A Group Contextual Model for Activity Recognition in Crowded Scenes

Khai N. Tran, Xu Yan, Ioannis A. Kakadiaris and Shishir K. Shah
University of Houston, Department of Computer Science, 4800 Calhoun Rd, TX 77204, U.S.A.

Keywords: Group Context Activity, Activity Recognition, Social Interaction.

Abstract: This paper presents an efficient framework for activity recognition based on analyzing group context in crowded scenes. We use graph based clustering algorithm to discover interacting groups using top-down mechanism. Using discovered interacting groups, we propose a new group context activity descriptor capturing not only the focal person's activity but also behaviors of its neighbors. For a high-level of understanding of human activities, we propose a random field model to encode activity relationships between people in the scene. We evaluate our approach on two public benchmark datasets. The results of both the steps show that our method achieves recognition rates comparable to state-of-the-art methods for activity recognition in crowded scenes.

1 INTRODUCTION

Typically, in crowded scenes, people are engaged in multiple activities resulting from inter and intra-group interactions. This poses a rather challenging problem in activity recognition due to variations in the number of people involved, and more specifically the different human actions and social interactions exhibited within people and groups (Ryoo and Aggarwal, 2011; Tran, 2013). Understanding groups and their activities is not limited to only analyzing movements of individuals in group. The environment in which these groups exist provides important contextual information that can be invaluable in recognizing activities in crowded scenes. Perspectives from sociology, psychology and computer vision suggest that human activities can be understood by investigating a subject in the context of social signaling constraints (Smith et al., 2008; Helbing and Molnár, 1995; Cristani et al., 2011). Exploring the spatial and directional relationships between people can facilitate the detection of social interactions in a group. Thus, activity analysis in crowded scenes can often be considered a multi-step process, one that involves individual person activity, individuals forming meaningful groups, interaction between individuals and interactions between groups. In general, the approaches to group activity analysis can be classified into two categories: bottom-up and top-down. The bottom-up approaches rely on recognizing activity of each individual in a group. Vice versa, top-down approaches recognize group activity by analyzing at the group level rather than at the individual level. Bottom-up approaches show the understanding of activities at the individual level, however they are limited in recognizing activities at group level. Top-down approaches show better contextual understanding of activities in groups but they are not robust enough to recognize activities at the individual level.

In this paper, we develop a social context framework for recognizing human activities in crowded scenes by taking advantage of both top-down and bottom-up approaches. Our hybrid framework localizes groups through social interaction analysis using a top-down approach and analyzes individual activity based on social context within the group using a bottom-up mechanism. We propose a novel group context activity descriptor capturing characteristics of individual activity with respect to the behavior of its neighbors along with an efficient conditional random field model to learn and classify human activities in crowded scenes.

The main contributions of our work are:

1. *A Group Context Activity Descriptor.* We use a top-down approach to dynamically localize interacting groups capturing behaviors of individuals. We form a group context activity descriptor that is a combination of individual activity and its neighbor's behavior, represented using the Bag-of-Words (BoW) representation.
2. *An Efficient Conditional Random Field Framework to Learn and Classify Human Activities in*

Tran K., Yan X., Kakadiaris I. and Shah S..
A Group Contextual Model for Activity Recognition in Crowded Scenes.
DOI: 10.5220/0005258600050012
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 5-12
ISBN: 978-989-758-090-1

context. We present a recognition framework that jointly captures the individual activity and its activity relationships with its neighbors.

The rest of the paper is organized as follows. We review related work on activity analysis in crowded scenes in section 2. Section 3 describes the human activity descriptor in group context along with the conditional random field model used to address the activity recognition task. Experimental results and evaluations are presented in Section 4. Finally, Section 5 concludes the paper.

