Learning Person–Person Interaction in Collective Activity Recognition

Xiaobin Chang, Wei-Shi Zheng, and Jianguo Zhang

***Abstract*— Collective activity is a collection of atomic activities (individual person's activity) and can hardly be distinguished by an atomic activity in isolation. The interactions among people are important cues for recognizing collective activity. In this paper, we concentrate on modeling the person–person interactions for collective activity recognition. Rather than relying on hand-craft description of the person–person interaction, we propose a novel learning-based approach that is capable of computing the class-specific person–person interaction patterns. In particular, we model each class of collective activity by an interaction matrix, which is designed to measure the connection between any pair of atomic activities in a collective activity instance. We then formulate an interaction response (IR) model by assembling all these measurements and make the IR class specific and distinct from each other. A multitask IR is further proposed to jointly learn different person–person interaction patterns simultaneously in order to learn the relation between different person–person interactions and keep more distinct activity-specific factor for each interaction at the same time. Our model is able to exploit discriminative low-rank representation of person–person interaction. Experimental results on two challenging data sets demonstrate our proposed model is comparable with the state-of-the-art models and show that learning person–person interactions plays a critical role in collective activity recognition.**

***Index Terms*— Collective activity recognition, interaction modeling, action analysis.**

I. Introduction

COLLECTIVE activity recognition in computer vision has received increasing attentions in recent years. Beyond the actions performed by individuals, collective activity is a collective behavior of a group of people in a scene, where the interactions between people are important features. Therefore, collective activity can hardly be identified by the action of individual person in isolation. For example, Fig. 1 shows the ambiguous role of the action performed by each single person in identifying different collective activities. As illustrated in Figs. 1(*a*) and 1(*b*), there are obvious interaction patterns existed in the collective activities, i.e., the scenario where two people are standing still and facing to each other might indicate that they are *talking* to each other, whilst the scenario where those are standing still and facing to the *same* direction provides a strong cue for the presence of a *queuing* collective activity. However, with the atomic action "stand still" only, it is impossible to distinguish the two collective activities. This example shows the incapability and limitation of using actions performed by individuals alone for collective activity recognition without considering person-person interaction patterns, as collective activities are often sharing the same or similar atomic actions such as standing still shown in Fig. 1. In this paper, we call the action of a single person and its influence to the surrounding people as the *atomic* activity. Similar terminology was also adopted in some recent literature [11], [12].

The research challenges and focus on collective activity recognition should differ significantly from the widely studied action recognition [21], [28], [31], [32], [35], [37], [41]–[43], where the actions performed by individuals are the main focus. It should also be distinguished from the crowd activity recognition [19], [27], [29], [30] in a way that collective activity is not to model a crowd scenario but rather to infer collective person-person interactions between several people. In comparison, the focus of crowd activity recognition is mainly on discovering regular and common moving patterns in a large public scene often containing a crowd of more than tens or hundreds of people or objects in a single view such as in a train station. What's more, the people in the crowd usually severely occluded by each others. In contrast, the people in collective activity are less occluded and the action of each person can be much more clearly observed. In addition, the number of people in collective activities is usually much less than that in crowd.

Existing methods on modeling collective activities have considered different types of interactions, mainly including: subject-time interaction [26], group-person interaction [24], person-object interaction [4], [5], and person-person interaction [9], [18], [22]. However, in most of the existing work, the interaction, especially the person-person interaction,

Manuscript received October 19, 2014; revised February 1, 2015; accepted February 15, 2015. Date of publication March 6, 2015; date of current version March 27, 2015. This work was supported in part by the National Natural Science of Foundation of China under Grant 61472456, Grant 61173084, and Grant U1135001, in part by the 12th Five-year Plan China Science and Technology Supporting Programme under Grant 2012BAK16B06, in part by the Guangzhou Pearl River Science and Technology Rising Star Project under Grant 2013J2200068, in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014265, and in part by the RSE-NSFC Joint Project under Grant 443570/NNS/INT. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ling Shao. *(Corresponding author: Wei-Shi Zheng.)*

X. Chang is with the School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510275, China, and also with the SYSU-CMU Shunde International Joint Research Institute, Shunde 510006, China (e-mail: littlesoliderchang@gmail.com).

W.-S. Zheng is with the School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510275, China, and also with the Guangdong Provincial Key Laboratory of Computational Science, Guangzhou 510275, China (e-mail: wszheng@ieee.org).

J. Zhang is with the School of Computing, University of Dundee, Dundee DD1 4HN, U.K. (e-mail: jgzhang@computing.dundee.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2409564

(a) (b) (c)

Fig. 1. The collective activity in (a) is talking, the collective activity in (b) is queuing, and the collective activity in (c) is running. One person in each picture is boxed, highlighting their individual actions and facing directions. All these boxed people are all facing left; the first two people are standing still, while the third one is running. The first two pictures demonstrate that the person-person interactions among the people are critical to collective activity recognition since we cannot distinguish their collective activity just from their atomic activities in isolation. What's more, the atomic activity also provides strong cues for collective activity, which can be shown by comparing (c) with (a) and (b).

(a) (b)

Fig. 2. The global collective activity in (a) is queuing, but the two people in the bounding boxes are talking. The global collective activity in (b) is walking, but the people in the bounding box are talking. Hence, one should predict the collective activity based on *collective* information from all pairs of person-person interactions rather than local ones in isolation.

is usually described by hand-crafted features, and the interaction descriptors are often not learned and quantified automatically for discrimination. Therefore, their discrimination powers are usually not guaranteed and generalized well across datasets. The intrinsic and discriminative person-person interaction patterns may not be well exploited.

We have also observed that the current trend of tackling the problem of collective activity recognition is to develop a model or framework with *increasing complexity* to jointly learn more subtasks simultaneously (e.g., detection, tracking, pose estimation, appearance modeling, and interaction) [11]. Despite some plausible aspects of this trend (e.g., a nice joint learning framework to tackle tasks simultaneously), the potential pitfalls within this regime will arguably span a couple of aspects including: 1) more complex models tend to be more difficult to optimize; 2) each subtask is not sufficiently explored and studied in depth; 3) it will be difficult to provide insights on which component is critical to solve the problem in order to guide future research efforts.

In this work, we take a different perspective and focus on one particular task of automatically learning person-person interactions. More specifically, we derive a learning-based approach to automatically mine the intrinsic person-person interaction patterns between atomic activities. In particular, our model assumes that two atomic activities in a collective activity are connected. In most cases, two connected atomic activities in one collective activity are either 1) quite similar and spatially close to each other to form a meaningful collective activity (e.g. two people are walking together); or 2) not quite similar but are strongly interacting to each other when participating a collective activity (e.g. facing each other when two people are talking, or fighting). In order to learn the connection, we propose to formulate such a connection into the form of a generalized inner product. To describe the collective information of the atomic activity in a clip, all of the pairwise connection potentials within an instance of collective activity are then aggregated together (Eq.(1)) to generate a final *response* score for further prediction. We call our model as an *interaction response* (IR) model. It is worth noting that different activities might be potentially related by sharing some atomic activities as shown in Fig. 1. In order to learn a more robust discriminative class-specific model, we further develop a *multi-task interaction response* (MIR) model, which jointly learns different class-specific interaction models by bridging them with a shared components.

In addition, we developed an optimization algorithm by re-formulating the model using a low-rank matrix factorization. In order to guarantee learning a suitable low-rank subspace, we impose a $-LogDet$ penalty to constrain the volume of the kernel matrix. We also call this model as the $-LogDet$ multiple low-rank subspace interaction model.

In summary, the main *contributions* of our work are:

1) We directly learn the person-person interaction between any two atomic activities in a collective activity. Such a kind of interaction is considered as the correlation (in the form of a generalized inner product) between the feature representations of two atomic activities, and an interaction response (IR) model and an extension called multi-task interaction response (MIR) are developed;
2) A *jointly learned* and *class-specific* $-LogDet$ based interaction modeling is proposed, with an attempt

TABLE I

COMPARISONS ON THE CHARACTERISTICS AND CAPABILITIES OF DIFFERENT MODELS. ✓ MEANS THE MODEL IS BASED ON SUCH INFORMATION OR HAS SUCH CAPABILITY. △ MEANS NOT ALL THE MODELS IN THE FIRST COLUMN OF THE SAME ROW USE SUCH INFORMATION OR HAVE SUCH A CAPABILITY

Models	Characteristics and Capabilities of Existing Models									
	Person's Facing Direction	Atomic Activity	Collective Activity of each person	Person-Person Interaction	Group-Person Interaction	Object-Person Interaction	Subject-Time Interaction	Spatial distribution	Video Feature	Notes
Lan's Model [23], [24]	✓		✓	✓	✓			✓		Interaction between people is outcome
Choi's Model [10], [11]	✓	✓		✓				✓		Multi-tracking + Collective Activity Recognize
AND-OR Graph Model [4], [5]	✓	✓			✓	✓				
Temporal Interaction Features [26]							✓			
Spatial Distribution [12], [13]	✓							✓		
Video Feature [3], [6], [33]			△						✓	
Person-person feature [9], [18], [22]	✓	△		✓	△			△		Modeling person-person interaction as feature
Our IR Model	✓			✓				✓		interaction patterns formed by interaction matrices

to better distinguish different types of collective activities.

