Group Context Learning for Event Recognition *

Yimeng Zhang† Weina Ge‡ Ming-Ching Chang‡ Xiaoming Liu‡
†School of Electrical and Computer Engineering, Cornell University
‡GE Global Research Center, 1 Research Circle, Niskayuna, NY
†yz457@cornell.edu ‡{gewe,changm,liux}@research.ge.com

Abstract

We address the problem of group-level event recognition from videos. The events of interest are defined based on the motion and interaction of members in a group over time. Example events include group formation, dispersion, following, chasing, flanking, and fighting. To recognize these complex group events, we propose a novel approach that learns the group-level scenario context from automatically extracted individual trajectories. We first perform a group structure analysis to produce a weighted graph that represents the probabilistic group membership of the individuals. We then extract features from this graph to capture the motion and action contexts among the groups. The features are represented using the "bag-of-words" scheme. Finally, our method uses the learned Support Vector Machine (SVM) to classify a video segment into the six event categories. Our implementation builds upon a mature multi-camera multi-target tracking system that recognizes the group-level events involving up to 20 individuals in real-time.

1. Introduction

Recognizing events of interest from surveillance videos is an important topic and has been extensively studied. Applications include monitoring transportation hubs, public venues, and yards for security and safety. In general, the efforts can be organized into three main categories: *(i) action recognition* [2, 12, 18, 20], such as recognizing if a person is walking or chatting, where the analysis of the body articulation is essential; *(ii) interaction recognition* [9, 14, 22] between a few individuals or with respect to an object, such as determining if two people are meeting or exchanging items, where both the overall motion and articulation are useful cues; and *(iii) crowd event recognition* [11, 13, 16, 19, 21, 26, 28, 31], such as detecting abnormal traffic or aggressive fight involving groups of individuals, where the scenario is most complicated, since a diverse range of activities could occur in a crowded scene.

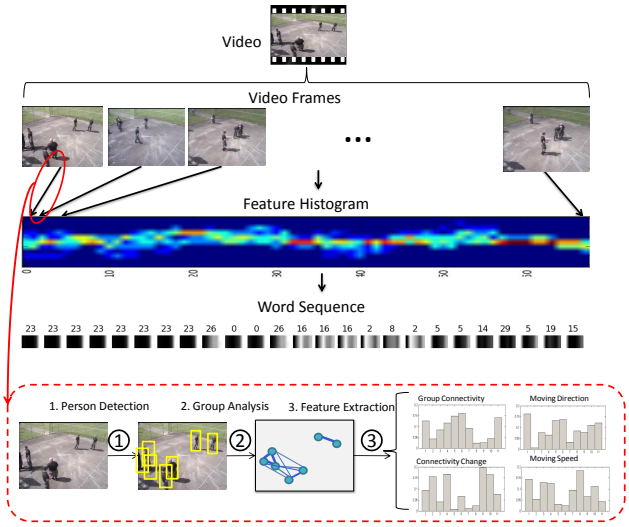


Figure 1. **Overview: Learning group context words.** Given a video we extract feature histograms for each frame to represent their group connectivity and motion features. To extract features from a frame, we go through three steps as illustrated at the bottom of the figure. The feature histogram image shown is column-wise stacked, where the red indicates large value and blue indicates small value. Group context words learned from the feature histograms are visualized in the middle, where each word is created with the histogram of four consecutive frames, and each row in a word depicts a histogram for a frame.

In this paper, we are interested in recognizing events involving groups of people and the interaction among them. Example scenarios including group dynamic analysis such as group formation, dispersion, and one or more groups ap-

*This work was done while the first author was visiting GE Global Research as an intern. This work was supported by the National Institute of Justice, US Department of Justice, under the award #2009-SQ-B9-K013. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

(a) group dispersion (b) group formation (c) approaching (d) group flanking (e) group fighting

Figure 2. Example scenarios we aim to recognize from videos. Typical behaviors of interest include group formation and dispersion, a group approaching, following or chasing another group, one or more groups aggressively surrounding another group (flanking), which likely suggests a fight, and the actual group fighting.

proaching another group. We are also interested in detecting group-level behaviors such as chasing and aggression among groups, which are likely related to potential fighting and security concerns. Our task is different from those crowd event recognition tasks that are mainly based on motion analysis of dense crowds [1,13,21,28], and focus on the macro-analysis of the crowd rather than the micro-dynamics of groups and the interaction of individuals. Figure 2 illustrates typical scenarios that we aim to recognize from surveillance videos.