2 RELATED WORK

In this section, we review related work on human activity analysis in crowded scenes that use a top-down or bottom-up approach. In bottom-up approaches, group context is used to differentiate ambiguous activities e.g. standing and talking, which are normally represented by the same local descriptors. Most approaches integrate contextual information by proposing a new feature descriptor extracted from an individual and its surrounding area. Lan *et al.* (Lan et al., 2012b) propose an Action Context (AC) descriptor capturing the activity of the focal person and the behavior of other people nearby. AC descriptor is computed by concatenating the focal person's action probability vector (computed using Bag-of-Words approach with SVM classifier), and the context action probability vectors capturing the activities of other neighborhood people. However, this AC descriptor only can capture spatial proximity information by using 'near by' context. Considering a more sophisticated contextual descriptor, Choi *et al.* (Choi et al., 2009) propose Spatio-Temporal Volume (STV) descriptor, which captures spatial distribution of pose and motion of individuals in a scene to analyze group activity. STV descriptor centered on a person of interest or an anchor is used for classification of the group activity. The descriptor is a histogram of people and their poses in different spatial bins around the anchor. These histograms are concatenated over the video to capture the temporal nature of the activities. SVM using pyramid kernels is used for classification. The same descriptor is leveraged in (Choi et al., 2011) but Random Forest classification is used for group activity analysis. In addition, random forest structure is used to randomly sample the spatio-temporal regions to pick most discriminative features. Recently, Amer *et al.* (Amer and Todorovic, 2011) introduced Bags-of-Right-Detections (BORD) seeking to remove noisy people detection in groups. BORD is a histogram of human poses detected in a spatio-temporal neighborhood centered at a point in the video volume. The BORD is not computed from all neighborhood people, but only from those detections that are considered to take part in the target activity. A two-tier MAP inference algorithm is proposed for the final recognition step.

In contrast to bottom-up approaches, top-down methods model the entire group as a whole rather than each individual separately. Khan and Shah (Khan and Shah, 2005) use rigidity formulation to represent parade activities. They modeled group shape as a 3D polygon with each corner representing a participating person. The tracks from person in group are treated as tracks of feature points in a 3D polygon. Using rank of track matrix, activities are classified as parade or just random crowds. Vaswani *et al.* (Vaswani et al., 2003) model an activity using a polygon and its deformation over time. Each person in the group is treated as a point on the polygon. The model is applied to abnormality detection in a crowded scene. Multi-camera multi-target tracks are used to generate dissimilarity measure between people, which in turn are used to cluster them into groups in (Chang et al., 2010). Group activities are recognized by treating the group as an entity and analyzing the behavior of the group over time. Mehran *et al.* (Mehran et al., 2009) built a 'Bag-of-Forces' model of the movements of people using social force model in a video frame to detect abnormal crowd behavior. Close to top-down approach, Ryoo *et al.* (Ryoo and Aggarwal, 2011) present an approach that splits group activity into subevents like person activity and person to person interactions. Each portion is represented using context free grammar and the probability of their occurrence given a group activity or time periods. A hierarchical recognition algorithm based on Markov Chain Monte Carlo density sampling technique is developed. The technique identifies the groups and group activity simultaneously.

Recently, several approaches that leverage social signaling cues for analyzing crowded scenes have been proposed. Group activities can be better inferred from valuable social interactions cues between people present in the scene. Several approaches are proposed to identify meaningful group from the videos using spatial and orientational arrangement of people in the scene as a cue based on social signaling principles (Farenzena et al., 2009b; Farenzena et al., 2009a; Tran et al., 2014). Lan *et al.* (Lan et al., 2012a) present a bottom-up approach integrating social role analysis to understand activities in crowd scene. Different from above approaches, our approach takes advantage of both bottom-up and top-down mechanisms by designing a group context activity descrip-

