# Activity Group Localization by Modeling the Relations among Participants

Lei Sun[1], Haizhou Ai[1], and Shihong Lao[2]

[1] Computer Science & Technology Department, Tsinghua University, Beijing, China
[2] OMRON Social Solutions Co. Ltd., Kusatsu, Shiga, Japan
ahz@mail.tsinghua.edu.cn

**Abstract.** Beyond recognizing the actions of individuals, activity group localization aims to determine "who participates in each group" and "what activity the group performs". In this paper, we propose a latent graphical model to group participants while inferring each group's activity by exploring the relations among them, thus simultaneously addressing the problems of group localization and activity recognition. Our key insight is to exploit the relational graph among the participants. Specifically, each group is represented as a tree with an activity label while relations among groups are modeled as a fully connected graph. Inference of such a graph is reduced into an extended minimum spanning forest problem, which is casted into a max-margin framework. It therefore avoids the limitation of high-ordered hierarchical model and can be solved efficiently. Our model is able to provide strong and discriminative contextual cues for activity recognition and to better interpret scene information for localization. Experiments on three datasets demonstrate that our model achieves significant improvements in activity group. localization and state-of-the-arts performance on activity recognition.

**Keywords:** Action recognition, group localization, graphical model.

## 1 Introduction

Vision-based human action and activity analysis have attracted much attention in computer vision literature. There has been quite a lot of work focusing on single-person action recognition [2], interactive activity between a person and objects [14,11], or pair-activities between two persons [16]. Collective activities, i.e. multiple persons performing activities in groups, however, is more common in real scenarios, with typical examples like: shopper queuing in a shopping store to get checked, pedestrians crossing a road, and friends talking together with their kids playing around. The analysis of such collective activity is of great practical importance for many applications such as smart video surveillance and semantic video indexing.

In this paper, we go beyond recognizing collective activities of individuals and focus on activity group localization in videos, which involves two distinct but related tasks: activity recognition and group localization. We seek to jointly solve

D. Fleet et al. (Eds.): ECCV 2014, Part I, LNCS 8689, pp. 741–755, 2014.

these two tasks by grouping individuals and reasoning activities at the group level. Noticeably, this incorporation of group information is in sharp contrast to most recent research in collective activity recognition, in which no group information is considered (e.g. regarding persons nearby as context for single person activity recognition [17,18,6,7] or modeling interactions or activity co-occurrences among some closely related persons [17,5]), leaving the whole relations among persons unclear.

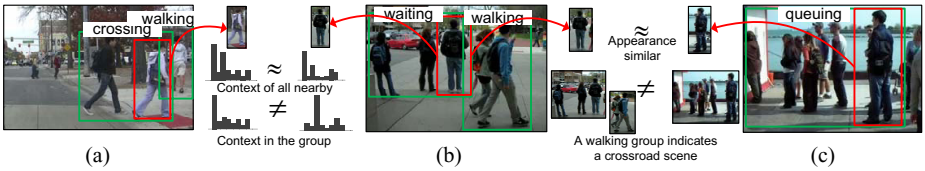



**Fig. 1.** Group helps action recognition. The green box denotes the activity group.

We argue that, instead of treating these two tasks separately, jointly addressing activity recognition and group localization enjoys many benefits. Firstly, it allows us to focus on recognizing activity on a group of persons and disregard those persons that are not discriminative or relevant. For example, in Fig. 1(a), the person in the red box is crossing. It will be confused to find his activities if we consider all his nearby persons as context. However, it will be much helpful if we only take the persons in his group and disregard the irrelevant persons in other groups. Secondly, it reduces the obscured relations of persons in the scene to person-person relations in each group and group-group relations among groups, thus enabling explicit modeling of such relations. In this way, by encapsulating individuals into groups, inter-group relations can better characterize the scene information. Take the queuing person boxed in red in Fig. 1(c) as an example, with similar appearance to the outlined person in Fig. 1(b), it still can be disambiguated since a co-existed "crossing" group implies a crossroad scenario. Last but not least, in perception, it is more sensible to discover activity groups than recognize individuals' actions. Group localization and activity recognition are mutually beneficial to each other. On one hand, group localization reveals the relations among participants in the scene, in which case more useful cues for activity recognition are obtained. On the other hand, activity recognition assists group localization in a more evident way, i.e., fusing activity information enables group localization at an activity level.

We propose a latent graphical model to jointly address two problems together, which we present in this work as a new problem called activity group localization. In particular, we employ a tree structure to represent each group and a fully connected graph to describe the relations among groups. Such graphical structure is quite sensible and is capable of capturing characteristics of group activities and co-occurrences among them. Then by dynamically inferring over this latent structure, the groups along with their activities can be consequently obtained.

Specifically, we treat it as an extended minimum spanning forest problem and utilize a max-margin framework to efficiently solve it.

The contributions of our work can be summarized in three-fold. Firstly, we advance prior work of individual activity recognition to activity group localization by jointly addressing group localization and activity recognition. Secondly, a relational graph is presented to model the relations among participants, which gives an interpretable description of the scene information and thus largely assists the activity recognition as well as group localization. Thirdly, we solve the graphical model as an extended spanning forest problem, and cast it into a max-margin framework which enables efficiently inference over the graph structure.