We propose a novel learning-based framework for group event recognition that automatically learns the *group context* for different event categories. The group context refers to the group-level interactions among people over time. An overview of our approach is illustrated in Figure 1. We first detect and track people in the video using standard methods [4, 30]. Based on the tracking, we analyze the group-level structure and motion, and then extract the group context features based on the probabilistic (soft) group structure analysis results (Figure 3). We can then recognize group-level behaviors and detect events of interest using the learned group context features.

A key step towards robust event detection is the ability to recognize the *temporal co-occurrence* of similar group context patterns that appear in videos, which can occur at different locations, scales, and times. To illustrate the challenge, a video containing group formation could possibly involve a variable number of people getting together in a variable length of time. Thus, the occurrence and co-occurrence of different group context patterns are the key to recognize this particular scenario. In order to achieve robust event recognition, we first define several robust features that model the interaction and motion between individuals among the groups (group context), which can be detected on a per-frame or per-segment basis (Figure 1). Second, to capture the temporal co-occurrences of these features exacted from different frames in a video, we adopt the "bag-of-words" scheme [7] by clustering them into group context words. Finally we train a SVM with bags of group context words to classify the videos into different event categories.

We summarize our contributions as follows: (1) we develop a novel machine learning based framework to robustly recognize group-level events from videos; (2) we propose robust features that model the group context of individuals with motion tracking; and (3) we implement our algorithm with a multi-camera tracking system and demonstrate it in a real-time event recognition system in surveillance applications.

2. Related Works

With the prevalence of surveillance cameras, event recognition has drawn increasing attention from the computer vision community. Some works consider the problem of recognizing actions performed by a single person, such as [2, 12, 18, 20, 23, 27, 29]. Activity categories such as walking or running are defined and detected straight from analyzing body part movements of a person.

More relevant to our work is the recognition of events involving multiple agents in a crowd scene. Existing works typically focus on events defined by the movements of individuals or the entire crowd. Typical applications include the detection of cars or people with abnormal movements in a traffic scene. There are mainly two types of approaches in this category. The first type is *object centric* [1, 24], where the trajectories of detected targets are analyzed for recognition. The second type is *view or flow centric* [11, 13, 16, 19, 21, 25, 26, 28, 31], which avoids object tracking, and instead models the crowd motions with dense optical flows, or the gradients and appearances of the spatio-temporal subvolumes. The non-object centric approach is also popular for *general event or action categorization* from movies or youtube videos [8, 15, 17]. Motion or appearance features are extracted from spatio-temporal subvolumes detected at interest points. Usually a visual word representation is followed for the final event categorization.

We focus on the type of events that are defined not only by the motion information but also by the interactions of groups or the individuals among groups, such as "group fighting" and "flanking" (a group of people surrounding another group). Previous works on this type of events mainly use the logic or rule based methods, which require manual creation of rules [3, 4, 10, 14, 22] for each event category. Chang *et al.* [4] recognize various group events by combining the results from probabilistic group structure analysis and motion analysis and checking against a list of event models, which are defined manually using scenario-specific

predicates. To explicitly model the temporal constraints pertaining to complex events, different probabilistic logical inference engines have been built, such as the Markov Logic Networks [22] and probabilistic event logic [3]. These works use rule or logic based approach, and thus require experts to manually create person-person or person-object rules with domain knowledge. Therefore, the performance of event recognition highly depends on how well the rules are defined. The learning curve for a general operator to define a set of compatible rules could be sharp. Moreover, the rules in these methods often rely on clean input observations, which are hardly the case for results obtained from automatic detection and tracking algorithms.