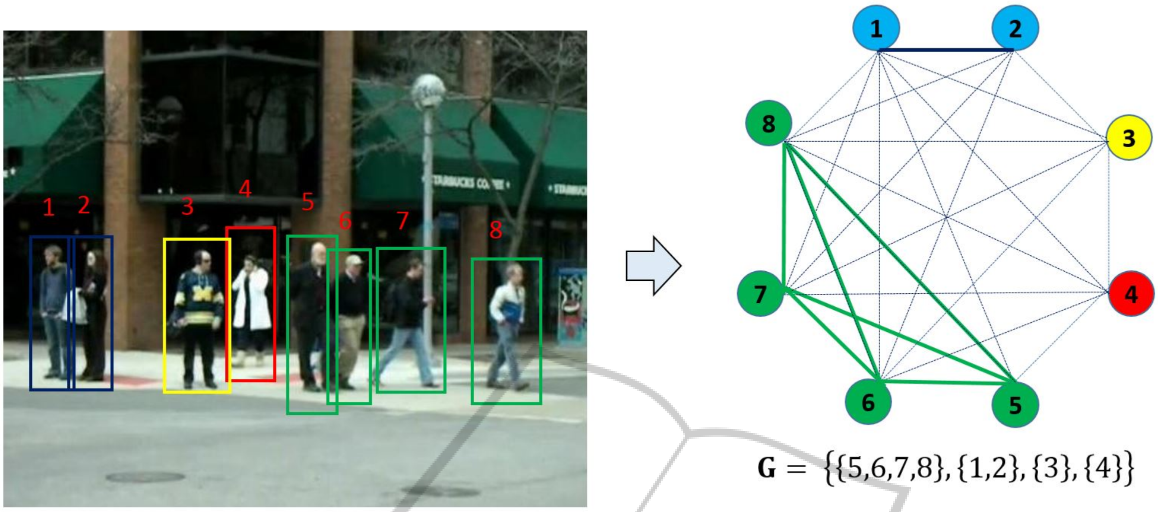


Figure 1: Illustration of group discovery. Human interactions in a group is represented as an undirected edge-weighted graph. Dominant set based clustering algorithm is used to localize interacting groups. There are four discovered groups from the scene: {5,6,7,8}, {1,2}, {3} and {4}.

tor capturing individual activity and behavior of its neighbors within its groups. Once meaningful groups are identified from the videos by using top-down approach, group context activity descriptor is built for each individual in discovered group. Using this descriptor a random field model is built to recognize individual activities using a bottom-up approach.

3 APPROACH

In this paper, we mainly focus on recognizing human activities in crowded scenes. Thus, we assume that people in a crowded scene have been detected and the trajectories of people in 3D space and the head poses are available or methods such as (Choi et al., 2009; Hoiem et al., 2006) can be used to obtain the same.

3.1 Group Discovery in Crowded Scene

In general, the analysis of complex activity in crowded scene is a challenging task, due to noisy observations and unobserved communication between people. In order to understand which people in the scene form meaningful groups, we employ a top-down approach proposed in (Tran et al., 2014) to discover socially interacting groups in the scene. This top-down approach basically represents all detected people as a graph where each vertex represents one person and weighted edges describe the social interaction between any two people in a group. The dominant set based clustering algorithm is used to discover the interacting groups (Pavan and Pelillo, 2007). Fig.1 depicts the overview of group discovery in the crowded scene.

3.2 Model Formulation

Given a set $\mathbf{N} = \{1,...,n\}$ of all the people detected in the scene, let $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ be the set of people activity descriptors; $\mathbf{a} = \{a_1, a_2, ..., a_n\}$ be the set of individual activity labels, where x_i is feature vector and $a_i \in \mathbf{A}$ is activity label associated with person $i \in \mathbf{N}$ ($\mathbf{A}$ is set of all possible activity labels). As a result of clustering people in the scene to different interacting groups, let us define $\mathbf{G} = \{\mathbf{G}_1, \mathbf{G}_2, ..., \mathbf{G}_m\}$ as the set of discovered groups where $\mathbf{G}_c$ is set of people clustered in group c and $\cup_{c=1}^{m} \mathbf{G}_c = \mathbf{N}$. We introduce a standard conditional random field model to learn the strength of the interactions between activities in discovered groups. The activity interaction is conditioned on image evidence, so that the model not only takes into account which activity each person is engaged in, but also higher-order dependencies between activities. Our model is represented as:

$$\Psi(\mathbf{a},\mathbf{x}) = \sum_{i \in \mathbf{N}} \phi(a_i, x_i) + \sum_{c=1}^{m} \sum_{(i,j) \in \mathbf{G}_c} \phi(a_i, a_j) \qquad (1)$$

where $\phi(a_i, x_i)$ is a singleton factor that models the probability of person i's activity label $a_i \in \mathbf{A}$ given its feature vector x_i. $\phi(a_i, a_j)$ is the pairwise factor that models the probability between pair of individual activities a_i and a_j, where (i,j) belong to the same group $\mathbf{G}_c$ discovered by using top-down approach described in section 3.1. A graphical illustration of our model discovering meaningful groups and formula-

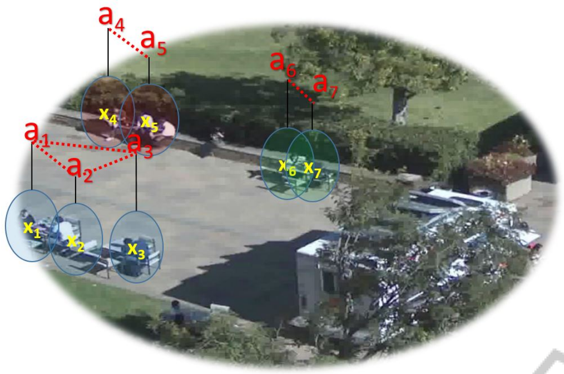


Figure 2: Illustration of our conditional random field model for each discovered interacting group. The activity-to-activity relationships in each group are represented by dashed lines.

tion of conditional random field model is shown in Fig.2.