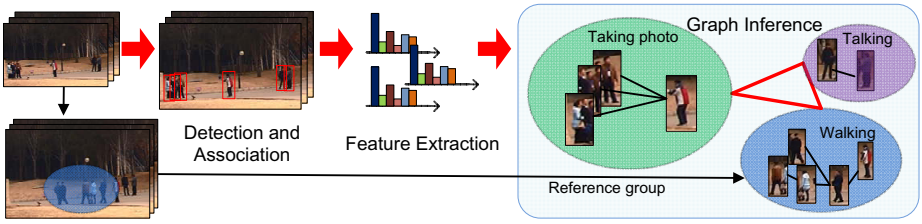



**Fig. 2.** System overview. First detect and associate the persons in the video, then a relational graph is constructed and inferred with respect to activities as well as groups.

## 2 Related Work

Many recent works on human action recognition model the context explicitly to assist recognition. For example, the contextual information is exploited by means of scenes[20], objects [14,11], or interactions between two or more objects[16]. The scene or role interactions are explored by many researchers using sophisticated models like dynamic Bayesian networks [23], CASE natural language representations [13], AND-OR graphs [12], and probabilistic first-order logic [21,3].

In group activity or collective activity recognition, context generally means what others are doing. Some methods attempt to provide contextual information for single person activity classification by concatenating the action scores of all the neighbor persons [17,18] or extracting spatiotemporal distributions of surroundings persons [6,7]. Some mid-level atomic interactions are captured to encode the relation between a pair of persons [5]. Unfortunately, such kind of interactions only provide useful information for interactive activities such as "talking" but rarely occur in other casual activities such as "crossing", and "waiting". Besides, involving the atomic interactions also complicates the problem. Rather than recognizing individual's activity in isolation, some approaches [17,18,5] attempt to jointly classify all people in a scene. In this case, a hierarchical model is often used to model the compatibility of the activities among person-person and person-group. To the best of our knowledge, none of previous approaches

explicitly captures the overall relations among participants, which we believe is of critical importance for activity recognition. Another related issue in collective activity recognition is using the tracking information [19], e.g., to formulate multi-target tracking and action recognition into a constrained minimum cost flow problem [15], or to integrate tracking, atomic activities, interactions and collective activities together to form a hierarchical model, and infers them by combining belief propagation and branch and bound [5].

There is not much work about grouping activity groups. They typically focus on one aspect of this problem, e.g., to determine the group location by developing contextual spatial pyramid descriptor while neglecting individual activity [22], or to infer the individual activity by a chain model [1]. Some other approaches [4] attempt to cluster individual with specific scenarios and strict rules, which is not suitable for collective activity.

Our work is to some extent related to the model in [17,18], where a hierarchical model is proposed to model the compatibility of image, action and activity. They also attempt to implicitly infer the person-person relations using sparse loopy graph structure. However, such a sparse structure does not completely characterize the relations. Our work here emphasizes on the structure of relations among participants, which leverages visual patterns, motions and activity compatibility in terms of intra-group relations and inter-group relations.

## 3 System Overview

The proposed framework is illustrated in Fig. 2. Our main objective is to localize activity groups in a video. Haven the persons detected and associated, for each single image, we construct a relational graph, which is then inferred with respect to groups as well as their activity labels. Notice that there are some reference groups participated in the relational graph. Such groups, coming from previous frames of the video, are often those that have been identified as reliable activity groups. In this case, They play a role of authority for further verification. We each time select one reference group to participate in the graph inference and take the relational graph with the highest score as the final result.

Here we emphasize on how to model and solve the relational graph, which attempts to encode the relations among persons and groups. We assume that in each group every person closely coordinates with only one another, i.e. a tree structure. As for inter-group relations, we remain groups fully connected (Fig. 3(a)). Notice that, solving this relational graph is reduced to a clustering problem if no activity recognition is required, and such clustering can efficiently be modeled as a minimum spanning forest problem. Therefore, we seek to solve our activity localization problem tailoring an extended version of minimum spanning forest problem, of which the difference is that each tree is with an extra activity label and is connected to every other one. In the next, we start by explaining how to model such a graph in Section 4, then describe the learning of the model in Section 5 and model inference in Section 6.

# 4 Modeling Activity Group Localization

## 4.1 Model Formulation

Given a set of detected persons $\mathbf{x}=\{x_1, x_2, ..., x_m\}$ in the image and a reference group $(\mathbf{x}_r, g_r, a_r)$, the objective is to find the groups $\mathbf{g}=\{g_1, g_2, ..., g_n\}$ with activity labels $\mathbf{a}=\{a_1, a_2, ..., a_n\}$, where $g_i=(g_{i1}, g_{i2}, ..., g_{im})$ with $g_{ik} \in \{0, 1\}$ indicating whether the $k$th person belongs to the group $g_i$ or not ($\sum_i g_i=1_m, \forall i, j, g_i g_j^T=0$), and $a_i \in A$ with $A$ being the set of all possible activity labels. Let $h$ denote the relational graph structure, as shown in Fig. 3(a). It consists of $n$ trees, $h=\{t_n\}$, each representing one activity group $(g, a)$. We use $F_{\mathbf{w}}(\mathbf{x}, h, \mathbf{g}, \mathbf{a})$ to measure the compatibility among activity groups $(\mathbf{g}, \mathbf{a})$, graph structure $h$ and persons $\mathbf{x}$. And by maximizing such a potential function, the optimum assignment of $(\mathbf{g}, \mathbf{a})$ for $\mathbf{x}$ can be obtained. Note that we include the reference group $(\mathbf{x}_r, g_r, a_r)$ into the current notation $(\mathbf{x}, \mathbf{g}, \mathbf{a})$ for simplicity, which will be discussed in detail in the following.