The learning-based approach of Choi *et al.* [5, 6] is relevant to us. They focus more on atomic actions such as people queuing and talking, thus only human pose and spatial distance cues are considered for event recognition. In comparison, our events of interest involve group context and require further analysis on the group-level motion and interaction cues that could possibly change over time.

3. Approach

We propose to extract robust group context features from video and adopt a bag-of-words learning scheme to recognize group-level events. Figure 1 illustrates our overall approach. Given an input video segment, we first perform person detection and tracking. We then perform group structure analysis of the tracked individuals, as a mean to extract group context features. Following [4], we retain a probabilistic group representation, such that the group-level information can be reliably captured. Specifically, the group analysis produces a weighted connectivity graph for each frame, where the nodes of the graph are the detected individuals and the weight of an edges is the probability of two individuals being in the same group. From the connectivity graph, we extract features that capture pair-wise group relationships among the individuals and their motion information. Finally we cluster the extracted features into group context words and use "bag of group context words" to train a SVM to classify the input video segment into an event category. In the following sections, we will explain the details of each step in our approach.

3.1. Video Tracking System

We briefly explain the multi-view, multi-target tracking system that is used as a base component. Note that the event recognition algorithm introduced in this paper is general and can be applied to tracking results from other systems.

We take the videos from three standard CCTV cameras of overlapping views. All cameras are calibrated and synchronized. Figure 3 gives a snapshot of our system in operation, where the movements of the individuals are tracked cooperatively across cameras. Person detections from each view are projected onto the ground plane in 3D and fed into a centralized tracker which is implemented with a Kalman filter.

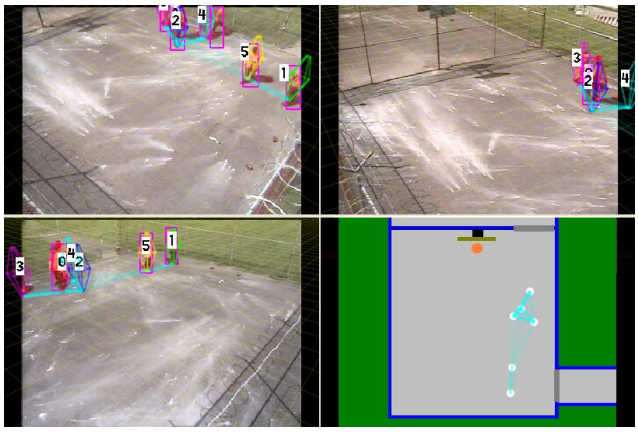


Figure 3. Video tracking system performing person detection and tracking from one or more views. A top-down synthesized view is generated to visualize the probabilistic grouping connectivity w_{ij} between individuals under tracking.

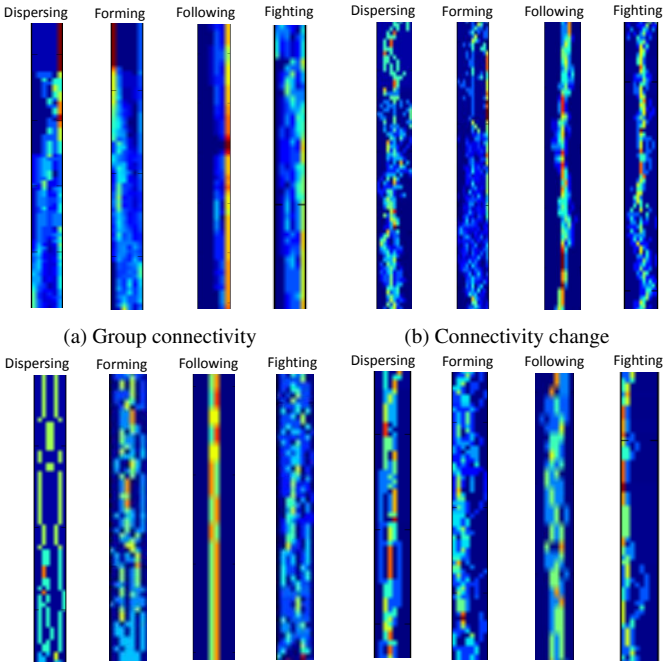
3.2. Group Analysis

Given the tracking of individuals, we incorporate a probabilistic grouping strategy similar to [4] to perform group analysis and to extract group-level motion and interaction features. As opposed to approaches that rely on hard, agglomerative or divisive clustering techniques to define groups, the probabilistic grouping without a hard segmentation of groups keeps more reliable information about the dynamics of groups that will be later on used to calculate the group context features.

