SMALL GROUP HUMAN ACTIVITY RECOGNITION

Yafeng Yin, Guang Yang, Jin Xu, Hong Man

ECE Department, Stevens Institute of Technology, Hoboken, NJ, 07030

ABSTRACT

Small group people activity recognition has attracted much attention in computer vision community in recent years, since it has great potential in public security applications. Comparing to single human activity recognition, group human activity recognition has much more challenges, such as mutual occlusions between different people, the varying group size, and the interaction within or between groups. In this paper, we propose a novel structural feature set to represent group behavior as well as a probabilistic framework for group activity learning and recognition. We first apply a robust multiple targets tracking algorithm to track each individual in the entire image region. Small groups are then clustered based on the output positions of the tracker. After that, we introduce a set of social network analysis based structural features to describe the dynamic behavior of small group people in each frame. A Gaussian Process Dynamical Model(GPDM) is then employed to learn the temporal activity of small group people overtime. After training, the new group activity will be identified by computing the conditional probability with each learned GPDM. Our experimental results indicate that our proposed features and behavior model can successfully capture both the spatial and temporal dynamics of group people behavior, and correctly identify different group activities.

Index Terms— human group action recognition,social network features, GPDM,

1. INTRODUCTION

Human action recognition has been studied for decades in the computer vision filed. Most current human action recognition research[1, 2, 3] focus on single human action identification, their experiment results showed most algorithms could achieve a very high recognition rate on two popular data sets: Weizmann human action dataset [4] and KTH human action dataset [5]. As the importance for public safety increases, recognizing interactions among different people, especially small human group(around 10 people)activity recognition has attracted much attention in human action recognition [6, 7, 8].

Different with single person action recognition, small groups often contain much richer inter-person interactions. Compared to crowd analysis, in which each person can be regarded as a point in a flow, a small human group includes much detail information about each individual in the group. Major challenges of small group activity analysis include mutual occlusions between different people, the varying group size, and the interaction within or between groups. Therefore small group activity recognition demands a descriptive feature to bridge the local description of single human and global description for crowd analysis, as well as addressing both the spatial dynamics (varying group size) and temporal dynamics (varying clip length).

In this paper, we propose a novel structural feature to represent group activities as well as a probabilistic framework for small group activity learning and recognition. Our framework consists of four stages. First, we apply a robust mean-shift [9] based tracker to locate each individual in a small group consequently. Second, the output coordinates of each tracker will be clustered to different small groups. By social network analysis based feature description, we extracted the structural features from each video clip in the third stage. Those feature vectors contain global structure of each group as well as local motion description of each group member, and they all have same size regardless the different number of people inside each group.In the last stage, the feature vectors from each frame will form a feature matrix for each video clip. A Gaussian process dynamical model is trained to model different group behaviors respectively. The group activity matrix will be projected to a low dimensional latent space and get a compact representation. A posterior conditional probability is compute with each trained model to identify different group behaviors. We validate our framework on two publicly available data set: BEHAVE data set [10] and IDIAP data set [11].

Our main contributions are listed as follows: First of all, we proposed a social network structure based feature set to represent dynamical of small group people. The structural feature characterizes both the global group distribution as well as local motion of each individual. In addition, this feature can keep a fixed length while handling vary number of members in a group, which is very helpful in recognition. Secondly, we established a probabilistic framework for human behavior classification. Different specific GPDM is trained for each group activity, then the conditional probability is computed for the new coming activity feature, and the one with the highest probability is selected as the group activity type. As there is no length constraint for input training

and testing feature, this GPDM based recognition framework can handle video clips with different length. Therefor, our proposed model can represent the dynamical characteristic of the similar activities with different duration.

The rest of this paper is organized as follows: Section 2 will describe the social network based feature for action representation. Section 3 presents the overall framework for small group human action recognition. All the experimental results and comparison are shown in the Section 4. Finally, conclusion and future work are summarized in the Section 5.

2. SOCIAL NETWORK ANALYSIS BASED FEATURE SET

As discussed in the previous section, feature extraction plays an essential role in the small group action recognition. Most features used for human action recognition fall into two big categories: general low level feature and middle level feature. General low level feature includes human motion, optical flow, 3D SIFT [12](Scale Invariant Feature transform) or STIP [13](Spatial Temporal Interest Points), which are directly computed on the entire image region. General low level features are good for single person action classification. As small group human behavior involving interactions between different members, it needs features capture local detail information as well as global structure description. Thus middle level feature [6], which characterizes the group structure information above general low level feature, has been developed for small group human action recognition.