Two kinds of potentials are developed to measure the compatibility function. The first regards to intra-group potential, which we attempts to model the compatibility of a pair of individuals' belonging to one activity group, while the second, inter-group potential, characterizes the compatibility of a pair of activity groups belonging to the same scene. Therefore, the potential function $F_{\mathbf{w}}(\mathbf{x}, h, \mathbf{g}, \mathbf{a})$ is formulated as

$$F_{\mathbf{w}}(\mathbf{x}, h, \mathbf{g}, \mathbf{a}) = \mathbf{w}_p^T \psi_p(\mathbf{x}, h, \mathbf{g}, \mathbf{a}) + \mathbf{w}_g^T \psi_g(\mathbf{x}, h, \mathbf{g}, \mathbf{a}), \tag{1}$$

where $\mathbf{w}_p^T \psi_p(\mathbf{x}, h, \mathbf{g}, \mathbf{a})$ measures intra-group compatibility, $\mathbf{w}_g^T \psi_g(\mathbf{x}, h, \mathbf{g}, \mathbf{a})$ scores inter-group compatibility. The model parameters are the combination of $\mathbf{w}_p^T$ and $\mathbf{w}_g^T$, $\mathbf{w} = [\mathbf{w}_p^T \ \ \mathbf{w}_g^T]^T$. The details of Eq. 1 are described in the following.

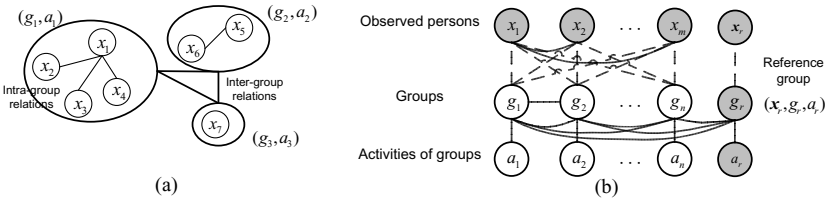



**Fig. 3.** (a) shows the relational graph. Grey node in (b) denote observable variants.

**Intra-Group Potential** $\mathbf{w}_p^T \psi_p(\mathbf{x}, h, \mathbf{g}, \mathbf{a})$: This function encodes the relation among a pair of persons and their belonged group. It is parameterized as:

$$\mathbf{w}_p^T \psi_p(\mathbf{x}, h, \mathbf{g}, \mathbf{a}) = \sum_{t \subseteq h} \sum_{(x_i, x_j) \in t} \sum_{b \in A} \mathbf{w}_{pb}^T \phi(x_i, x_j) \mathbf{1}(a_t = b), \tag{2}$$

where $\mathbf{1}(.)$ is the indicator function, and $\phi(x_i, x_j)$ denotes the person-person descriptor (Sec. 4.2). The parameter $\mathbf{w}_p$ is simply the concatenation of $\mathbf{w}_{pb}$ for all $b \in A$.

**Inter-Group Potential** $\mathbf{w}_g^T \psi_g(\mathbf{x}, h, \mathbf{g}, \mathbf{a})$: This function characterizes the relation between all pairs of groups. It is parameterized as:

$$\mathbf{w}_g^T \psi_g(\mathbf{x}, h, \mathbf{g}, \mathbf{a}) = \sum_{(t_i,t_j)\in h} \sum_{b,c\in A} \mathbf{w}_{gbc}^T \varphi(t_i, t_j) \mathbf{1}(a_{t_i} = b) \mathbf{1}(a_{t_j} = c), \quad (3)$$

where $\varphi(t_i, t_j)$ denotes the group-group descriptor (Sec. 4.2). By adding reference group in this term, additional group pairs are modeled with knowledge of reference group's activity label.

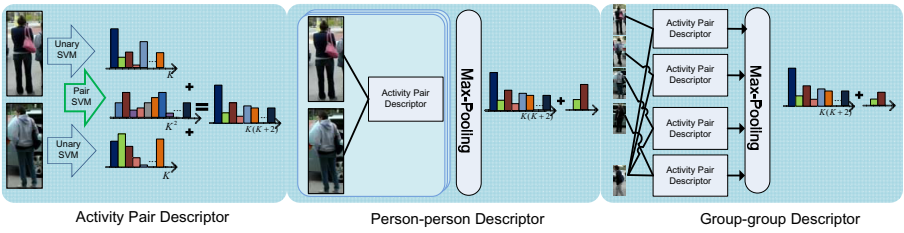



**Fig. 4.** Activity pair descriptor

### 4.2 Activity Pair Descriptor

We build this descriptor in two stages. Firstly, we train a multi-class SVM classifier (*unary* SVM) based on the person descriptors (e.g. HOG [8]) and their associated action labels, then each person can be represented as a $K$-dimensional vector, where $K$ is the number of activity classes. Secondly, we train a multi-class SVM classifier (*pair* SVM) on a pair of persons and their activity labels. Each person pair is represented as an $K_2$-dimensional vector. Our activity pair descriptor is computed by concatenating two person's action descriptor and the pairwise action descriptor, which ends up with a $K(K+2)$-dimensional vector, as shown in Fig. 4. The feature used to train the *pair* SVM for a person pair $(x_i, x_j)$ is denoted as

$$\mathbf{f}_{(x_i,x_j)} = [\mathbf{d}_{x_i}\ \mathbf{d}_{x_j}\ \mathbf{d}_{x_i} - \mathbf{d}_{x_j}\ \mathbf{d}_{x_i} \otimes \mathbf{d}_{x_j}\ \mathbf{c}], \quad (4)$$

where $\mathbf{d}_{x_i}$, $\mathbf{d}_{x_j}$ are the person descriptors of the person $x_i$ and $x_j$, respectively. The operator $\otimes$ means element-wise multiplication. $\mathbf{c}$ is the bag-of-words representation of the scene's context.