The model described in Eq.1 only captures high-level activity-to-activity relationships of people in discovered groups. This limit us from analyzing the detailed interaction of individual activity and the behavior of its neighbors within a group. Thus we introduce a low-level group context activity descriptor that encodes detailed individual activity interactions within a group.

3.3 Group Context Activity Descriptor

An ideal context activity descriptor can efficiently incorporate the focal person's activity in spatio-temporal relationship with activities in its spatial proximity. Lan *et al.* (Lan et al., 2012b) propose an Action Context (AC) descriptor capturing the activity of the focal person and the behavior of other people nearby. AC descriptor uses spatial proximity as an indicator of context. They do not consider whether the people near-by are engaged in meaningful interactions or not, effectively leading to a semantically noisy descriptor. Moreover, we argue that AC is represented by concatenating a set of probability vectors computed using Bag-of-Words approach with SVM classifier that adds to ambiguity already existent in the representation (BoW) for each person. Choi *et al.* (Choi et al., 2009) employs well-known shape context idea (Belongie et al., 2002) to propose Spatio-Temporal Volume (STV) descriptor, which captures spatial distribution of pose and motion of individuals in a scene. The descriptor centered on a person of interest or an anchor is represented as histograms of people and their poses in different spatial bins around the anchor. STV descriptor can effectively capture higher-level spatial relationship of individual interactions. Nonetheless, it is too coarse to capture finer semantically driven contextual relationship of individual activities in detail.

We develop a novel group context activity (GCA) descriptor that exploits strategies from above approaches. Our descriptor is centered on a person (the focal person), and describes the behaviors of focal person and its semantic neighbors represented by arranging individual activity descriptors in polar view. Let $\mathbf{f} = \{f_1, f_2, ..., f_n\}$ be the set of local activity descriptors formed using Bag-of-Words representation for all people in the scene, where f_i is K-dimensional vector representing person i's activity (K is number of visual codewords). Dense trajectory based descriptors have shown to be efficient for representing actions in video, thus we employ approach proposed in (Wang et al., 2011) to extract motion boundary histogram (MBH) as local activity descriptors. Given the i-th person in discovered group $\mathbf{G}_c$ as the focal person, we divide its context region into P sub-polar context regions characterized by number of orientation bins and radial bins (Belongie et al., 2002). Using spatial relationship between people in discovered group $\mathbf{G}_c$, we extract descriptors in each sub-polar context region around the focal person. As a result, the group context activity descriptor x_i for person i is represented as a $(P+1) \times K$ dimensional vector including focal activity descriptor computed as follows:

$$x_i = [f_i, \sum_{j \in S_1(i)} f_j, \sum_{j \in S_2(i)} f_j, ..., \sum_{j \in S_P(i)} f_j] \qquad (2)$$

where $S_p(i)$ is set of people in the p-th sub-polar context region of person i. Fig.3 shows the extraction of group context activity descriptor for a selected person in a discovered group.

3.4 Inference and Learning

Our model is a standard Conditional Random Field (CRF) with no hidden variables. We train a multi-class SVM classifier based on GCA descriptors and their associated labels to learn and compute singleton factor $\phi(a_i, x_i)$. Given an observation x_i, we use SVM parameters to compute probabilities for all possible activity labels. From training data, we use top-down approach to discover interacting groups in the scene. All pairs of activity labels in discovered groups are counted to compute pairwise factor $\phi(a_i, a_j)$.

Given a new testing scene, our inference task is to find best activity label assignments for all people detected in the scene. The prediction assignment $\mathbf{a}^*$ is computed by running MAP inference on the network as:

$$\mathbf{a}^* = \arg\max_{\mathbf{a}} \Psi(\mathbf{a}, \mathbf{x}) \qquad (3)$$

where $\Psi(\mathbf{a}, \mathbf{x})$ is specified in Eq.1.