3) We show that without the joint learning with other tasks (e.g. individual's action recognition and tracklet estimation), a purely learned interaction model between atomic activities outperforms the state-of-the-art in most cases or achieves comparable results.

The rest of this paper is organized as follows: Section II reviews the related work on collective activity recognition. Section III introduces the proposed interaction model, the optimization algorithm and inference strategy. Section IV extends the model into a multitask learning framework. Section V describes in details the implementation and provides an in-depth analysis of the person-person interaction properties learned by the proposed method. Section VI presents the experimental results and Section VII concludes the whole paper.

II. RELATED WORK

Since a collective activity consists of multiple atomic activities, the information from atomic activity in isolation is not sufficient to characterize the whole activity. In recent years, different models have been proposed to explore additional cues such as contextual information and interaction. In Table I, we summarize the characteristics of most of the existing methods in collective activity recognition.

Early work on collective activity recognition focused on contextual learning [12], [13], where the collective activities are described by the spatial distributions of the atomic activities and classified by random forest trees based on the spatial distributions. There are also some works focus on extracting video features for collective activity recognition. Todorovic [33] formulated the representation of a collective activity into a video graph (*kronecker graphs* [25]), where the extracted video features form the nodes and the relations between the features are represented by the edges. The work in [3] detected the video parts where the collective activities occur and made use of these local visual cues in the detected parts for recognition. Recently, Antic et al. [6] proposed to automatically learn activity constituents that are meaningful for collective activity recognition from video. Instead of detecting the video parts from the whole video as in [3], this work focused on the semi-local characteristics and the interrelation between different persons. The trajectory-based model presented in [27] totally relied on the trajectories of people and extracted the trajectory feature for collective activity recognition.

The holistic representation of contextual information though proven effective, may have limited descriptive power due to the diversity of interactions. Therefore, different types of interactions are specifically considered and mined. The group-person interaction was modeled in [23] and [24] by a latent SVM [40], with the focus on the patterns between the collective activity and the atomic activity. Later on, an efficient optimization algorithm of this model is proposed in [38]. Although the model by Lan et al. [23], [24] is able to tell

whether any two people are connected in the collective activity according to the latent variables in the modeling, it does not seek to characterize and quantify the interaction patterns in person-person interaction for collective activity recognition.

In addition to interactions between people, interactions between people and object are also explored in some of the existing work. Mohamed et al. focused on the object-person interaction and group-person interaction [4], [5]. A three-layer hierarchical model (AND-OR graph) was developed to associate objects, people and collective activity together, where collective activity occupies the highest level, the person activity comes the second and the objects locate at the lowest level. Different interactions are modeled as the connections between levels. For example, the connection between the top two layers represents the group-person interaction. However, this model requires a multitude of detectors at different levels, resulting in a time-consuming inference operation. The Monte Carlo tree search in [4] is therefore proposed to tackle this problem.

The motion pattern in collective activity was studied by Li et al. [26] and a compact and discriminative descriptor was proposed to characterize the subject-time interaction, which depends on many factors, such as the trajectories of people and the associated atomic activities. The corresponding feature is computed as a temporal interaction matrix (TIM) followed by a discriminative temporal interaction matrix (DTIM), which describes the properties of the subject-time interactions among multiple subjects in a collective activity. In recent developments, multiple people tracking and collective activity recognition are simultaneously considered in one framework [10], [11], with a hierarchy of three different levels representing collective activities, interactive activities and atomic activities respectively. The interactive activities can be considered as person-person interactions.

Recently, some work [9], [18], [22] started to consider person-person interaction for activity recognition. The interactive phrase, a latent mid-level feature, was proposed by Kong et al. [22] to describe the person-person interaction, which captures the interaction patterns by exploiting motion relationships between body parts. Tran et al. [18] modeled people interactions using an undirected graph, where each person is treated as a vertex and the person-person interaction is represented by the edge. A descriptor is then created to capture the motions and interactions of people within the graph. Finally, a bag-of-word approach is used to represent group activity. Cheng et al. [9] proposed features that make use of not only the person-person interaction but also the interactions at group level with the spatial distribution of group being encoded as features.

In summary, existing methods have considered different types of interactions. However, they mainly depend on hand-crafted features and do not explicitly exploit class-specific and learning-based technique to automatically mine person-person interaction features. Our method *differs* significantly from them and *learns* the person-person interaction between atomic activities via an interaction response model, which is capable of computing a set of class-specific and low-rank interaction features. Although one can infer whether two persons are linked in a collective activity as in [23] and [24], the person-person interaction features are not quantified directly.

III. Modeling Person-Person Interactions in Collective Activity

A. Atomic Activity Modeling

For each collective activity, the first task is to detect and track all the people in each video clip, which could be achieved by leveraging existing methods in [7], [16], [20]. Suppose N_q people are detected and tracked in a video clip V_q. Let $P_{1,q}, P_{2,q}, ..., P_{N_q,q} (N_q \geqslant 2)$ denote the N_q people. We then extract a low level feature $f_{n,q} \in R^\ell$ to describe the atomic activity associated to each person $P_{n,q}$, $n = 1, ..., N_q$. Since the *atomic activity* associated to person $P_{n,q}$ consists of its action and its influence [12], $f_{n,q}$ is composed of two types of features: the motion-based features and spatial distribution of the other people around this person.

B. Interaction Modeling for Atomic Activities

Suppose there are M collective activities and we denote its label set as γ. In order to directly model the interaction between atomic activities associated to any two people (i.e. person-person interaction) in the video clip V_q, our collective activity interaction response $R_{m,q}$, is defined as follows:

$$R_{m,q} = \sum_{i,j=1,\ i\neq j}^{N_q} f'_{i,q} \Omega_m f_{j,q}, \tag{1}$$

where m indicates one specified collective activity class, and $f'_{i,q}$ is the transpose of $f_{i,q}$. The person-person interaction pattern for class m is captured by the matrix $\Omega_m \in R^{\ell\times\ell}$, which is called the *interaction matrix*. Then the generalized inner product $f'_{i,q}\Omega_m f_{j,q}$ measures the connection between the two atomic activities associated to person i and person j acting for collective activity class m in video clip V_q. We would like to sum the effects of all person-person interactions in the video clip in order to consider the global collective activity. This can be further illustrated in Fig.2. Therefore, $R_{m,q}$ is the response that measures the contributions of all the person-person interactions in the context of collective activity class m in the video clip V_q. We call the model (Eq. (1)) the interaction response (IR) model.

In this work, we assume there is only one collective activity instance in each video clip and expect that if this class-specific person-person interaction is appearing in the m^{th} activity class, $R_{m,q}$ should output a higher score, otherwise a small value. Consequently, the inference of the collective activity class of a video clip V_q's can be casted as the following optimization problem:

$$\hat{l}_q = \arg\max_{m\in\gamma} R_{m,q}, \tag{2}$$

where γ is the set of all possible activity class labels and $\hat{l}_q$ is the predicted collective activity class label of V_q. It is obvious that for a given video clip V_q, the prediction of its class label depends only on the person-person interaction responses $R_{m,q}$, which are computed by the interaction matrix Ω_m.