**Person-Person Descriptor** $\phi(x_i, x_j)$: To compute the person-person descriptor for a person pair $(x_i, x_j)$,we do not only consider the visual appearance in the current frame, but also take advantage of association which locates the persons

in the neighbor frames. Let $N(x)$ be the set of tracked human across neighbor frames for person $x$, then we compute activity pair descriptors of all possible person pairs $P(x_i, x_j)$, of which the first is from $N(x_i)$ and the second is from $N(x_j)$. Note that, only reliable tracklets are used in our work, so if none of the tracklets coveres the person $x$, $N(x)$ will only have one element $x$. Finally we calculate the person-person descriptor as follows

$$\phi(x_i, x_j) = [\max_{p \in P(x_i, x_j)} S_{p,1}, ..., \max_{p \in P(x_i, x_j)} S_{p,K(K+2)}, \ \ l_x, \ \ l_y], \tag{5}$$

where $S_{p,k}$ denotes the $k$th value of the activity pair descriptor, $l_x$ and $l_y$ are the average relative deviations of all pairs at the x and y coordinates, respectively.

**Group-Group Descriptor** $\varphi(t_i.t_j)$: For all person pairs $(x_m, x_n)$, where $x_m$ comes from one group $t_i$ and $x_n$ comes from the other group $t_j$, we compute the person-person descriptors $\phi(x_m, x_n)$. The final group-group descriptor is obtained using the following equation

$$\varphi(t_i, t_j) = [\max_{x_m \in t_i, x_n \in t_j} \phi(x_m, x_n)_1, ..., \max_{x_m \in t_i, x_n \in t_j} \phi(x_m, x_n)_{K(K+2)}, \ \ l_x, \ \ l_y], \tag{6}$$

where $\phi(x_m, x_n)_k$ is the $k$th value in the person-person descriptor, $l_x$ and $l_y$ are the average relative deviations of all pairs at the x and y coordinates, respectively.

### 4.3 Reference Groups

Reference groups are those that have been identified as reliable activity groups in the previous frames. They, in a sense, serves as some explicit scene information. Given a *crossing* group in the scene, it is more likely to tell a group of standing persons to be a *waiting* group rather than a *talking* group. Specifically, the activity groups with confidence that exceeds a threshold (set empirically) are pushed into a reference group pool. And concerning the computation, we select a subset of reference groups with little overlap with the current groups' region (the total number of reference groups is discussed in Section 7.2). Such strategy is reasonable since our model favors seeing complete relational graph located in various regions.

## 5 Model Learning

Our scoring function can be converted into an inner product $\langle \mathbf{w}, \psi(\mathbf{x}, h, \mathbf{g}, \mathbf{a}) \rangle$, where $\mathbf{w} = [\mathbf{w}_p^T \ \ \mathbf{w}_g^T]^T$, $\psi(\mathbf{x}, h, \mathbf{g}, \mathbf{a}) = [\psi_p(\mathbf{x}, h, \mathbf{g}, \mathbf{a}) \ \ \psi_g(\mathbf{x}, h, \mathbf{g}, \mathbf{a})]$.

Given a set of $N$ training examples $(\mathbf{x}^n, \mathbf{g}^n, \mathbf{a}^n)$ $(n = 1, 2, ..., N)$, we train the model parameter $\mathbf{w}$ to produce the correct groups $\mathbf{g}$ and activity labels $\mathbf{a}$. Note that the groups and activity labels can be observed on training data, but the graph structure $h$ is unobserved. We adopt the latent SVM [10] formulation to train this model, which in our case can be written as follows

$$\begin{aligned}\mathbf{w}^* = \arg\min_{\mathbf{w}}\{\tfrac{1}{2}||\mathbf{w}||^2 - C\sum_{i=1}^{N}\mathbf{w}^T\psi(\mathbf{x}^i, h^i, \mathbf{g}^i, \mathbf{a}^i) \\ +C\sum_{i=1}^{N}\max_{(\hat{\mathbf{g}},\hat{h},\hat{\mathbf{a}})}[\mathbf{w}^T\psi(\mathbf{x}^i, \hat{h}^i, \hat{\mathbf{g}}^i, \hat{\mathbf{a}}^i) + \Delta(\mathbf{g}, \hat{\mathbf{g}}, \hat{h}, \mathbf{a}, \hat{\mathbf{a}})]\},\end{aligned} \tag{7}$$

where $C$ controls the tradeoff betwen the errors in the training model and margin maximization and $\Delta(\mathbf{g}, \hat{\mathbf{g}}, \hat{h}, \mathbf{a}, \hat{\mathbf{a}})$ is the loss function. Naturally, this function need penalize both incorrect groups and incorrect activity labels. We define it as follows

$$\Delta(\mathbf{g}, \hat{\mathbf{g}}, \hat{h}, \mathbf{a}, \hat{\mathbf{a}}) = n(\mathbf{g}) - \sum_{(x_i,x_j)\in\hat{h}} l(\mathbf{g}, \mathbf{a}, \hat{\mathbf{a}}, (x_i, x_j)), \tag{8}$$

where $n(\mathbf{g})$ is the difference of the number of nodes and the number of groups. The function $l(\mathbf{g}, \mathbf{a}, \hat{\mathbf{a}}, (x_i, x_j))$ returns 1 if $(x_i, x_j)$ belongs to the same group with the correct activity, returns 0 if $(x_i, x_j)$ belongs to different groups but with the correct activity, and -1 otherwise. It is easy to show that such a loss function equals zero if and only if the individuals are clustered into correct groups and with correct activities.