For each frame t, we define a connectivity graph $\mathcal{G}^t$ to represent the connectivity (or the probability) of two individuals i and j belonging to a group at frame time t. Specifically, for each edge e_{ij}^t in $\mathcal{G}^t$, the edge weight w_{ij}^t represents the probability that individuals i and j belong to the same group, $0 \leq w_{ij}^t \leq 1$. The definition of the connectivity w_{ij}^t is motivated by two lines of thoughts: (1) a *track-to-track* connectivity that considers the motion of the two individuals i and j under tracking, including the spatial distance and moving direction calculated from a small period of time in the tracking history, and (2) a *path-based* connectivity that considers the existence of neighboring individuals that increase the overall grouping strength of nearby individuals all together. The bottom-right image in Figure 3 illustrates an example of the probabilistic grouping graph $\mathcal{G}^t$ from a synthesized top-down view of the tracked individuals.

3.3. Feature Extraction

In order to achieve the recognition of group-level events that could occur in variable time scales, we propose to use robust features that can be efficiently extracted and can cap-



(a) Group connectivity (b) Connectivity change

(c) Moving direction (d) Moving speed

Figure 4. Feature histograms for example videos from four event categories. Red denotes higher value, and blue denotes lower value. See text for explanation of how the histograms capture group context features.

ture information about the group structure, motion, and dynamics. Our solution is to extract the following four types of features: (1) group connectivity, (2) connectivity change, (3) motion direction, and (4) motion speed (Figure 4).

Group connectivity. This feature models the group structures in a frame t by creating a histogram of the edge weights w_{ij}^t from the group connectivity graph $\mathcal{G}^t$. The histogram is normalized by the total number of edges so that the number of people does not bias the measurement. Figure 4(a) shows the group connectivity histograms of example videos from four event categories (group dispersion, formation, following, and fighting), where each row depicts a histogram at a frame, and the column axis indicates time. Observe that in the beginning of a group dispersing event, the bins corresponding to high connectivity values have more counts, and as the event unfolds in time, the bins of low connectivity values receive more counts. In other words, the strength of group connectivities decreases over time. In the contrary, for the group forming event, the strength of group connectivity increases over time. For the group following and fighting events, people mostly have high connectivity values, since people maintain tight groups in these events. These observations verify that this novel feature captures discriminative cues to distinguish various group-level events.

Connectivity change. This feature models the group connectivity change between the current frame t and a previous frame t' ($t' = t - 1/S$ second), for each pair of individuals i and j who are detected in both frames. The connectivity difference for i and j is $\Delta_{ij}^t = w_{ij}^t - w_{ij}^{t'}$, where w_{ij}^t is the weight for the edge between individual i and j in $\mathcal{G}^t$. The connectivity change feature is the histogram of such differences of all person pairs in the current frame, and again the histogram is normalized by the total number of edges. Figure 4(b) shows the connectivity change histograms for four event categories. The center of a histogram represents no connectivity change ($\Delta_{ij}^t \approx 0$); bins to the left correspond to negative changes ($w_{ij}^t < w_{ij}^{t'}$), that is, the group connectivity of the current frame is smaller than the one in the previous frame; bins to the right correspond to positive changes. The figure shows that the connectivity changes are mostly negative for the group dispersing event and positive for the group forming event. The connectivity change is almost 0 for the group fighting and following events since the group structures are reasonably stable during these events.