C. Learning the Interaction Matrix Ω_m

Given the training set T and the t^{th} training sample $I_t \in T$ with class label $l_t \in \gamma$, we would like to train our model to learn the parameters $\mathbf{\Omega} = \{\Omega_1, \Omega_2, ..., \Omega_{|\gamma|}\}$, where γ is the set of all collective activities class labels. In order to learn an interaction matrix Ω_m for activity class m, $\forall m \in \gamma$, the training set T is divided into two subsets, where all of the training instances in class m are considered as a positive class, and the rest as a 'negative' class (similar to the one-vs-rest setting). For simplicity, we introduce a division notation y_t^m for each training instance I_t, where $y_t^m = 1$ if the training instance I_t is labeled with class m and $y_t^m = -1$ if not. Then, the positive index set is $\mathcal{P}_m = \{t | y_t^m = +1,\ l_t = m\}$ and the negative index set is $\mathcal{N}_m = \{t | y_t^m = -1, l_t \neq m\}$.

For any video clip V_q from the positive set, its interaction response $R_{m,q}$ to the positive class should be larger than its response to the negative set. As a result, we can achieve this by finding the interaction matrix Ω_m that makes the interaction response to be positive on the positive training subset (where $y_t^m = +1$) and interaction response to be negative on the negative subset (where $y_t^m = -1$). Therefore, to learn an optimal interaction matrix Ω_m, we formulate a loss function as follows:

$$\sum_{t=1}^{T} max(0, 1 - y_t^m (\sum_{i,j=1, i\neq j}^{N_t} f_{i,t}' \Omega_m f_{j,t}))^2 \tag{3}$$

We would like the value of $y_t^m (\sum_{i,j=1,i\neq j}^{N_t} f_{i,t}' \Omega_m f_{j,t})$ to be as large as possible in order to achieve a minimum loss. To reduce the risk of over-fitting, we add the regularization term, namely minimizing the Frobenius norm of the Ω_m, into the loss function above (in a similar fashion to L2-SVM). More importantly, we want to learn a low-rank matrix Ω_m such that it can implicitly quantify the person-person interaction in a low-dimensional space. This is achieved by adding a low-rank constraint on the matrix to the cost function. Therefore, we have the following objective function:

$$\min_{\Omega_m} \frac{1}{2}\|\Omega_m\|_F^2 + C\sum_{t=1}^{T} max(0, 1 - y_t^m (\sum_{i,j=1,i\neq j}^{N_t} f_{i,t}' \Omega_m f_{j,t}))^2$$
$$s.t\ rank(\Omega_m) < \upsilon \tag{4}$$

where $C > 0$ and υ is the constant to control the rank of interaction matrix Ω_m. In the experiments, we demonstrate that a suitable low rank interaction matrix would lead to better result.

However, the object function in Formula (4) is not easy to optimize due to the low rank constraint. To solve this problem, we develop a simplified model by assuming Ω_m being a symmetric matrix with the following factorization:

$$\Omega_m = L_m * L_m' \tag{5}$$

where $L_m \in R^{\ell \times d}$ and usually $d \ll \ell$. L_m' is the transpose of L_m. Therefore, the interaction response in Eq.(1) equals to

$$R_{m,q} = \sum_{i,j=1,\ i\neq j}^{N_q} f_{i,q}' L_m L_m' f_{j,q}, \tag{6}$$

It is worth noting that due to the matrix factorization, $rank(\Omega_m) = rank(L_m * L_m') = rank(L_m) \leq d$. This means, $rank(\Omega_m)$ can be controlled by d (the number of columns of L_m) as an upper bound. The constraint $rank(\Omega_m) < \upsilon$ in Formula (4) becomes redundant and can be removed. Consequently, Formula (4) can be converted into the following unconstrained optimization problem:

$$\min_{L_m} \frac{1}{2}\|L_m\|_F^2 + C\sum_{t=1}^{T} max(0, 1 - y_t^m \times (\sum_{i,j=1,i\neq j}^{N_t} f_{i,t}' L_m L_m' f_{j,t}))^2 \tag{7}$$

There are two main advantages of decomposing Ω_m into L_m:

1) We derive a low-rank representation of atomic activity associated to each person, as a new representation $r_{i,t,m} = L_m' f_{i,t}$ of the atomic activity for each person can be computed by projecting $f_{i,t}$ onto the column space spanned by L_m once L_m is learned. More importantly, $r_{i,t,m}$ is a function of m, which means that, for different activity class, the same atomic activity shared by different classes could be projected and represented differently. Only the one corresponding to the true collective activity class will give the prominent interaction response to that class.
2) From the optimization point of view, much less parameters are required to optimized, as $d \ll \ell$ and the number of parameters in L_m is much smaller than that of Ω_m. For example, If $d = 64$ and $\ell = 500$, only 32,000 parameters need to be learned in L_m, significantly less than 250,000 parameters in the original Ω_m with a ratio of only 12.8%. Hence, using a *low-rank* representation can be considered as a regularization model to avoid over-fitting as much less parameters need to be optimized. Our experimental results confirm this and suggest a suitable d is preferred.

A LogDet Regularization: It is important to note that $rank(L_m) \leq d$ is implicitly a loose constraint in Formula (7), as we have no tight control of the rank of L_m. If $rank(L_m) = r \leq d$, there are only r independent columns. Thus, a number of columns $(d - r)$ can be written as a linear combination of others, and become *redundant*. This will degrade the learning power of L_m. We term this as a *redundancy problem*. Therefore we prefer an ideal L_m with $rank(L_m) = d$, to ensure that all of the columns in L_m are as linearly independent as possible in the column space.

In order to solve the redundancy problem, we introduce a $-LogDet$ regularization term, $-log\ det(L_m' L_m)$, into the objective function in Formula (7), where det is the determinant operator. An intuitive motivation is that the value of $-log\ det(L_m' L_m)$ becomes $+\infty$ and consequently increases the cost of the objective function in Formula (7) if L_m is not of full column rank. Moreover, a larger $det(L_m' L_m)$ means more diversity of columns in L_m, which implies more discriminative information can be preserved in L_m. Hence, the proposed

objective function becomes:

$$\min_{L_m} J(L_m) = \min_{L_m} \frac{1}{2}\|L_m\|_F^2 - \frac{\beta}{2} log\ det(L_m' L_m) + C\sum_{t=1}^{|T|} \max(0, 1 - y_t^m(\sum_{i,j=1}^{N_t} f_{i,t}' L_m L_m' f_{j,t}))^2 \quad (8)$$

where $\beta \geqslant 0$, $C \geqslant 0$.

D. Optimization Algorithm

We use the Gradient Descent to solve the unconstrained optimization in Formula (8). The gradient of the term $-log\ det(L_m' L_m)$ is calculated as follows:

$$\frac{\partial - log\ det(L_m' L_m)}{\partial L_m} = -2(L_m^+)',$$

where L_m^+ is the pseudo-inverse of L_m. Then the gradient, $\nabla J(L_m)$, of the whole objective function $J(L_m)$, could be obtained as follows

$$\begin{aligned}
&\nabla J(L_m) \\
&= L_m - \beta(L_m^+)' \\
&\quad + C\frac{\partial(\sum_{t=1}^{|T|}(\max(0, 1 - y_t^m(\sum_{i,j=1, i\neq j}^{N_t} f_{i,t}' L_m L_m' f_{j,t})))^2)}{\partial L_m} \\
&= L_m - \beta(L_m^+)' \\
&\quad -2y_t^m C\sum_{t=1}^{|T|}\Big\{\max(0, 1 - y_t^m(\sum_{i,j=1, i\neq j}^{N_t} f_{i,t}' L_m L_m' f_{j,t})) \\
&\qquad \times [\sum_{i,j=1, i\neq j}^{N_t} (f_{i,t} f_{j,t}' + f_{j,t} f_{i,t}')] L_m\Big\} \quad (9)
\end{aligned}$$

Iterative gradient descent method is then applied to optimize the objective function.

Complexity Analysis: The computational complexity lies in two aspects: training and testing. On learning, assume given a dataset of P training video clips with an average of N persons present in each video clip. Let $L_m \in R^{\ell \times d}$, where $d \ll \ell$. Then the complexity of our optimization algorithm for one iteration step is $O(P \times N^2 + \ell^2 \times d + d^3)$. Excluding the time for computing the features and fixing $d = 384$, one round training of our algorithm takes about 8.39 seconds per training sample on a machine of 12-core with 256G memory. Note that the training stage can be performed offline and thus it would not affect its practical use too much. Our testing is pretty fast once the interaction matrix Ω_m is learned. It takes about 0.02 second for a test sample. In contrast, it is noted the method in [4] took more than 100 seconds on inference per frame, mainly due to its expensive searching procedures.