Social network analysis [14] is originally designed to model the social structure of individuals and relationships among people in real world societies. It maps the social individuals or "actors" as *nodes* and relationships between them as *ties* into a graphic based network. Inspired by the social network analysis, we proposed several features similar to social network analysis to capture the dynamic properties of a small group structure. To our best knowledge, this is the first time that social network analysis is used to model group behavior in the surveillance videos. Similar to the original definitions of betweenness, Closeness, and Centrality[15, 16] in social network analysis, we define several group structure features for human group activity recognition.

1. *group center:* Suppose there are n people in a group, group center $m = \left(\frac{1}{n}\sum_{i=0}^{n} x_i, \frac{1}{n}\sum_{i=0}^{n} y_i\right)$ is defined as the mass center of the group.

2. *motion histogram:* Motion vector of a person in a group $\mathbf{M}_t = \{m_i\}_t (i = 1, \ldots, n)$ is defined as the position difference of each individual between two consecutive frames. The magnitude of m_i is then accumulated into orientation histograms and normalized at each direction. As the orientation has been divided to 8 bins, the motion histogram is a 8-dimension vector for each group in each frame.

3. *closeness histogram:* Closeness describes how close an individual is near to all the other nodes, directly or indirectly in a network. In our experiment, closeness vector $\mathbf{C}_t = \{c_i\}_t (i = 1, \ldots, n)$ is defined as the directional vector between every two different people. Similar to motion histogram, the magnitude of c_i is accumulated into 8-bin orientation histograms and normalized at each direction.

4. *centrality histogram:* Centrality was originally used for describing the overall network structure based on each node's location in a network. Group centrality vector $\mathbf{Ce}_t = \{ce_i\}_t (i = 1, \ldots, n)$ is defined as the directional vector which from the position of each person toward the group mass center. Similar as motion histogram, the magnitude of c_i is accumulated into 8-bin orientation histograms and normalized at each direction.

As described above, a 26 dimensional feature vector is extracted from each frame, including group center, motion histogram, closeness histogram and centrality histogram. Suppose the length of a group activity clip(total frame number) is m, then the size of the feature matrix is $26 \times m$.

3. PROPOSED FRAMEWORK FOR BEHAVIOR CLASSIFICATION

The proposed small group activity recognition framework consists of four stages: adaptive mean-shift tracking, small group clustering, group feature extraction and group activities recognition.

3.1. Adaptive Mean-shift Tracking

One of the important factor for small group human activities analysis is the accuracy and robustness of tracking each individual in the group. As the development of multiple camera systems, the accurate tracking of each individual can be well addressed. In this paper we apply adaptive mean-shift tracking[9] on the two data sets.

Compared to general mean-shift tracking, on-line feature selection is applied during the adaptive mean-shift tracking. In [9], the feature consisted of linear combination of pixel valves at R, G, B channels: $F \equiv \omega_1 R + \omega_2 G + \omega_3 B$, where $\omega_i \in [-2, -1, 0, 1, 2], i = 1, \ldots, 3$. By pruning all redundant coefficients of ω_i, the feature set was cut down to 49. Linear discriminative analysis (LDA) was then used to determine the most descriptive feature for target tracking. To reduce the computational complexity, we just update the feature set every 50 frames instead of updating at each frame. In addition, we extend the single mean-shift tracking algorithm for multiple targets tracking. As the cameras were fixed in these two data sets, a simple motion detector is applied to detect each new person coming into scene. Once a person comes in the

scene, a new tracker will be allocated and track that person overtime. Since our focus of this paper is not reliable multiple targets tracking, we just reinitialize each target manually if the tracking algorithm fails for some reason.

3.2. Small Group Clustering

After obtaining all the positions of each target, a group clustering algorithm [8] will be applied to locate small groups. We first calculate the closeness of each person and use the Minimum Span Tree (MST) clustering to obtain the distribution of each group.After that, we follow the hierarchical clustering method described in [8] to locate the mass center of each small group.