Motion direction. We also record the moving direction of each person i by the velocity direction $d_i^t \in [0, 2\pi)$. To deal with camera rotations and view point changes, we normalize the direction d_i^t of each person by subtracting the mean of them d_μ^t in the $[0, 2\pi)$ periodic space: $\hat{d}_i^t = d_i^t - d_\mu^t$. This normalized motion direction $\hat{d}_i^t$ is used to compute the motion direction histogram. Figure 4(c) shows example motion direction histograms for different event categories. In a group following event, people tend to have similar motion directions, since that direction is the one all are heading towards. For other events, the moving directions have a wider distribution.

Motion speed. This feature captures the motion speed s_i^t (magnitude of velocity) for each person i. Observe in Figure 4(d) that people do not move much when they are engaged in a fight; while for the dispersion and following events, people show larger motion speeds over time.

All four features can be extracted directly and efficiently from trajectories obtained from the tracker. These features robustly capture group structure and dynamic changes over time. We will next describe how we formulate our bag-of-words learning scheme based on these features.

3.4. Learning Group Context Words

After feature extraction, a video can be represented as a sequence of feature histograms (Figure 1). Direct learning of classifiers on these sequences is difficult, especially when we have a long video. Moreover, the video of an event can have variable lengths, and the starting and ending time of the event are unknown, rendering the problem more difficult. We propose to cluster the feature histograms into a few representative clusters, which we refer to as *group context words*.

To create such words, we first represent each frame by

concatenating the histograms of the current and previous consecutive T frames. This concatenated histogram models local histogram changes and smoothes out the noise in the observations from a single frame. Then we cluster the concatenated histograms using K-means into a vocabulary of $|V|$ words. A word represents a certain pattern of the local histograms. Since we have four types of features, we create a vocabulary for each feature type. Thus for each feature type, a video will be represented as a sequence of words.

We adopt the "bag-of-words" model, which represents a video as a histogram of words. We create a bag-of-words histogram for each feature type and concatenate them together. These concatenated word histograms are used to train a SVM to classify the input video into different event categories.

4. Experiments

We evaluate our approach on part of the Mock Prison Riot (MPR) dataset (http://mockprisonriot.org) as in [4]. The dataset has 19 surveillance videos taken in an abandoned prison yard in West Virginia. In these videos, several volunteer correction officers enact typical behaviors of the prisoners. The length of each video varies from 3 to 6 minutes. Example snapshots of the dataset can be found in Figure 2. We report our performance for the following six categories of group-level events: (1) group formation, (2) group dispersion, (3) group following, (4) group chasing, (5) group flanking, and (6) group fighting.

4.1. Event Recognition

The first experiment we performed is to classify an entire input video into one of the six pre-defined event categories, *i.e.*, to determine whether an event occurs or not. For this experiment, we manually segmented the videos in the dataset into 177 non-overlapping small video segments of 2 to 30 seconds. For each video segment, we label all events occurred in the segment. Some events may overlap with others and occur at the same time. If no events of interest occurred in a segment, we label it as "random", which serves as negative examples for all other categories. Note that we do not need to label a clear start and end point for each event in the video. We randomly select 60% of the videos for training and the rest for testing.

We illustrate the words of the "connectivity change" feature that occur most frequently in the videos of the four event categories (dispersing, forming, following, and fighting) in Figure 5. As described in Section 3.4, a word is constructed with the histograms of consecutive T frames (we use $T = 4$). In the figure, each row of a word image represents a "connectivity change" feature histogram exacted from a frame. Notice these feature histograms show similar observations as those visualized in Figure 4(b). The most frequent words for the group dispersing event are those that

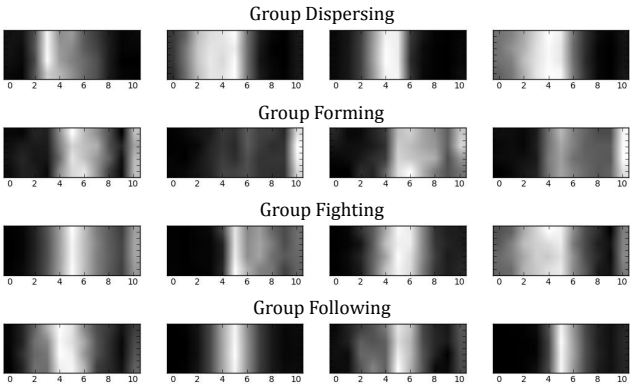


Figure 5. The "connectivity change" words that occur most frequently in the videos of different event categories. Each word is created with the histograms from four consecutive frames. For each word image, one row corresponds to a histogram exacted from one frame.