## 6 Model Inference

Given the model parameter $\mathbf{w}$, the inference problem is to find the best group locations $\mathbf{g}$ along with the corresponding activity label $\mathbf{a}$ for each input $\mathbf{x}$. Using the latent SVM formulation, it can be written as:

$$F_{\mathbf{w}}(\mathbf{x}, \mathbf{g}, \mathbf{a}) = \max_{a}\max_{g,h} F_{\mathbf{w}}(\mathbf{x}, h, \mathbf{g}, \mathbf{a}). \tag{9}$$

Since groups $\mathbf{g}$, graph structure $h$ and activities $\mathbf{a}$ are not independent with each other, the optimization of Eq. 10 is NP-hard. When the number of persons and activities are small and some spatial restrictions can be incorporated, we encourage a combinatorial search to generate exactly inference. In other cases, we approximately solve it by iterating the following three steps:

- Holding activities $\mathbf{a}$ and groups $\mathbf{g}$ fixed, optimize the graph structure $h$, using a standard spanning tree algorithm such as Kruskal's algorithm.
- Holding graph structure $h$ and groups $\mathbf{g}$ fixed, optimize the activities $\mathbf{a}$ by enumerating all possible activities.
- Holding activities $\mathbf{a}$ fixed, generate new optimal groups $\mathbf{g}$ by merging two trees or splitting one tree with the same activity in the current structure $h$.

The three steps are iterated until converged. While this algorithm cannot guarantee a globally optimum solution, in our experiments it works well to find good solutions.

## 7 Experiments

**Datasets.** We evaluate our method on two collective activity datasets from [6,7] and a newly recorded dataset collected by ourselves. The first collective activity dataset is composed of 44 video clips with 5 activities, *crossing*, *walking*, *queuing*, *talking* and *waiting*. While the second is an extended dataset of the former. It includes two more classes of *dancing* and *jogging* and removes the ill-defined *walking* class, which results in 6 class of activities. We refer these two collective activity datasets as 5-class collective dataset and 6-class collective dataset, respectively. We use the activity annotations provided by [6] and further annotate the groups with bounding boxes. We also collect several 10-minute videos from a outdoor touring environment, and we segment them into 52 video clips, each having 800 to 1000 frames. Typical collective activities include *walking*, *bicycling*, *taking photos*, *standing*, and *talking*. This new dataset is referred as touring dataset. We annotate the activity label for each person and the groups in every tenth frame (4560 annotated frames including 5067 *walking*, 3126 *bicycling*, 3228 *taking photos*, 3850 *standing*,3027 *talking*).

**Evaluation Metric.** We stress that our objective is to localize activity groups, two aspects are evaluated: activity recognition and group localization. For localization, we compute a ratio of the intersection and union of detected and ground-truth bounding boxes of people participating in activities. The activity group is correct only if ratio $> 0.5$ and activity is correct.

**Implementation Details.** For the 5-class collective dataset and 6-class collective dataset, we apply the pedestrian detector in [9], and obtain some reliable tracklets by simply associating the detected bounding boxes in two neighboring frames using spatio-temporal locations and appearance similarity. For touring dataset, however, pedestrian detectors are not enough, we additionally apply background subtraction in [24] using Gaussian Mixture Model to obtain foreground objects. Instead of using raw features (e.g. HOG), we follow the setting in [17] to extract Action Context (AC) descriptor as the person descriptor. Also, the $\mathbf{c}$ in Eq. 4 is constructed by computing the histogram of visual-words within the persons appearing in neighbor frames. Specifically, we extract HOG feature of persons and apply $k$-means to generate 200 codewords.

### 7.1 Activity Recognition

In this part, we concentrate on activity recognition task. First we demonstrate the effective of our model by comparing to several baselines. Then we make comparisons with the state-of-the-art approaches with respect to three different validation schemes. We also analyze the behavior of our model in terms of learnt weights and total number of reference groups.

First we construct several baselines to demonstrate the capability of our model to interpret context in terms of a relational graph, activity pair descriptor and reference groups, which largely improves activity recognition. To

evaluate the performance of our relational structure, three baselines with different graph structures are considered as shown in Fig. 5. The first (*unary person*) is a latent SVM model based on AC descriptor. It simply regards all nearby persons as context and attempts to infer their activities. It can be formulated as $F_{\mathbf{w}}(\mathbf{x},\mathbf{g},\mathbf{a})=\sum_i \sum_b \mathbf{w}_b x_i \mathbf{1}(a_i = b)$, where $\mathbf{x}$ denotes all persons in a euclidean distance. The second (*sparse link*) adopts the structure in [18]. To our knowledge, [18] is the only work that has mentioned about the structure of participates, and in particular, they tend to find sparse but important links between persons by maximize the summation of all pairwise activity potential under a maximum limitation of each vertex's degree, which can be formulated as $F_{\mathbf{w}}(\mathbf{x},\mathbf{g},\mathbf{a})=\sum_{(i,j)} \sum_{(b,c)} \mathbf{w}_{bc}\psi_p(x_i,x_j)\mathbf{1}(a_i{=}\, b)\mathbf{1}(a_j{=}\, c)$, $s.t.\ \forall i, d(x_i) \leqslant q$. $d(x_i)$ denotes the degree of the vertex while $q$ is a threshold. The third baseline (*unary group*) ignores the pairwise group structure, which is equivalent to our model in Eq. 1 by removing inter-group potential term.