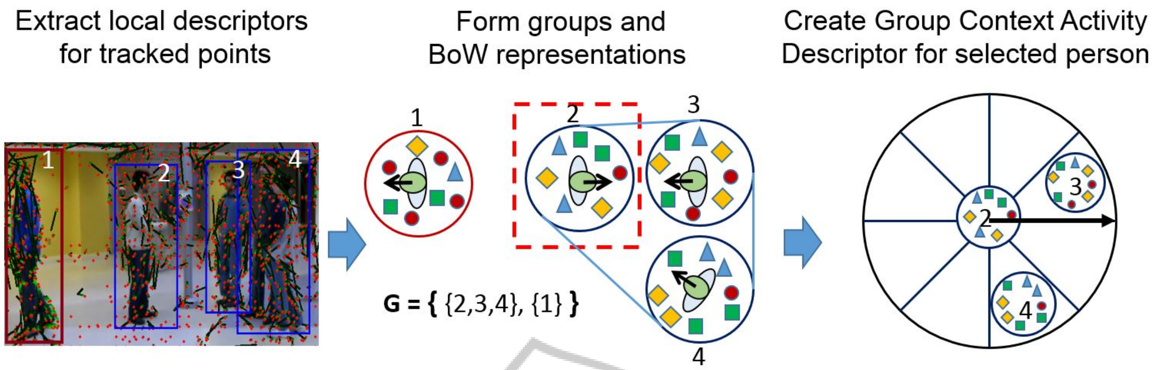

Extract local descriptors for tracked points
Form groups and BoW representations
Create Group Context Activity Descriptor for selected person
1
2
3
4
G = { {2,3,4}, {1} }

Figure 3: Depiction of Group Context Activity (GCA) descriptor extraction. From left to right, people are localized in different groups using a top-down approach from (Tran et al., 2014); local descriptors are extracted from dense trajectories (Wang et al., 2011); local BoW is computed for each person's activity; GCA descriptor is extracted for a selected person in a discovered group by computing descriptor for each sub-polar context bin.

4 EXPERIMENTS AND RESULTS

In this section, we describe the experiments designed to evaluate the performance of the proposed group context activity (GCA) descriptor and framework for human activity recognition in crowded scenes.

4.1 Datasets

In this work, we choose to use two challenging benchmark datasets to evalute our proposed approach in recognizing human activities in a crowded scene. The first benchmark dataset is Collective Activity dataset (Choi et al., 2009). The old version of dataset contains 5 activities in group (*Crossing, Waiting, Queuing, Walking* and *Talking*) and recently, the authors presented a new version of dataset including two additional activities (*Dancing* and *Jogging*). HOG based human detection and head pose estimation along with a probabilistic model is used to estimate camera parameters (Choi et al., 2009). Extended Kalman filtering is employed to extract 3D trajectories of people in the scene. These automatically extracted 3D trajectories and head pose estimates are provided as a part of the dataset. Thus, the dataset represents real world, noisy observations with occlusions and automatic person detection and trajectory generation.

The second benchmark dataset is UCLA Courtyard dataset recently introduced by Amer *et. al* (Amer et al., 2012). This dataset contains 106 minutes of high resolution videos at 30 fps from a bird-eye view of a courtyard at the UCLA campus. The annotations in term of bounding boxes, poses, and activity labels are provided for each frames in video. The dataset contains 10 primitive human activities which are *Riding Skateboard*, *Riding Bike*, *Riding Scooter*, *Driving Car*, *Walking*, *Talking*, *Waiting*, *Reading*, *Eating*, and *Sitting*.

4.2 Model Parameters

In using the group discovery algorithm (Tran et al., 2014), we set parameters that maintain the ratio proposed in (Was et al., 2006) and the social distance function is modeled as the power function $F_s(r) = (1-r)^n, n > 1$. We define 56 activity labels (8 head poses × 7 activity labels) for new version of Collective Activity dataset and 40 activity labels (8 head poses × 5 activity labels) for UCLA Courtyard dataset by combining the head poses and activity labels. We train a multi-class SVM classifier which is used to compute singleton factors by utilizing the libSVM library (Chang and Lin, 2011) with linear kernel on GCA descriptor. Using discovered groups from top-down approach, respectively, matrices of size 56×56 and 40×40 are used to learn and look up pairwise factors for Collective Activity and UCLA Courtyard datasets. For recognition, we use libDAI (Mooij, 2010) to perform inference in our conditional random field model.