represent the frames where the group connectivities among people decrease as compared to the previous frames. On the contrary, the top words for the group forming event are those that show the opposite pattern. For the group fighting and group following events, the top words correspond to the histograms where the group connectivity change is close to zero.

We train an one-vs-all SVM for each event category, and evaluate the recognition performance with the ROC curves drawn with the probabilistic scores generated by the SVMs. Figure 6 shows the ROC curve for three example event categories using different feature types. The "combined" one uses all four feature types concatenated as described in Section 3.4. As shown in the figure, for the group forming event, the group connectivity and connectivity change features are more useful, compared to the motion direction and speed features. While for the group fighting event, the group connectivity feature outperforms other feature types. The combined one achieves the best performance. We also show the ROC curve for the random group, which indicates the performance for distinguishing behaviors of interest from normal behaviors. We show the AUC (area under curve) scores for all categories in Figure 7. Similar to the results for the forming event, "group connectivity" and "connectivity change" features are more discriminative for recognizing group dispersing. The speed feature performs better for group chasing events, since people move fast in the chasing events, which is a very distinct feature for this particular event category as compared to the other five. The speed feature also works well for the group following category, since people keep relatively constant speed over time when following each other. The performance with the combined features is the best for all event categories except for the fighting category, where the combined feature performs worse but still comparable to the connectivity feature. We achieve more than 90% AUC scores for all categories.

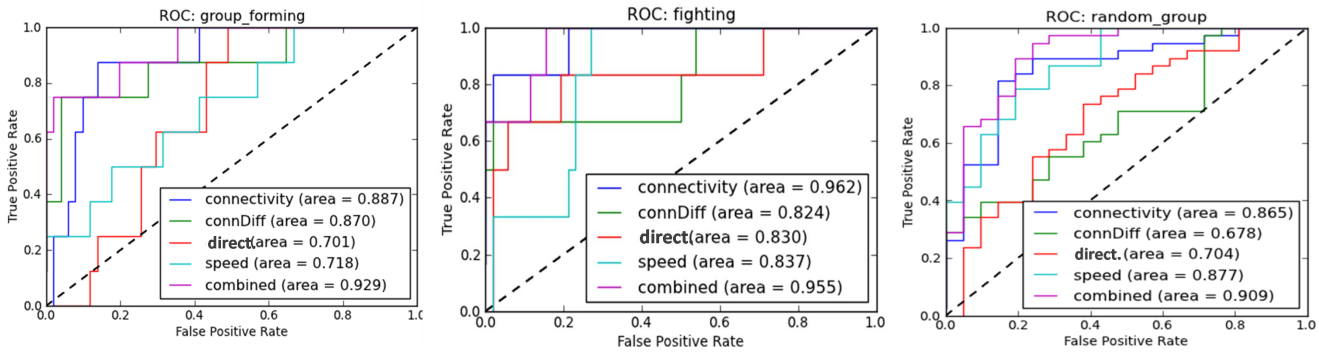


Figure 6. The ROC curves using different types of features for event categories: (left) group forming, (middle) group fighting, and (right) random group. The "random group" ROC curve shows the performance for classifying the events of interest vs. normal behaviors.

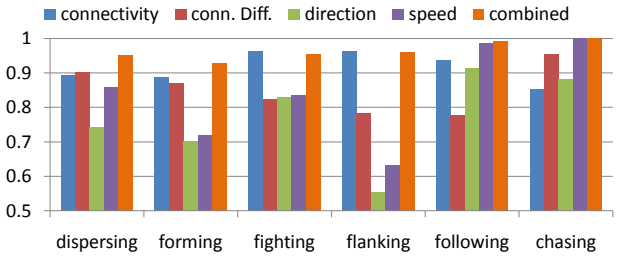


Figure 7. The AUC scores of the ROC curves for different event categories using different types of features.