IV. A Multi-Task Extension

A. Motivation

It was observed that different classes of collective activities could share some common aspects, i.e., different collective activities may have the same type of atomic activities, or different collective activities are performed by the same group of people. For example, the collective activities of chasing and gathering could share the common element of walking, though facing to different directions. In addition, the collective activities could further share the element of people's spatial distribution. In order to better find out the discriminative information in each collective activity, we exploit the idea of multi-task learning [15], which is designed to tackle different but related learning tasks in one framework, aiming to give better performance. In our modeling, we treat learning each collective activity's interaction matrix as a task, and a multi-task interaction response (MIR) model is proposed by introducing a shared component among interaction matrices. The MIR model jointly learns all the interaction matrices of different collective activities and the shared component. As a result, the learned interaction matrix for each collective activity can preserve more distinctive information for the corresponding collective activity. In comparison, the single task interaction response model (Eq. (8)) is optimized with respect to different classes individually, i.e., class by class.

B. Multi-Task Extension

Specifically, we represent each $\Omega_m, m = 1, ..., \gamma$ by $\Omega_m = (1 - \alpha)\overline{\Omega}_0 + \alpha\overline{\Omega}_m$, where $\alpha \in [0, 1]$. On the one hand, the $\overline{\Omega}_0$ can be seen as the shared component of different Ω_m. On the other hand, the class-specific discriminative person-person interaction information among different collective activities can be preserved in the corresponding $\overline{\Omega}_m$. α is a parameter, which gives the freedom to control the trade-off between the shared component $\overline{\Omega}_0$ and $\overline{\Omega}_m$ in Ω_m.

In order to learn $\overline{\Omega}_0$ and $\overline{\Omega}_m$, we also impose a matrix factorization on each of them. Assuming $\overline{\Omega}_0 \in S^{\ell \times \ell}$ and $\overline{\Omega}_m \in S^{\ell \times \ell}$, then we have $\overline{\Omega}_0 = \overline{L}_0 * \overline{L}_0'$, where $\overline{L}_0 \in R^{\ell \times d}$ and $\overline{\Omega}_m = \overline{L}_m * \overline{L}_m'$, where $\overline{L}_m \in R^{\ell \times d}$. The multi-task learning objective function becomes:

$$\begin{aligned}
&\min_{\overline{L}_0, \overline{L}_m} J(\overline{L}_0, \overline{L}_m) \\
&\quad = \frac{1}{2}\sum_{k=0,m}\{\|\overline{L}_k\|_F^2 - \frac{\beta}{2} log\ det(\overline{L}_k' \overline{L}_k)\} \\
&\qquad + C\sum_{t=1}^{|T|} Loss(\overline{L}_0, \overline{L}_m) \quad (10)
\end{aligned}$$

where $m = 1, \ldots, \gamma$, $\beta \geqslant 0$, $C \geqslant 0$, $\alpha \in [0, 1]$ and the loss function $Loss(\overline{L}_0, \overline{L}_m)$ is defined as:

$$\begin{aligned}
&Loss(\overline{L}_0, \overline{L}_m) \\
&= \max(0, 1 - y_t^m(\sum_{i,j=1}^{N_t} f_{i,t}'((1 - \alpha)\overline{L}_0 \overline{L}_0' + \alpha \overline{L}_m \overline{L}_m') f_{j,t}))^2
\end{aligned} \quad (11)$$

We also make use of the $-logdet$ term in Formula (10) in order to guarantee the full column rank of $\overline{L}_0$ and different $\overline{L}_m$s. Formula (10) becomes equivalent to Formula (8) if $\alpha = 1$. We develop an alternating optimization

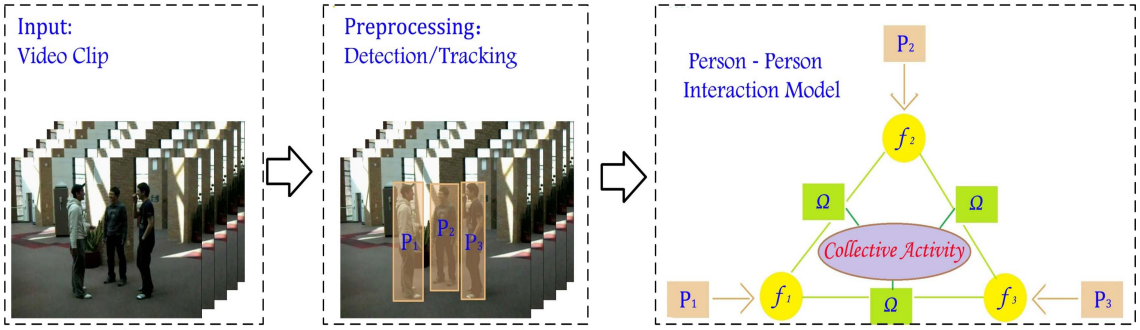

Fig. 3. The flow chart of our model. We focus on modeling the person-person interaction in collective activity recognition. We exploit the features of person (f_i) in pairs and model their relation as a generalized inner product $f_i'\Omega f_j$. The interaction pattern is then captured by the interaction matrix Ω, which is learned by maximizing the interaction responses in a collective activity. Best viewed in color.

Algorithm 1 Optimization of MIR Model (Eq. (10))

1: Fix $\overline{L}_0$ and optimize the objective function in Formula (10) on different $\overline{L}_m$, $m = 1, \ldots, \gamma$ respectively by one-step gradient descent
2: Optimize $\overline{L}_0$ when fixing all the $\overline{L}_m$, $m = 1, \ldots, \gamma$. The objective function of $\overline{L}_0$ is the sum of all the $J(\overline{L}_0, \overline{L}_m)$, $m = 1, \ldots, \gamma$ in Formula (10) and becomes

$$\overline{J}(\overline{L}_0) = \sum_{m=1}^{\gamma} J(\overline{L}_0, \overline{L}_m) \tag{12}$$

3: Optimize $\overline{J}(\overline{L}_0)$ w.r.t. $\overline{L}_0$ by one-step gradient descent;
4: Repeat 1), 2) and 3) until maximum number of iterations or convergence criterion is met.

procedure to solving for $\overline{L}_0$ and different $\overline{L}_m$s and our algorithm is summarized in Algorithm 1.

V. Implementation Details & Model Analysis

In this section, we describe the implementation details and present an analysis of two important aspects of our model including 1) the characteristics of the learned interaction matrices; and 2) the projected representation of the atomic activity in the learned low-rank subspace.

A. Implementation Details

The flow chart of our model is shown in Fig. 3, where the features of people are modeled in pair, and the person-person interaction patterns in collective activity are captured by the interaction matrix Ω. During the preprocessing step, multiple people could be detected and tracked in the video clips by employing the state-of-art object detection [16] and tracking algorithms [7], [20]. In order to have a fair comparison with exsiting methods, in the experiment, we directly use the people detection and tracking results provided by two public datasets: the CAD dataset [1], and the Choi's dataset [2] (see Section VI-A for detailed description of the two datasets).

In the feature extraction stage, a low-level feature f_i of the ith person is extracted based on local motions and spatial distributions. Specifically, the motion-based features [36] are extracted from the video. We only make use of the motion-based features that are located in the bounding boxes surrounding the person to form the feature vectors of the corresponding person. We learn a code book by a k-means algorithm [34], [39] based on the motion-based features above. Hence, a bag-of-video-words (BoV) representation is generated for each person. We also extract the spatial distribution of people around each person by using the STL feature [12]. The STL feature in our model forms a 96D vector. Finally, we concatenate the STL feature with the BoV motion-based histogram to form a feature vector. What's more, we apply the principal component analysis (PCA) [17] on the feature vectors and the number of principal components is determined based on preserving 99% of the total variance (the resulting feature dimensonality is 833 on the Collective activity dataset (CAD) [1] and 603 on Choi's dataset [2]). Some low-level feature examples are shown in Fig. 5.

B. Model Analysis

Since our model is discriminative, different person-person interaction patterns should be learned for different classes of collective activities, i.e., the learned interaction matrix Ωs should be intrinsically different. In order to provide insights into the pattern of Ω, we visualize the learned matrix Ω as an image for each class of collective activity. Figs. 4(a)-(e) and Figs. 4(l)-(q) illustrate the learned matrix for different classes of activities in two different datasets. We can see the patterns in Ω are different from class to class, which confirms that our model is *class specific*.

