Dynamics of the Configuration of Interacting Objects", IEEE Computer Vision and Pattern Recognition (CVPR), 2003.