4.2. Event Detection

In the second experiment we perform online event detection, that is, to determine whether any event of interest occurs at each frame in the input video. We label the start and end points of the occurred events for the videos in our dataset. This scenario is useful to provide real-time alerts to the operators. Since a clear start and end points are difficult to determine for several events, we label the one second period around the start and end points of an event as ambiguous frames, and do not use them for evaluation. We randomly select 60% of the 19 videos in the dataset for training, and the rest for testing. We make prediction at every 4^{th} frame in a video, using observations from a four-second temporal window ($[t-4s, t]$), *i.e.*, the previous 4 seconds. Other aspects of the algorithm remain the same.

We compare the performance of our approach against the state-of-the-art approach introduced in [4], which adopts a rule-based method for event detection. Probabilistic rules are created manually for each event category and probabilistic decisions are made at each frame. Figure 8 shows the prediction results on an example test video. The rule-based method generates much more false positives than our method. One main reason is that the rule-based method is more sensitive to the person detection errors, since the errors are not considered when creating the rules, whereas our method can tolerate more observation noise by learning from the training data. The other reason is that the rules

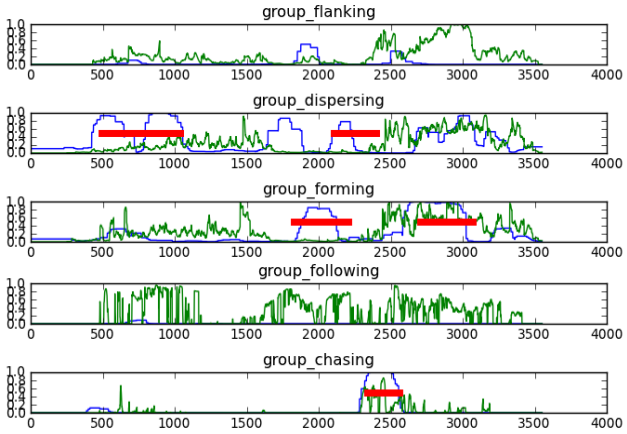


Figure 8. The predicted probabilities for a test video. Red lines represent the ground truth labels. Green lines denote the probabilities at each frame using the rule-based method [4]. Blue lines denote the probabilities using our approach.

	Dispersing	Forming	Flanking	Following	Chasing	Fighting
rule-based	0.592	0.658	0.921	0.667	0.981	N/A
ours	**0.926**	**0.811**	**0.959**	**0.827**	**1.000**	**0.834**

Table 1. The AUC scores of the ROC curves using the rule-based method [4] and our method.

in [4] only consider the past several frames when making the decision for the current frame, while our method uses a much larger temporal window (4 seconds) and is thus able to remove some local noise. Table 1 shows the AUC score comparison for different event categories. Our method outperforms the rule-based method [4] on all categories. Note that the "fighting" event is not defined in [4], since its occurrence in a video usually spans a long duration. As we already discussed, the rules in [4] are usually defined with only a few frames.

We implement our approach with C++ and Python. On an Intel 2.4G dual-core computer, the entire event detection system, including person detection and tracking, takes around 0.02 second to process a 640x480 frame. Therefore,

we can detect events of interest in real-time.

5. Conclusion

We proposed a novel learning based framework for group-level event recognition. Unlike most existing event recognition works, which define the events based on the movements of an individual or the entire crowd, the events discussed in this paper focus more on the interactions among people. We designed robust features that can capture the group context of individuals in a video. We built a system with the proposed algorithm, which can process a video and detect the events in real-time. The performance of the system significantly outperforms the state-of-the-art method on a challenging dataset.