The multi-task extension of our model focuses not only on the different person-person interaction patterns but also on the shared information among different classes of collective activities. As shown in Figs. 4(g)-(k) and 4(s)-(x), the interaction matrices $\overline{\Omega}_m$ learned by the multi-task extension are still different from class to class. The shared component $\overline{\Omega}_0$ is the same for all of the classes as shown in Fig. 4(f) and Fig. 4(r) for the two datasets [1], [2] respectively. This means that some related information in different collective activities is implicitly accommodated by $\overline{\Omega}_0$.

Based on the factorization and the low-rank constraint, our model actually learns a class-specific and low-dimensional representation of the atomic activity feature. The person-person interactions under different collective activities emphasize on different aspects of atomic activities. For example, the people

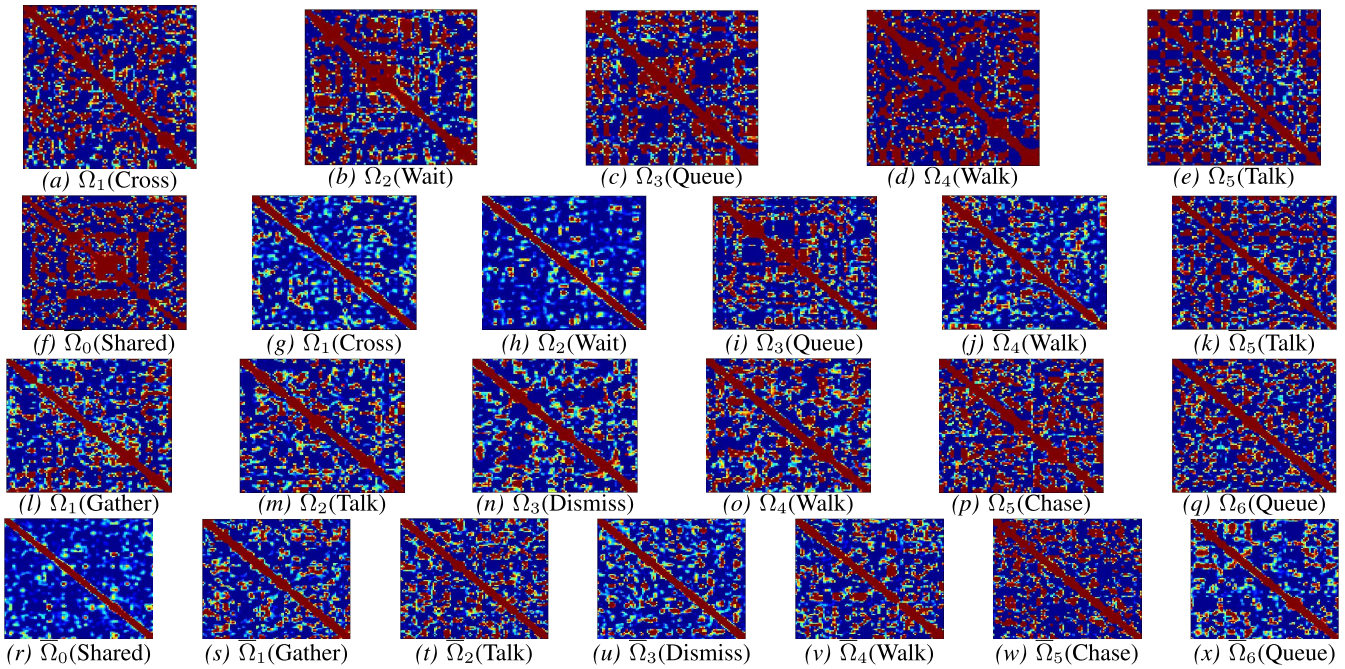

Fig. 4. These figures illustrate different interaction matrices of collective activities in dataset [1], [2] that are learned by our model and the multi-task extension. Since our model is class specific, each $\mathbf{\Omega}_m$ corresponds to collective activity m. Besides the class specific components $\overline{\mathbf{\Omega}}_m$s, the multi-task extension also learns a shared component $\overline{\mathbf{\Omega}}_0$ for all collective activities in the training set. Fig. 4(a) $\sim$ 4(e) depict the interaction matrices of our IR model learned from dataset [1]. The shared component $\overline{\mathbf{\Omega}}_0$ of the multi-task extension learned from dataset [1] is shown in Fig. 4(f) and different $\overline{\mathbf{\Omega}}_m$s for different classes are shown from Figs.4(g) $\sim$ 4(k). The learned interaction matrices of our IR model on dataset [2] are shown in Figs. 4(l) $\sim$ 4(q). The shared component $\overline{\mathbf{\Omega}}_0$ and different $\overline{\mathbf{\Omega}}_m$s of multi-task extension on dataset [2] are shown in Figs. 4(r) $\sim$ 4(x).

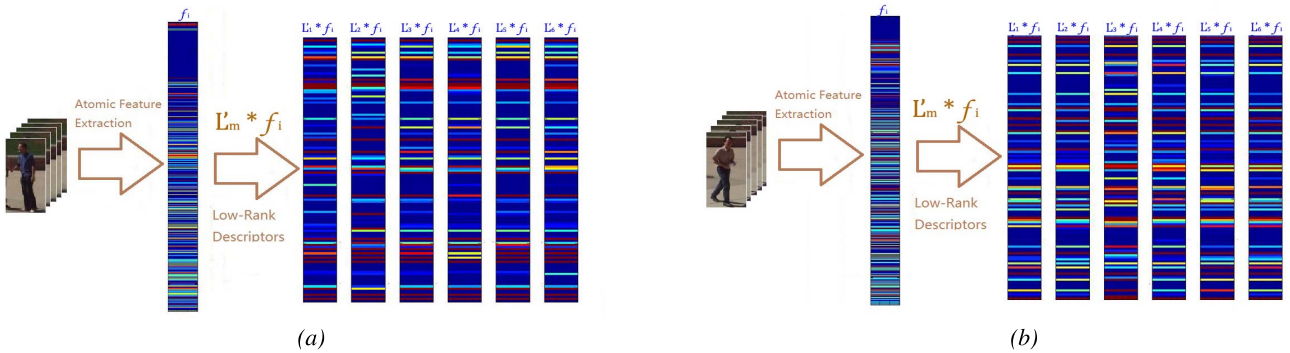

Fig. 5. The person shown in (a) is talking and shaking hands with others. The person shown in (b) is running. Their corresponding atomic activity features f_is are different from each other. The low-rank descriptors $r_{i,m}$ under six different collective activities m ($m = 1, 2, ..., 6$) are also shown for the Choi's dataset [2]. For each person, $r_{i,m}$ can be computed by projecting the atomic activity feature f_i onto each column of the matrix L_m. The resulting low-rank descriptors are different and class-specific, although they are computed from the same atomic activity feature (the same person). This confirms that our model is discriminative.

who are talking may have stronger upper body movements (e.g., shaking hands) than those queuing. The people who are running have their legs moving faster than those who are walking. This means that the features of the atomic activity performed by a person should be different under different collective activities. Let the ith person's atomic activity feature be f_i, then the corresponding *projected* low-rank descriptor $r_{i,m}$ under collective activity m is computed from $r_{i,m} = L'_m f_i, r_{i,m} \in R^d$. The resulting low-rank descriptors $r_{i,m}$ of f_i for different class m are illustrated in Fig. 5.

By comparing the atomic activity feature f_i and their corresponding projected low-rank descriptors $r_{i,m}$ in Figs. 5(a) and 5(b), we observe that the low-rank descriptors under different collective activities are notably different. Since different low-rank descriptors are computed by employing different L_m, the low-rank descriptors are *class specific*. In other words, different low-rank descriptors are capable of preserving the discriminative aspects of atomic activity projected from the raw atomic activity feature which could in return benefit our model.

C. Discussion

In the following, we discuss the relationships of the proposed model with relevant existing machine learning algorithms: Support Vector Machine (SVM) and Canonical Correlation Analysis (CCA).

1) Relation to SVM Model: The proposed interaction response (IR) model is a *class-specific* model and built based on the hinge-loss function. The class-specific modeling is similar to the one-vs-rest strategy used in SVM with the hinge-loss function. However, the main difference is that our interaction response model is to measure the *connection* between any two atomic activities and the class-specific formulation is to distinguish the pooled interaction response over all pairs of atomic activities in a video clip rather than just classifying each pair. This is in contrast with the *large margin* SVM which aims to find a decision boundary in the feature space (e.g., a holistic representation of a video) and does not seek to directly model person-person interactions. Therefore, traditional SVM is incapable of measuring the connection between any two atomic activities. In the experiment, we constructed a baseline method using SVM with a holistic representation, and compare its performance with ours.