3.3. Small Group Activity Recognition

Gaussian Process Dynamical Model [17] was derived from Gaussian Process Latent Variable Model(GPLVM) [18], which provided a probabilistic mapping from high-dimensional observation data to low-dimensional latent space and represented the joint distribution of observation data. The small group activity recognition can be divided to two phases: group activity training and group activity classification. In the training stage, a GPDM $\{\Lambda_i, i = 1, \ldots, n\}$ will be trained for each small group activity $\{A_i, i = 1, \ldots, n\}$. Suppose we have k samples of a group activity A_i, the length of each each sample is m, then we have k feature matrices of size $26 \times m$. To learn a specific GPDM for A_i, we will first compute the mean value $\bar{Z}$ of k feature matrices, and utilize the mean for training.

GPDM is applied to learn the specific trajectories of a group activity.The probability density function of latent variable X and the observation variable $\bar{Z}$ are defined by the following equations. The basic procedure Gaussian Process Dynamical Model training is described as below:

1. *Creating GPDM:* GPDM $\Lambda = \{\bar{Z}^T, X^T, \bar{\alpha}, \bar{\beta}, \Omega\}$ is created on the basis of the trajectory training data sets, i.e. extracted structural feature, where $\bar{Z}^T$ is the training observation data, X^T is the corresponding latent variable sets, $\bar{\alpha}$ and $\bar{\beta}$ are hyperparameters.

2. *Jointly initializing the model parameters:* The latent variable sets and parameters $\{X^T, \bar{\alpha}, \bar{\beta}\}$ are obtained by minimizing the negative log-posterior function $-lnP(X^T, \bar{\alpha}, \bar{\beta}, \Omega|\bar{Z}^T)$ of the unknown parameters $\{X^T, \bar{\alpha}, \bar{\beta}, \Omega\}$ with scaled conjugate gradient (SCG) on the training datasets.

3. *Train GPDM for each group activity:* For each group activity $\{A_i, i=1,\ldots,n\}$, repeat the procedure 1 and 2, create a corresponding GPDM: $\{\Lambda_i, i = 1, \ldots, n\}$.

After training, we have a set of GPDMs: $\{\Lambda_i, i = 1, \ldots, n\}$ for the human group activities. When a new human group activity $Z^\star$ coming in, we will compute the conditional probability with respect to each trained GPDM, and select the one with highest conditional probability.

1. *Calculate the conditional probability with each trained GPDM:* For each trained GPDM $\{\Lambda_i\}$, compute $X_i^\star$ by using the learned parameters: $\{\bar{\alpha}_i, \bar{\beta}_i\}$. This can be obtained by minimizing the negative log-posterior function $-lnp(X^T, \bar{\alpha}_i, \bar{\beta}_i, \Omega|Z^\star)$ with scaled conjugate gradient (SCG) on the training datasets. After that, we can calculate the conditional probability $P\left(Z_i^{(*)}, X_i^{(*)}|\Lambda_i\right)$.

2. *Select the GPDM with the highest conditional probability:* The new group activity can be determined by the following equation:

$$argmax_{i=1,\ldots,n} P\left(Z_i^{(*)}, X_i^{(*)}|\Lambda_i\right) \quad (1)$$

As we discussed in the previous section, the length of new observation can be different with the size of training data, which means that the number of frames in test clips can be different with training clips. Therefore our trained model can address the dynamics in the temporal dimension. As the duration of an activity may change under different situation, it is important that the classifier can handle the testing sequences with varying lengths.

4. EXPERIMENTAL RESULTS

We test our framework on two popular group activity data sets. The first one is the recently released BEHAVE data set [10], which contains the ground truth for each group activity. The second data set is IDIAP data set [11] , which was originally captured for multiple human tracking.

4.1. Results on BEHAVE data set

The BEHAVE data set consists of four video clips, and $76,800$ frames in total. This video data set is recorded at 26 frames per second and has a resolution of 640×480. Different activities include: InGroup, Approach, WalkTogether, Split, Ignore, Following, Chase , Fight, RunTogether, and Meet. There are 174 samples of different group activities in this dataset.