To compute the MBH descriptors, we set the neighborhood size $N = 32$ pixels, the spatial cell $n_\sigma = 2$, the temporal cells $n_\tau = 3$, trajectory length $L = 10$, and dense sampling step size $W = 5$ for dense tracking. This setting claims to empirically give good results for a wide range of datasets (see (Wang et al., 2011) for parameter details). In designing GCA descriptor, we select codebook size of $K = 200$ by clustering a subset of $100,000$ randomly selected training features using k-means. In addition, we evaluate our proposed model in different settings of P, which is

Table 1: Recognition rates of various proposed methods on Collective Activity dataset (Choi et al., 2009).

									Accuracy (%)
Approach	Year	Walk	Cross	Queue	Wait	Talk	Jog	Dance	Avg
Choi (Choi et al., 2009)	2009	57.9	55.4	63.3	64.6	83.6	N/A	N/A	65.9
Choi (Choi et al., 2011)	2011	N/A	76.5	78.5	78.5	84.1	94.1	80.5	82.0
Amer (Amer and Todorovic, 2011)	2011	72.2	69.9	96.8	74.1	99.8	87.6	70.2	81.5
Amer (Amer et al., 2012)	2012	74.7	77.2	95.4	78.3	98.4	89.4	72.3	83.6
Lan (Lan et al., 2012b)	2012	68.0	65.0	96.0	68.0	99.0	N/A	N/A	79.1
Our Method		60.4	60.6	89.1	80.9	93.1	93.4	95.4	82.9

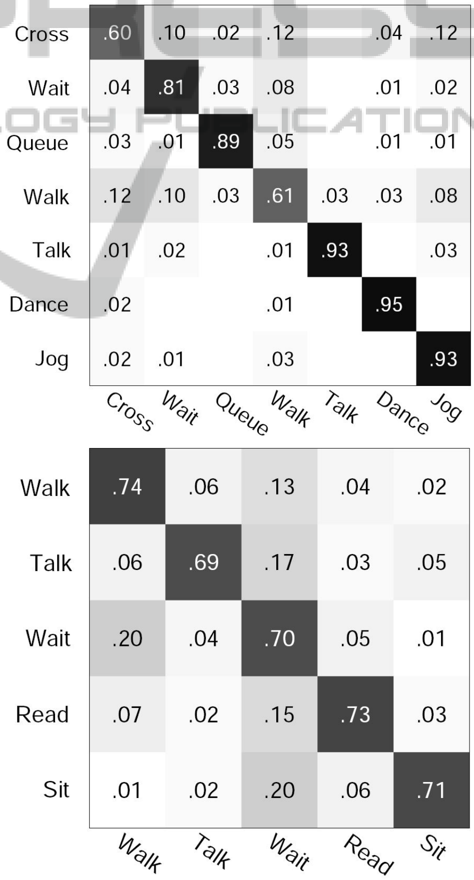
Table 2: Recognition rates of proposed methods on UCLA Courtyard dataset (Amer et al., 2012).

											Accuracy (%)
Approach	Walk	Wait	Talk	Drive Car	Ride S-board	Ride Scooter	Ride Bike	Read	Eat	Sit	Avg
Amer (Amer et al., 2012)	69.1	67.7	69.6	70.2	71.3	68.4	61.4	67.3	71.3	64.2	68.1
Our Method	74.3	69.9	70.0	N/A	N/A	N/A	N/A	72.8	N/A	70.8	71.4

number of sub-polar context regions around a focal person. Basically $P = R \times O$ where R is number of radial bins and O is number of orientation bins. However, given a focal person within his discovered group, context activity descriptor differs from others by discriminating in orientation distribution rather than radial distribution. Thus in our case, R is set to 1 and our GCA descriptor is controlled by $P = O$ number of orientation bins. For the special case when $P = 0$, GCA descriptor amounts to the focal person local activity descriptor without using context ($x_i = f_i$). This is the same for a non-group person in the scene who does not belong to any discovered groups. Experiments show that $P = 4$ and $P = 16$ achieves the best performances in Collective Activity and UCLA Courtyard datasets, respectively.