2) Relation to CCA: In the proposed method, the interaction response (Eq.(6)) is computed via an inner product between two different atomic activities in a low-rank subspace. The idea of using inner product for correlation modeling is also employed in CCA [8]. However, CCA is hardly suitable to learn the person-person interaction in a discriminative way in our case and we explain below. Firstly, the main purpose of CCA model is to maximize the correlation between a pair of samples across modalities in a latent feature space (usually, the dimensionalities of the features from different sets are different in the case of CCA). However, the interaction learning we have considered is not aiming for matching of cross-modality features, but to explore the intrinsic relationship between two atomic activities in the same feature space. Secondly, CCA finds two universal projections for all pairs of cross-modality data points and this is not suitable for collective interaction modeling, because it is apparent that different person-person interactions should have their specific characteristics that are different from each other. Hence, our IR and MIR models learn class specific person-person interaction patterns. Thirdly, there are always a number of person-person interactions in a video frame rather than just a single one, and therefore the proposed IR and MIR models are actually considering the collective person-person interaction response rather than a single one, which is not considered in CCA.

VI. Experiments

In order to test the performance of our approach, we conduct a group of experiments and compare our model with the state-of-the-art methods as well as a baseline implementation on two public datasets including the collective activity dataset [1] and the Choi's dataset [2]. We further present a detailed analysis of the results and test the effects of different model parameters.

A. Datasets and Setting

Collective Activity Dataset (CAD) [1]: It contains 44 video clips labeled with 5 different collective activities (*crossing, waiting, queuing, walking and talking*). There are eight facing directions (right, right-front,..., right-back) of people presented in this dataset.

We choose the experimental setting used in [24], which splits one fourth of this dataset for testing and the rest for training. We have observed that with a limited number of splits, the averaged overall accuracies were not stable. Unfortunately, most of the existing methods did not clarify how many number of splits were used in their settings. To compensate this ambiguity, we tested our algorithm by increasing the number of splits until the averaged results having no significant change. We observed that the averaged results become stable when the number of splits is larger than 20, at which the results of IR and MIR are reported in Table II.

Choi's Dataset [2]: It consists of 32 video clips with 6 collective activities: gathering, talking, dismissal, walking together, chasing, and queuing. There are eight poses similar to the CAD dataset. We follow the standard experimental protocol of the 3-fold cross validation, suggested by Choi et al. [2]. This is a challenging dataset due to the large inter- and intra-class variations.

Baseline Method: An intuitive idea is to simply construct a holistic representation for each video clip and ignore the person-person interactions. Therefore, the popular bag-of-words representation with SVM is considered as baseline for further comparison, and its formulation for collective activities classification is described as follows:

$$\min_{\nu_m,\xi_t} \frac{1}{2} \parallel \nu_m \parallel_2^2 + C \sum_{t=1}^{|T|} \xi_t$$
$$s.t. \quad y_t^m (\nu_m' F_t) \geqslant 1 - \xi_t, \quad t = 1, 2, ..., |T| \qquad (13)$$

where ν_m is the model parameter of the collective activity m. F_t is the video feature of the tth training instance. In this work, F_t can be obtained by aggregating all the video-words histogram in the video.

B. Experiment on Collective Activity Dataset

Table II compares the results of the proposed method, its multi-task extension, the baseline method and other state-of-the-art approaches under the same experimental settings on the CAD dataset. It can be seen that our interaction response (IR) model outperforms the baseline method by a gain of 6.9%. Note that the baseline method does not take into account person-person interactions, and therefore it does not achieve a satisfactory classification result. This confirms the important role of person-person interaction in collective activity recognition. On the other hand, our approach namely the multi-task interaction response (MIR) model further improves the performance to 83.3% on the CAD dataset.

By comparing the results of the proposed model with the state-of-the-art shown in Table II, we can see that both the proposed model and its multi-task extension outperform the existing ones. Lan *et al.* [23] achieves 77.5%. It attempts to automatically figure out the interactions among the people form their atomic actions and models the person-person interactions as the latent variables. However, it did not seek to model the spatial relationship of people in their approach.

TABLE II
CLASSIFICATION ACCURACY (%) ON THE COLLECTIVE ACTIVITY DATASET (CAD) [1]

Class	Baseline	[24]	[23]	[26]	IR	MIR
Crossing	62.3	68.0	65.0	**77.0**	72.3	65.9
Waiting	55.5	69.0	60.0	63.0	76.3	**82.2**
Queuing	**98.6**	76.0	96.0	70.0	90.0	91.9
Walking	66.8	80.0	68.0	73.0	77.5	**81.4**
Talking	91.9	**99.0**	**99.0**	88.0	93.3	95.2
Average	75.0	78.4	77.5	74.2	81.9	**83.3**

TABLE III
CLASSIFICATION ACCURACY (%) ON CHOI'S DATASET [2]

Class	Baseline	[12]	[11]	[4]	IR	MIR
Gathering	**64.1**	50.0	43.5	48.1	55.2	59.9
Talking	96.5	72.2	82.2	81.3	94.3	**97.0**
Dismissal	76.4	49.2	77.0	55.3	**91.8**	90.5
Walking	90.4	83.2	87.4	89.1	93.4	**94.3**
Chasing	21.6	95.2	91.9	**95.9**	42.2	53.9
Queuing	78.7	95.9	93.4	**96.7**	84.3	86.3
Average	71.3	74.3	79.2	77.7	76.9	**80.3**

What's more, modeling the person-person interactions as latent varibales makes the problem hard to be optimized. As a result, these limit the performance of the model. Lan *et al.* [24] takes one step further and considers the interactions among people not only by modeling them as latent variables but also by capturing them in the feature level. This combined approach achieves a classification accuracy of 78.4%, which is slightly better than that in [23]. On this dataset under the same experimental setting, [11] and [26] also report their performances, with 75.7% and 74.2% respectively, which are inferior to ours.

C. Experiment on Choi's Dataset

The comparison of different models on Choi's dataset [2] is shown in Table III including: 1) the proposed model; 2) the proposed multi-task extension; 3) the method by Choi et al. [12]; 4) the method by Choi and Savarese [11]; 5) Amer et al. [4]; and 6) the baseline method.

From Table III, it can be observed that the multi-task extension outperforms the original model on most of the classes and improves the overall accuracy from 76.9% to 80.3%. The multi-task extension benefits from learning all tasks simultaneously instead of learning each class of collective activity independently as in our original model.

The baseline approach performs the worst. The multi-task extension of our model improves the performance and even slightly better than the state-of-the-art. The work in [11] achieves the closest results to our multi-task extension on this dataset (79.2% [11] vs. 80.3% (ours)).

Since the method by Choi's [11] achieves the closest performance to our multi-task extension, we present an analysis of these two models by examining their confusion matrix as shown in Fig.6. First, we can see that our multi-task extension model achieves better performance on recognizing 4 out of 6 collective activities than that of Choi and Savarese [11]. For example, the multi-task extension has an accuracy of 97% on recognizing the Talking activity, which is 15% higher than that of Choi and Savarese [11].

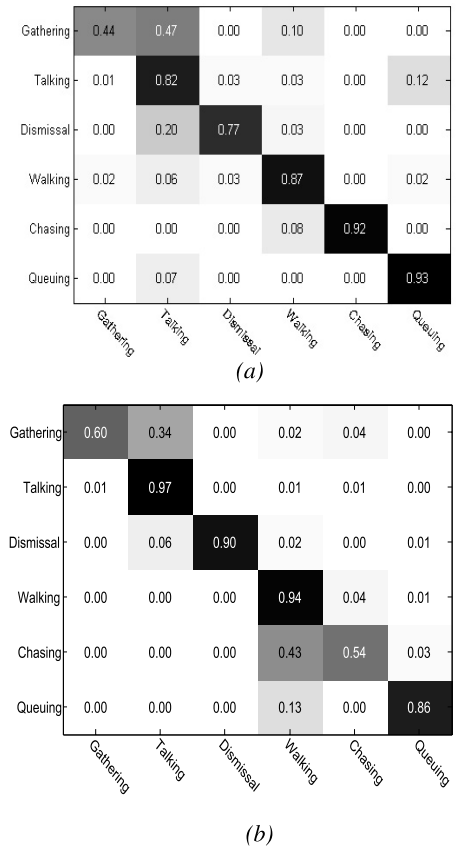

Fig. 6. Confusion matrices on Choi's dataset [2]: the proposed multi-task extension versus Choi's [11]. (a) Confusion matrix of [11]. (b) Confusion matrix of our multi-task method.