We also evaluate the performance of our activity pair descriptor by replacing it with a concatenated vector by two AC descriptors of a person pair in *full* model, which we called *2-AC*. And the performance of reference group (*non-reference*) is evaluated by removing them from our *full* model.

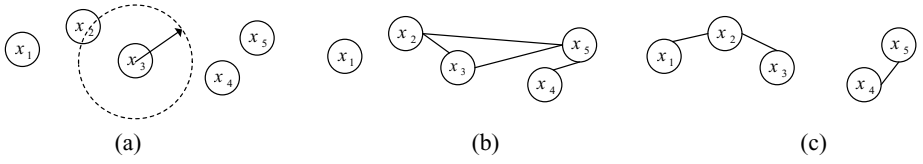



**Fig. 5.** (a)-(c) shows the graph structure of three baselines: *unary person model*, *sparse link model* and *unary group model*, respectively

**Table 1.** Mean average activity classification accuracy on three datasets

| Dataset | unary person | sparse link | unary group | 2-AC | non-reference | full |
|---|---|---|---|---|---|---|
| 5-class collective dataset | 54.8% | 69.4% | 62.3% | 72.9% | 71.6% | **74.8**% |
| 6-class collective dataset | 67.3% | 80.2% | 72.1% | 83.2% | 82.9% | **85.8**% |
| Touring dataset | 49.2% | 58.4% | 53.6% | 67.1% | 66.2% | **68.3**% |

Table 1 summarizes the results using leave-one-video-out validation strategy. We can see that our model significantly outperforms all baselines with respect to all the three datasets. Consider the structural baselines, the *unary person* model is almost unachievable, especially in a more complex scenario of touring dataset. This can be attributed to the unstructured context, drawing all persons nearby as context information introduces much noise as well as irrelevant persons. Compared to the *unary person* model, *unary group* model and *sparse link* model achieve a large improvement of performance, which further proves the effect of modeling the relations among participants. We can also see that

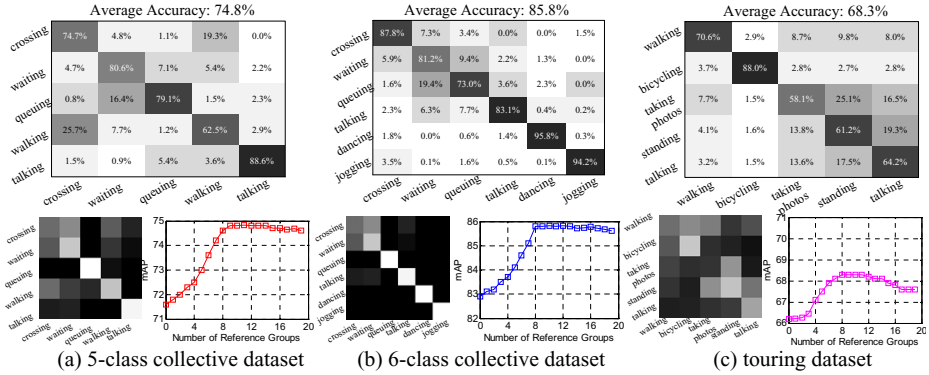
*sparse link* has a better performance by about 6% than *unary group* model. It is quite reasonable, since *unary group* model imposes quite a few links on the relations while *sparse link* model might find multiple important pairwise relations. In comparisons to these three baselines, our *full* model has a significant boost in performance, showing the advantage of integrating groups into activity recognition and modeling pairwise group relations. We also observe a slight degradation in performance occurs when a combination of AC descriptor is used instead of activity pair descriptor, suggesting that our model is able to perform competitively well even with "poorer" descriptors. Moreover, we find that reference groups leads to better results by comparing *non-reference* model to *full* model. We attribute this to the complementary scene information from the video provided by reference groups.

Then we make quantitative comparisons with other state-of-the-art approaches on the 5-class and 6-class collective datasets, including RSTV approach in [7], a joint tracking and recognition flow model in [15], a complex hierarchical model in [5], a bayesian BORD method in [1] and a discriminative latent model in [17]. To be comparable to these reported results, we adopt their respective training/testing schemes and evaluation criteria.

We summarize the results using three validation schemes in Table 2. The first scheme is the leave-one-video-out (LOO) training/testing scheme and per-person activity classification is evaluated, which is used in [6,7,5,15]. Our model outperforms all approaches by achieving an overall accuracy of 74.8% on the 5-class collective dataset and 85.8% on the 6-class collective dataset. Notice that, the model from [5] yields competitive results as our model for the first dataset. However, it employs a complex hierarchical model, which requires additional pose orientations of each person, 3D trajectories and some interactive atomic actions. The second experiment is to train the model on three fourths of the dataset while testing on the remaining fourth, and to evaluate per-scene activity classification. We follow the same split of dataset suggested by [17], and achieve 81.2% on the 5-class collective dataset. It is superior than 79.1% and 80.4% reported in [17] and [5]. The last experiment adopts the scheme in [1], which merges 5-class collective dataset and 6-class collective dataset to form a 7-class collective dataset ( *walking* activity is not removed). They use 2/3 and 1/3 of the videos from each class for training and testing. Our model reports 83.7% accuracy which is 2.2% higher than [1]. To demonstrate the effective of our model in a more complex scenario, we re-implement the adaptive structured latent SVM method in [18], and achieve 61.5% accuracy on touring dataset using leave-one-video-out validation scheme, which is 6.8% lower than our performance.