4.3 Human Activity Recognition Evaluation

We summarize the recognition results obtained using our method and other approaches in Table 1 for Collective Activity dataset using standard 4-folds cross-validation scheme. As we can see, our proposed approach achieves recognition rates comparable to state-of-the-art methods in the new version of Collective Activity dataset. Fig. 4(Top) shows the confusion matrices obtained on Collective Activity dataset. It lists the recognition accuracy for each activity individually. The low values of the non-diagonal elements imply that the descriptor is highly discriminative with very low decision ambiguity between different activities. The confusion matrix also shows the most confusion between *Walking* and *Crossing* activities in Collective Activity dataset, which can be explained because both are essentially *Walking* activity but with different scene semantic. Since most of *Waiting* activities are in social interaction with *Walking* activities in both datasets, so there is a relatively high confusion between *Waiting* and *Walking*. Overall, the confusion matrix shows very high accuracy rates in recognizing



	Cross	Wait	Queue	Walk	Talk	Dance	Jog
Cross	.60	.10	.02	.12		.04	.12
Wait	.04	.81	.03	.08		.01	.02
Queue	.03	.01	.89	.05		.01	.01
Walk	.12	.10	.03	.61	.03	.03	.08
Talk	.01	.02		.01	.93		.03
Dance	.02			.01		.95	
Jog	.02	.01		.03			.93

	Walk	Talk	Wait	Read	Sit
Walk	.74	.06	.13	.04	.02
Talk	.06	.69	.17	.03	.05
Wait	.20	.04	.70	.05	.01
Read	.07	.02	.15	.73	.03
Sit	.01	.02	.20	.06	.71

Figure 4: Confusion matrices for Collective Activity dataset (Top) and UCLA Courtyard dataset (Bottom).

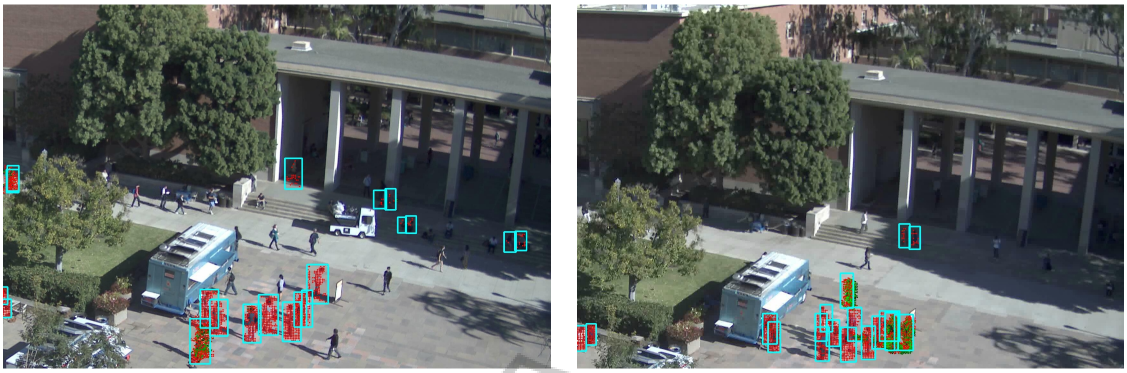


Figure 5: Depiction of computed dense tracks in UCLA Courtyard dataset. Due to low resolution, there are few tracked trajectories extracted for people in shadow regions. Thus, there are very few local activity descriptors extracted for people in those regions.

Queue, *Talk*, *Dance* and *Jog* activities. This can be explained because our group context activity descriptor efficiently encodes activities in different contexts.

For UCLA Courtyard dataset, Table 2 shows our recognition rate in comparison with other proposed methods, and Fig. 4(Bottom) shows the recognition confusion matrix. As we can see, our proposed approach achieves recognition rates that outperform the state-of-the-art methods in recognizing selected activities in UCLA Courtyard dataset. However, there is a limitation in using our framework to UCLA Courtyard dataset. The dense track algorithm proposed in (Wang et al., 2011) does not perform well across all observations in the UCLA Courtyard videos. There are very small number of dense trajectories extracted from people in shadow regions in comparison to other regions (see Fig. 5). Thus, there are not enough extracted descriptors to build GCA descriptor for those people in shadow regions. Using alternate feature detectors could alleviate this problem and hence the limitation towards computing the local activity descriptor for UCLA Courtyard videos. Due to this limitation, not all activities are included in our evaluation. Some activities such as *Riding Skateboard*, *Riding Bike*, *Riding Scooter*, *Driving Car*, and *Eating* are limited and hence do not provide sufficient exemplars for learning.