The performance gap of 15% is mainly a resultant from misclassifying the Talking to the Queuing activity by the method in [11]. It is worth noting that the atomic activities of people in these two collective activities are quite similar. They are all *standing still*. Therefore, they can be better distinguished by their interaction patterns than by the atomic activities alone. The scenario where people facing each others suggests a talking collective behavior, whilst the scenario where the people facing roughly the same direction indicates a queuing activity. Therefore, this suggests that our model is capable of extracting the distinctive person-person interaction patterns for different collective activities, even though the atomic activities in them are similar.

Moreover, by examining the confusion matrix as shown in Fig. 6, our multi-task extension performs significantly better than the method of Choi and Savarese [11] (13% margin) on discriminating the collective activity Gathering from Talking. The collective activity Gathering and Talking are hard to distinguish without a discriminative description of person-person interactions, since the person-person interactions are very similar (e.g. people may always face to each other) in those cases, which again confirms that our model is able to extract the class discriminative information from person-person interactions. Although the previous work (see [11], [12]) makes use of the person's facing directions to describe the atomic activity and the spatial distribution is also considered, they cannot distinguish the collective

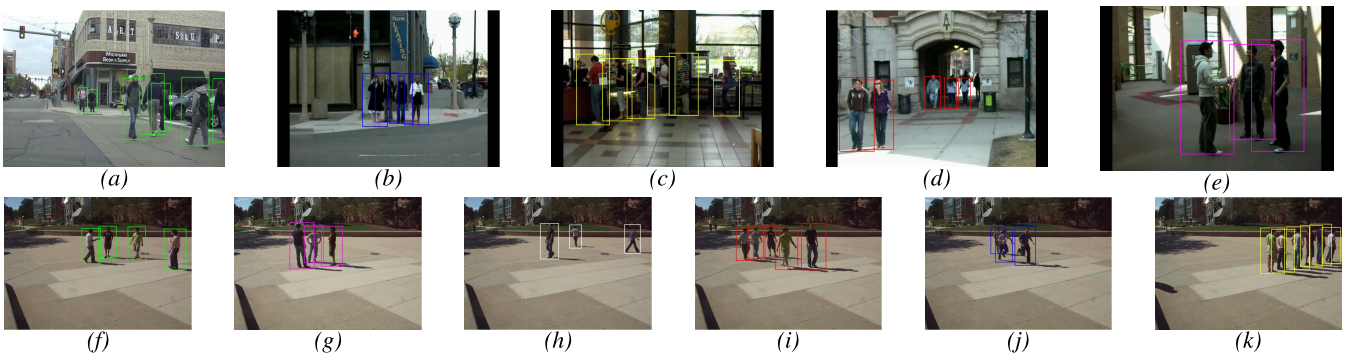

Fig. 7. Illustrate of correctly recognized samples. The first row shows the results on the CAD dataset [1]. The different colors of the bounding boxes represent the different collective activities: Crossing (Green), Waiting (Blue), Queuing (Yellow), Walking (Red), Talking (Magenta). The second row shows the results on the Choi's dataset [2]. Similarly, different colors represent different activities: Gathering (Green), Talking (Magenta), Dismissal (White), Walking (Red), Chasing (Blue), Queuing (Yellow). (a) Crossing. (b) Waiting. (c) Queuing. (d) Walking. (e) Talking. (f) Gathering. (g) Talking. (h) Dismissal. (i) Walking. (j) Chasing. (k) Queuing.

activity Talking and Gathering due to lack of a principled strategy to extract the discriminative information through learning. It is worth noting that our model performs much worse on the Chasing class compared to the method by Choi and Savarese [11]. This can be explained by several reasons. First of all, although the experimental results illustrate that person-person interactions play a central role in some collective activity situations, the person-person interactions may fail to be discriminative under some collective activity scenarios, such as walking and chasing. Both the atomic activities and person-person interactions in these two collective activities are very similar. Therefore, it is a challenging task for our model to distinguish these activities only based on the person-person interactions. On the contrary, [11] considers a unified framework on collective activity recognition. It captures not only the person-person interactions and atomic activities of each person but also other information, such as the tracklets of the people, which is very important in distinguishing collective activities like walking and chasing. What's more, the definition of some collective activities are not very clear. Some collective activities are hard to be distinguished one from the others (e.g. some activities in chasing looks very similar to walking). We present some further insights on this aspect by analyzing some mis-classified samples in next section.

However, it is worth noting that directly comparing different existing methods might not be possible and feasible, even under the same experimental settings because the baseline features of different methods are usually different. The approach in [12] was mainly based on the spatial distribution features, whereas Amer et al. [4] employed a diverse set of features to describe the collective activity on different levels. The method of Choi and Savarese [11] made use of the spatial distribution of people, which is similar to ours. The difference is that we employed the features for describing atomic activities developed by Wang et al. [36], while Choi and Savarese [11] used the spatial-temporal features (STF) developed by Dollár et al [14]. In order to have a fair comparison with [11], we switch to the spatial-temporal features developed in [14], which is the same as used in [11], and run another experiment on the Choi's dataset.

Fig. 8. Illustration of misclassified samples. The ground truth label of (a) is gathering. Our model predicts its label as talking. This example is used to illustrate the classification challenges that are caused by the transitions of collective activities. The ground truth label of (b) is chasing while we recognize it as walking. This example illustrates the classification challenges that are caused by the similarity between different collective activities.

Our multi-task extension model could achieve an accuracy of 82.0%, which is much better than the sate-of-the-art reported in [11] (79.2%) based on the same type of features.

D. Visual Analysis

Fig. 7 shows some examples that are correctly recognized by our model (one example per class) on the two datasets, and Fig. 8 shows some wrongly classified examples.

By examining the misclassified examples in Fig. 8, we present two main reasons to explain the cases that lead to the wrong prediction. Firstly, the transitions between different collective activities can cause wrong recognition. This is illustrated in Fig. 8(*a*). This example actually depicts the transition from a Gathering to a Talking activity. However, its ground truth is labeled as Gathering in the dataset. Our model recognizes it as Talking, because their talking interactions (hand shaking) are strong. Secondly, the differences between some collective activities are subtle and they share similar atomic activities and person-person interactions to some extent. Sometimes it is hard to distinguish them due to the strong similarity. For example, the main differences between Chasing and Walking are the moving speed and the gestures of the people. However, their differences become indistinguishable under some circumstances. For instance, we recognize the scenario in Fig. 8(*b*) as Walking because the atomic activities of the people look more like walking than

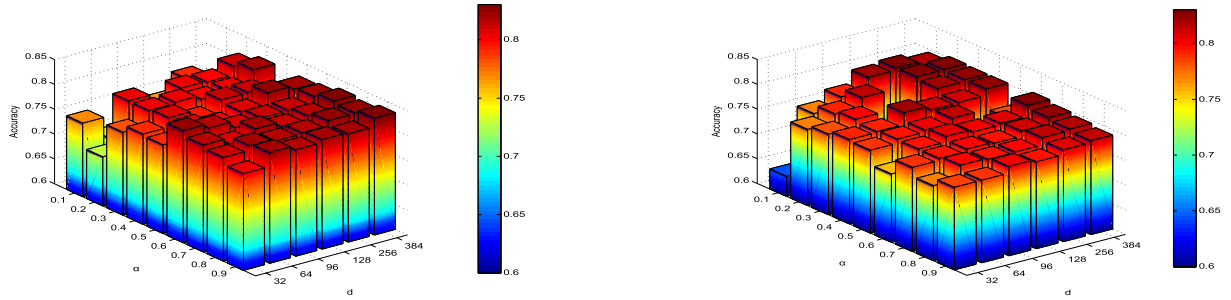

Fig. 9. Effects of parameters d and α on our model. When fixing the parameters $\beta = 0.3$ and $C = 20$, we vary the multi-task method parameter α from 0.1 to 0.9 with a step of 0.1. The parameter d picks the value of 32, 64, 96, 128, 256 and 384 respectively. Best viewed in color. (a) Collective Activity Dataset [1]. (b) Choi's Dataset [2].

running with nearly the same interaction (facing to the same directions), but the dataset defines its collective activity as Chasing because the people are walking at a high speed.