**Table 2.** Comparisons with the state of the art on two collective datasets

| Validataion | Approaches | RSTV [7] | RSTV+MRF [7] | AC+Flow[15] | T.+A.+I.[5] | AC+LSVM[17] | BORDS[1] | full model |
|---|---|---|---|---|---|---|---|---|
| LOO | 5-class | 67.2% | 70.9% | 70.9% | 74.4% | – | – | **74.8%** |
| LOO | 6-class | 71.7% | 82.0% | 83.7% | – | – | – | **85.8%** |
| one fourth | 5-class | – | – | – | 80.4% | 79.1% | – | **81.2%** |
| one third | 7-class | – | – | – | – | – | 81.5% | **83.7%** |




(a) 5-class collective dataset (b) 6-class collective dataset (c) touring dataset

**Fig. 6.** The confusion matrixes (upper row), the learnt weights across different activity groups and precision-reference groups curves (lower row)

At last, to give a clear insight into our model, we give the confusion matrixes and the learnt pairwise weights across different activity groups in Fig. 6. The learnt weights encode some scene information, which further demonstrates the benefits of including pairwise group relations. For example, *walking* groups are more likely to be co-existed with *crossing* groups, while *queuing* groups tend to be appear alone. *standing* groups have high co-occurrences with *taking photos* groups. Besides, the performance with different number of reference groups are illustrated in Fig. 6. It indicates that 8 to 15 is optimal, in which case larger causes computation issue while smaller leads to insufficient scene cues.

### 7.2 Activity Group Localization

In this part, we evaluate our model for the task of activity group localization. To our knowledge, there is only one work [22] about activity group localization, so we re-implement their method (*CSPM*), and compare our results with it. In order to investigate the capability of our model to localize activity groups, we construct two step-wise baselines: a) we estimate the activity label of each person (use the re-implemented version of [18]), followed by a mean-shift clustering algorithm (*activity-cluster*), and b) we remove the latent activity term in Eq. 1, to formulate a clustering method based on our activity pair descriptor (no reference groups), and then use a multi-class SVM to classify the activity of each group using the max-pooled AC descriptors within each group (*cluster-activity*).

As Fig. 7 shows, our model achieves a significant improvement with respect to all activity groups over [22] as well as two baselines. Such good performance resides in not only explicit activity inference but also the pairwise group relations modeling. The work in [22] proposes a contextual spatial pyramid descriptor and attempts to localize one particular group at one time. Though it might implicitly characterize the variations of activity, it lacks the ability to account for the correspondences between groups. The first baseline is a conventional step-wise