5 CONCLUSION

In this paper, we have proposed an efficient framework for recognizing human activities in crowded scenes. We have introduced a novel group context activity descriptor efficiently capturing focal person's activity and its neighbor's behavior. Along with group context activity descriptor, we also proposed a high-level recognition framework jointly captures the individual activity and its activity relationships with its neighbors. We evaluated our approach in two public benchmark datasets. The results demonstrate that our approach obtains results comparable to state-of-the-art in recognizing human activities in crowded scenes.

REFERENCES

Amer, M. R. and Todorovic, S. (2011). A chains model for localizing participants of group activities in videos. In *Proc. IEEE International Conference on Computer Vision*.

Amer, M. R., Xie, D., Zhao, M., Todorovic, S., and Zhu, S.-C. (2012). Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *Proc. European Conference on Computer Vision*, pages 187–200.

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, pages 27:1–27:27.

Chang, M.-C., Krahnstoever, N., Lim, S., and Yu, T. (2010). Group level activity recognition in crowded environments across multiple cameras. In *Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 56–63, DC, USA.

Choi, W., Shahid, K., and Savarese, S. (2009). What are they doing? : collective activity classification using spatio-temporal relationship among people. In *Proc. Visual Surveillance Workshop, ICCV*, pages 1282 – 1289.

Choi, W., Shahid, K., and Savarese, S. (2011). Learning context for collective activity recognition. In *Proc. Computer Vision and Pattern Recognition*, pages 3273 –3280, Spring CO, USA.

Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Bue, A. D., Menegaz, G., and Murino, V. (2011).

Social interaction discovery by statistical analysis of f-formations. In *Proc. British Machine Vision Conference*, pages 23.1–23.12.

Farenzena, M., Bazzani, L., Murino, V., and Cristani, M. (2009a). Towards a subject-centered analysis for automated video surveillance. In *Proc. International Conference on Image Analysis and Processing*, pages 481–489, Berlin, Heidelberg.

Farenzena, M., Tavano, A., Bazzani, L., Tosato, D., Pagetti, G., Menegaz, G., Murino, V., and Cristani, M. (2009b). Social interaction by visual focus of attention in a three-dimensional environment. In *Proc. Workshop on Pattern Recognition and Artificial Intelligence for Human Behavior Analysis at AI*IA*.

Helbing, D. and Molnár, P. (1995). Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286.

Hoiem, D., Efros, A., and Hebert, M. (2006). Putting objects in perspective. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages 2137–2144.

Khan, S. M. and Shah, M. (2005). Detecting group activities using rigidity of formation. In *Proc. ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 403–406, New York, NY, USA. ACM.

Lan, T., Sigal, L., and Mori, G. (2012a). Social roles in hierarchical models for human activity recognition. In *Proc. Computer Vision and Pattern Recognition*, pages 1354 –1361.

Lan, T., Wang, Y., Yang, W., Robinovitch, S., and Mori, G. (2012b). Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Mehran, R., Oyama, A., and Shah, M. (2009). Abnormal crowd behavior detection using social force model. In *Proc. Computer Vision and Pattern Recognition*, pages 935 –942.

Mooij, J. M. (2010). libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173.

Pavan, M. and Pelillo, M. (2007). Dominant sets and pairwise clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 167 –172.

Ryoo, M. and Aggarwal, J. (2011). Stochastic representation and recognition of high-level group activities. *International Journal of Computer Vision*, pages 183–200.

Smith, K., Ba, S., Odobez, J.-M., and Gatica-Perez, D. (2008). Tracking the visual focus of attention for a varying number of wandering people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1212 –1229.

Tran, K. (2013). *Contextual Descriptors for Human Activity Recognition*. PhD thesis, University of Houston.

Tran, K., Gala, A., Kakadiaris, I., and Shah, S. (2014). Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. *Pattern Recognition Letters*, 44(0):49 – 57. Pattern Recognition and Crowd Analysis.

Vaswani, N., Roy Chowdhury, A., and Chellappa, R. (2003). Activity recognition using the dynamics of the configuration of interacting objects. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages II – 633–40 vol.2.

Wang, H., Klaser, A., Schmid, C., and Liu, C.-L. (2011). Action recognition by dense trajectories. In *Proc. Computer Vision and Pattern Recognition*, pages 3169 –3176.

Was, J., Gudowski, B., and Matuszyk, P. J. (2006). Social distances model of pedestrian dynamics. In *Cellular Automata for Research and Industry*, pages 492–501.