As our focus is the small group activity analysis, we select 118 samples from all the group activities data set, and all the samples contain three or more people in the scene. The selected group activities include InGroup (IG), WalkingTogether(WT), Split(S) and Fight(F) as our group activities. For each activity, we divide the samples to ten-fold, with nine-fold for training and one fold for testing, the classification result is shown in the Table 1. Two of learned GPDMs are shown in the Figure 1. Each point in the latent space indicate one of the

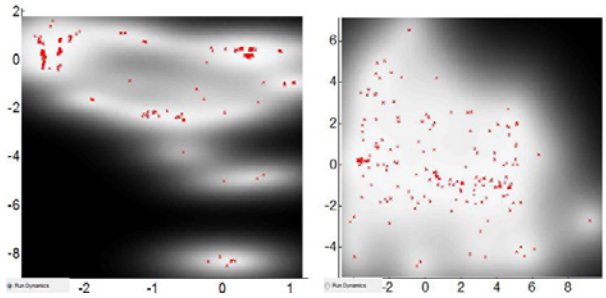

Fig. 1. Visualization of trained GPDMs, the left one the In-Group, and the right one is Group Fight

Table 1. Classification results of our method

	IG	WT	F	S
Our method	94.3%	92.1%	95.1%	93.1%

feature vector in a frame. The distribution of InGroup activity is prone to have some local clusters in the latent space, while the distribution of GroupFight activity is similar to a random distribution.

We also compare our results with the classification results in [10]. As in [10], the training and testing data is divided to $50/50$, our proposed method can achieve 93.1%, comparing to 92.1% of HMM based method[10]. It should be noted that, the recognition rate is the average rate for all the activities, and the window size for calculating feature in [10] is 60. In addition, our proposed algorithm can adaptively recognize human group action with different length, although the method in [10] can reach a higher recognition rate when window size is increased to 100.

4.2. Results on IDIAP data set

IDIAP data set is firstly used in [11] for multiple targets tracking. The data set contains 37182 frames in total. We manually select 46 clips with different lengths for human group activity recognition. As there is no Fight activity in the IDIAP data set, we just evaluate three group activities: InGroup, WalkTogether,and Split. To validate the robustness of our framework, we directly apply the trained GPDMs in the BEHAVE data set for activity recognition on the IDIAP data set, and the overall average classification rate is 92.3%. The expremenat results indicate that our proposed framework is robust to identify human group activities under different scenariors.

5. CONCLUSION

In this paper, we propose a novel structural feature to describe the small group activity. Based on the structural feature, we also propose a Conditional Gaussian Process Dynamic Model for group activity recognition. The proposed structural feature can be adapted to many other applications, since its dynamic characteristics can be used to describe different features. The framework can also be used for abnormal group activity detection in surveillance systems.

6. ACKNOWLEDGMENTS

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under Grant FA8650-11-1-7152.

7. REFERENCES

[1] J. C. Niebles, H.Wang, and F. Li, "Unsupervised learning of human action categories using spatial-temporal words," in *BMVC 2006*.

[2] S.-F. Wong and C. Roberto, "Extracting spatiotemporal interest points using global information," in *ICCV 2007*.

[3] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR 2008*.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *ICCV 2005.*, pp. 1395–1402.

[5] S. Christian, L. Ivan, and C. Barbara, "Recognizing human actions: a local svm approach," in *ICPR 2004*.

[6] B. Ni, S. Yan, and A. Kassim, "Recognizing human group activities with localized causalities," in *CVPR 2009*.

[7] W. Ge, R. T. Collins, and R. Barry, "Automatically detecting the small group structure of a crowd," in *WACV 2009*.

[8] M. Chang, N. Krahnstoever, S. Lim, and T. Yu, "Group level activity recognition in crowded environments across multiple cameras," in *AVSS2010*.

[9] R. T. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Trans. on PAMI*, vol. 27, no. 10, pp. 1631 –1643, 2005.

[10] S. Blunsden and R. B. Fisher, "The behave video dataset: ground truthed video for multi-person," *Annals of the BMVA*, vol. 4, pp. 1–12, 2010.

[11] K. Smith, D. Gatica-perez, and J. Odobez, "Using particles to track varying numbers of interacting people," in *CVPR 2005*.

[12] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," 2007.

[13] I. Laptev and T. Lindeberg, "Space-time interest points," in *IEEE 6th International Conference on Computer Vision, ICCV 2003*.

[14] D. J. Watts and D. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, June 1998.

[15] S. Wasserman and K. Faust, *Social Networks Analysis: Methods and Applications*. Cambridge University Press., 1994.

[16] V. Krebs, "The social life of routers," in *Internet Protocol Journal*, dec. 2000, pp. 14–25.

[17] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamic models for human motion," *IEEE Trans. on PAMI*, vol. 30, pp. 283–298, 2008.

[18] N. Lawrence, "Probabilistic non-linear principal component analysis with gaussian process latent variable models," *Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.