Future work. We would like to explore a broader range of event categories for various scenarios of interest in surveillance and forensic applications. We are also interested in developing algorithms that can wisely combine different types of features, rather than a straightforward concatenation. Finally, instead of the "bag-of-word" scheme, we plan to develop algorithms that can further model the occurrence and temporal relationship between feature words, in order to improve the scenario recognition performance.

References

[1] A. Basharat, A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. In *CVPR*, 2008. 250

[2] W. Brendel and S. Todorovic. Learning spatio temporal graphs of human activities. In *ICCV*, 2011. 249, 250

[3] W. Brendel, S. Todorovic, and A. Fern. Probabilistic event logic for interval-based and holistic event recognition. In *CVPR*, 2011. 250, 251

[4] M.-C. Chang, N. Krahnstoever, and W. Ge. Probabilistic group-level motion analysis and scenario recognition. In *ICCV*, 2011. 250, 251, 253, 254

[5] W. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Visual Surveillance Workshop, ICCV*, 2009. 251

[6] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *CVPR*, 2011. 251

[7] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004. 250

[8] L. Duan, D. Xu, I. W. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *CVPR*, 2010. 250

[9] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury. A string of feature graphs model for recognition of complex activities in natural videos. In *ICCV*, 2007. 249

[10] A. Hakeem and M. Shah. Learning, detection and representation of multi-agent events in videos. *Artifical Intelligence*, 171:586–605, June 2007. 250

[11] T. Hospedales, S. Gong, and T. Xiang. A Markov clustering topic model for mining behaviour in video. In *ICCV*, 2009. 249, 250

[12] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, 2007. 249, 250

[13] D. Kuettel, M. D. Breitenstein, L. J. V. Gool, and V. Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *CVPR*, 2010. 249, 250

[14] S. Kwak, B. Han, and J. H. Han. Scenario-based video event recognition by constraint flow. In *CVPR*, 2011. 249, 250

[15] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, I. Rennes, I. I. Grenoble, and L. Ljk. Learning realistic human actions from movies. In *CVPR*, 2008. 250

[16] J. Li, S. Gong, and T. Xiang. Scene segmentation for behavior correlation. In *ECCV*, 2008. 249, 250

[17] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *CVPR*, 2009. 250

[18] J. Liu, M. Shah, B. Kuipers, and S. Savarese. Cross-view action recognition via view knowledge transfer. In *CVPR*, 2011. 249, 250

[19] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, 2010. 249, 250

[20] S. Maji, L. Bourdev, and J. Malik. Action Recognition from a Distributed Representation of Pose and Appearance. In *CVPR*, 2011. 249, 250

[21] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009. 249, 250

[22] V. I. Morariu and L. S. Davis. Multi-agent event recognition in structured scenarios. In *CVPR*, 2011. 249, 250, 251

[23] F. Nater, H. Grabner, and L. V. Gool. Exploiting simple hierarchies for unsupervised human behavior analysis. In *CVPR*, 2010. 250

[24] I. Saleemi, K. Shafique, and M. Shah. Probabilistic modeling of scene dynamics for applications in visual surveillance. *PAMI*, 31:1472–1485, August 2009. 250

[25] S. Saxena, F. Brémond, M. Thonnat, and R. Ma. Crowd behavior recognition for video surveillance. In *ACIVS*, 2008. 250

[26] D. Tran and J. Yuan. Optimal spatio-temporal path discovery for video event detection. In *CVPR*, 2011. 249, 250

[27] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. In *CVPR*, 2011. 250

[28] S. Wu, B. E. Moore, and M. Shah. Chaotic invariants of lagrangian particle trjectories for anomaly detection in crowded scenes. In *CVPR*, 2010. 249, 250

[29] A. Yao, J. Gall, and L. V. Gool. A hough transform-based voting framework for action recognition. In *CVPR*, 2010. 250

[30] T. Yu, Y. Wu, N. Krahnstoever, and P. Tu. Distributed data association and filtering for multiple target tracking. In *CVPR*, 2008. 250

[31] G. Zen and E. Ricci. Earth Movers's Prototypes: a Convex Learning Approach for Discovering Activity Patterns in Dynamic Scenes. In *CVPR*, 2011. 249, 250