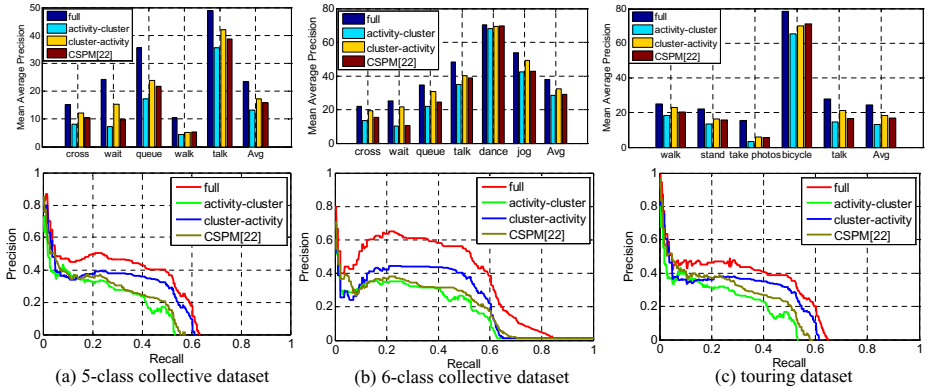



**Fig. 7.** The mean average precisions and precision-recall curves of localization

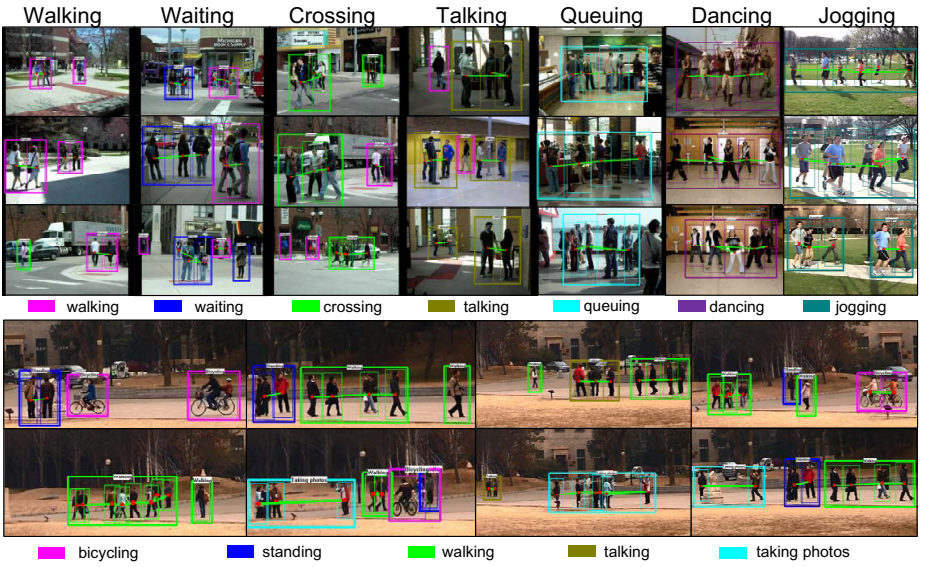



**Fig. 8.** Results on two collective datasets (upper row) and touring dataset (lower row). Bold rectangle denotes an activity group, with its color indicating its activity. Red line denotes an edge in the tree structure.

method to localize activity groups. It clearly suffers from the unreliable activity recognition. The second baseline, on the other hand, attempts to first cluster the groups and then to recognize their respective activities, of which the poor performance implies that clustering on visual cues is not sufficient.

We visualize the activity group localization results and the learned structure among participants in Fig. 8. Some interesting inner group tree structures are learnt, like a chain structure which connects all people for queuing activity,

one or two connections between people facing each other for talking and short links between people nearby having similar standing pose for waiting. As can be seen from Fig. 8, this kind of structure reveals some discriminative relations and disregards irrelevant ones, and also mitigates the impact from occlusions by only linking the overlapped person to one other person. Furthermore, our model, unlike previous approaches that often attempts to recognize the dominant activities, favors seeing different activity groups, thus can effectively disambiguate non-dominant activities and is more suitable for complex scenes.

## 8 Conclusions

In this paper, we aim at activity group localization including two tasks: group localization and activity recognition. A relational graph is proposed to model the relations among participants, which is solved as an extended problem of minimum spanning forest. We demonstrated that the incorporation of group helps to classify collective activities, and it is especially useful for structure-rich activities. With context structured by a relational graph, our proposed model can achieve competitive results comparing with the state-of-the-art approaches using three different validation schemes. In return, the activity group localization accuracy is also significant improved by jointly inferring the activities. In future work, we plan to exploit this group structure to mine group activity in long surveillance videos.

**Acknowledgements.** This work is supported in part by the 973 Program of China under Grant No.2011CB302203 and is also supported by a grant from Omron Corporation.

## References

1. Amer, M.R., Todorovic, S.: A chains model for localizing participants of group activities in videos. In: ICCV (2011)
2. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV (2005)
3. Brendel, W., Todorovic, S., Fern, A.: Probabilistic event logic for interval-based event recognition. In: CVPR (2011)
4. Chang, M.C., Krahnstoever, M., Lim, S., Yu, T.: Group level activity recognition in crowed environments across multiple cameras. In: Workshop on Activity Monitoring by Multi-camera Surveillance System (2010)
5. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 215–230. Springer, Heidelberg (2012)
6. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: VSWS (2009)
7. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: CVPR (2011)

8. Dalal, N., Triggs, B.: Histogram of oriented gradients for human detection. In: CVPR (2005)
9. Duan, G., Huang, C., Ai, H., Lao, S.: Boosting associated pairing comparison features for pedestrian detection. In: Workshop of ICCV (2009)
10. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
11. Gupta, A., Davis, L.S.: Objects in action: An approach for combing action understanding and object perception. In: CVPR (2007)
12. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In: CVPR (2009)
13. Hakeem, A., Shah, M.: Learning, detection and representation of multi-agent events in videos. Artificial Intelligence 171(8), 586–605 (2007)
14. Jain, A., Gupta, A., Davis, L.S.: Learning what and how of contextual models for scene labeling. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 199–212. Springer, Heidelberg (2010)
15. Khamis, S., Morariu, V.I., Davis, L.S.: A flow model for joint action recognition and identity maintenance. In: CVPR (2012)
16. Lan, T., Wang, Y., Mori, G., Robinovitch, S.N.: Retrieving actions in group contexts. In: Kutulakos, K.N. (ed.) ECCV 2010 Workshops, Part I. LNCS, vol. 6553, pp. 181–194. Springer, Heidelberg (2012)
17. Lan, T., Wang, Y., Wang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: NIPS (2010)
18. Lan, T., Wang, Y., Yang, W.L., Robinovitch, S.N., Mori, G.: Discriminative latent models for recognizeing contextual group activities. TPAMI 34(8), 1549–1562 (2012)
19. Liu, L., Ai, H.: Learning structure models with context information for visual tracking. Journal of Computer Science and Technology 28(5), 818–826 (2013)
20. Marszalek, M., Laptev, I., Shimid, C.: Actions in context. In: CVPR (2009)
21. Morariu, V.I., Davis, L.S.: Multi-agent event recognition in structured scenarios. In: CVPR (2011)
22. Odashima, S., Shimosaka, M., Kaneko, T., Fukui, R., Sato, T.: Collective activity localization with contextual spatial pyramid. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part III. LNCS, vol. 7585, pp. 243–252. Springer, Heidelberg (2012)
23. Xiang, T., Gong, S.: Beyond tracking: modeling activity and understanding behavior. IJCV 67(1), 21–51 (2006)
24. Xing, J., Liu, L., Ai, H.: Background subtraction through multiple life span modeling. In: ICIP (2011)