It is worth noting that our method performs worse on the Chasing class than the approach in [11] as shown in Fig. 6. A large portion of the error is due to the misclassification of the Chasing activity into the Walking activity, and we would like to point out that the Walking activity is more like an atomic activity rather than a collective activity, which therefore could potentially lead to the confusion with other collective activities (e.g., the Chasing activity with similar atomic activity).

E. Effects of Parameters

In this section, we carry out a set of experiments to test the effects of different parameters on the performance in the context of the multi-task extension including the trade-off parameter α and the number of columns, d, of the matrix L_m. We increase α from 0.1 to 0.9 by a step of 0.1, and the value d is chosen as 32, 64, 96, 128, 256, 384. Parameters β and C are set as 0.3 and 20 empirically. Fig. 9 shows the results against different values of α and d on the two datasets. From this figure, we can see that there is an optimal value of d which gives the best results on the two datasets. Either larger or smaller value may decrease the performance. On the one hand, a small value of d does not give much discriminative power. On the other hand, a larger value of d tends to cause the over-fitting problem.

The parameter α in the multi-task extension controls the trade-off between the shared component $\overline{\mathbf{\Omega}}_0$ and different $\overline{\mathbf{\Omega}}_m$s. The multi-task extension degenerates into the original model if $\alpha = 1$, i.e., independent learning. When α decreases, the corresponding multi-task extension algorithm focuses more on joint learning. However, the balance of independent learning and joint learning in different problems are different. Generally, we may conclude that the performance of our multi-task extension model tends to be stable as α increases. As shown in Fig. 9, the multi-task model achieves the best result with $\alpha = 0.9$ on the CAD dataset [1] while the best result on the Choi's dataset [2] is achieved with $\alpha = 0.2$. By cross-checking the results from Fig. 9, Table II and Table III, we can confirm the multi-task extension is capable of achieving better results than the original model ($\alpha = 1$).

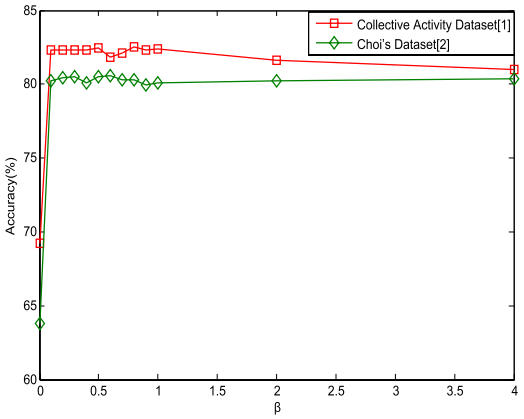

Fig. 10. Effects of parameter β on our model. When fixing the parameters $\alpha = 0.6$, $d = 384$ and $C = 20$, we vary the parameter β from 0 to 4. The red curve shows the results on Collective Activity Dataset [1] and the green curve shows the results on Choi's Dataset [2]. Best viewed in color.

We further test the effect of parameter β as shown in Fig.10. It can be observed that the lowest performances are achieved at $\beta = 0$, with accuracies of 69.3% on Collective Activity Dataset [1] and 63.8% on Choi's Dataset [2], respectively. When β increases, the performances start to increase rapidly until saturated. Bear in mind that $\beta = 0$ implies that LogDet regularization has no impact in the objective function. The introduction of LogDet regularization has a positive impact on the performance, mainly due to its capability of solving the redundancy problem. When β reaches a certain value of 0.3, the performances start to remain stable, which means our algorithm is insensitive to the value of β in a reasonable range. This trend is consistent on both datasets.

VII. Conclusion

In this paper, we focus on learning the person-person interaction and develop a discriminative interaction response (IR) model for collective activity recognition. The main characteristic of our modeling is to formulate the person-person interaction as the generalized inner products of two atomic activities and the discriminative person-person interaction patterns of different classes of collective activities are captured by different interaction matrices. By employing

the low-rank matrix factorization, the class-specific model also helps exploit low-rank representation of atomic activity, which can better describe the person-person interactions under certain collective activity. A multi-task formulation with an alternating optimization procedure is proposed, which boosts the performance by making use of the shared information among different collective activities. Our study shows that

1) Without jointly learning with tracking, detection, pose estimation and etc., an effective person-person interaction learning is also able to achieve state-of-the-art or comparable performance for collective activities recognition on two benchmarking datasets.
2) Learning the person-person interaction based on mid-level/raw features can mine more intrinsic relationship between atomic activities.

In this work, we consider all the person-person interactions that appear in the short video clip. Although the number of persons in a collective activity is rather small, compared to the large group size in the task of recognizing activities of crowd, the number of person-person interactions grows quadratically as the number of persons increases. But it will not affect the size of the interaction matrix, thus no increase of number of variables to estimate. However, this indeed affects the total number of summation that needs to be calculated in Eq. (1), which could be handled by distributed computing, e.g., parallel computing.

ACKNOWLEDGEMENTS

The authors would like to thank all reviewers' constructive comments on improving the manuscript.

REFERENCES

[1] *Collective Activity Dataset*. [Online]. Available: http://www.eecs.umich.edu/vision/activity-dataset.html, accessed Sep. 2013.
[2] *Dataset: A Unified Framework for Multi-Target Tracking and Collective Activity Recognition*. [Online]. Available: http://www-personal.umich.edu/~wgchoi/eccv12/wongun_eccv12.html, accessed Feb. 2014.
[3] M. R. Amer and S. Todorovic, "A chains model for localizing participants of group activities in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 786–793.
[4] M. R. Amer, S. Todorovic, A. Fern, and S.-C. Zhu, "Monte Carlo tree search for scheduling activity recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1353–1360.
[5] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu, "Cost-sensitive top-down/bottom-up inference for multiscale activity recognition," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 187–200.
[6] B. Antic and B. Ommer, "Learning latent constituents for recognition of group activities in video," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 33–47.
[7] S.-H. Bae and K.-J. Yoon, "Robust online multiobject tracking with data association and track management," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2820–2833, Jul. 2014.
[8] M. B. Blaschko, J. A. Shelton, A. Bartels, C. H. Lampert, and A. Gretton, "Semi-supervised kernel canonical correlation analysis with application to human fMRI," *Pattern Recognit. Lett.*, vol. 32, no. 11, pp. 1572–1583, 2011.
[9] Z. Cheng, L. Qin, Q. Huang, S. Yan, and Q. Tian, "Recognizing human group action by layered model with multiple cues," *Neurocomputing*, vol. 136, pp. 124–135, Jul. 2014.
[10] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *Proc. 12th Eur. Conf. Comput. Vis.*, vol. 4. 2012, pp. 215–230.
[11] W. Choi and S. Savarese, "Understanding collective activities of people from videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1242–1257, Jun. 2014.
[12] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops*, Sep./Oct. 2009, pp. 1282–1289.
[13] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3273–3280.
[14] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. 2nd Joint IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, Oct. 2005, pp. 65–72.
[15] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 109–117.
[16] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[17] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA, USA: Academic Press Professional, Inc., 1990, p. 592.
[18] K. N. Tran, A. Gala, I. A. Kakadiaris, and S. K. Shah, "Activity analysis in crowded environments using social cues for group discovery and human interaction modeling," *Pattern Recognit. Lett.*, vol. 44, pp. 49–57, Jul. 2014.
[19] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1003–1016, May 2012.
[20] A. Heili, A. López-Méndez, and J. M. Odobez, "Exploiting long-term connectivity and visual motion in CRF-based multi-person tracking," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3040–3056, Jul. 2014.
[21] J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong, and T. Xiang, "Exemplar-based recognition of human-object interactions," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[22] Y. Kong, Y. Jia, and Y. Fu, "Interactive phrases: Semantic descriptions for human interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 9, pp. 1775–1788, Sep. 2014.
[23] T. Lan, Y. Wang, W. Yang, and G. Mori, "Beyond actions: Discriminative models for contextual group activities," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran & Associates Inc., 2010, pp. 1216–1224.
[24] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1549–1562, Aug. 2012.
[25] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *J. Mach. Learn. Res.*, vol. 11, pp. 985–1042, Mar. 2010.
[26] R. Li, R. Chellappa, and S. K. Zhou, "Recognizing interactive group activities using temporal interaction matrices and their Riemannian statistics," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 305–328, 2013.
[27] W. Lin, H. Chu, J. Wu, B. Sheng, and Z. Chen, "A heat-map-based algorithm for recognizing group activities in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1980–1992, Nov. 2013.
[28] L. Liu, L. Shao, F. Zheng, and X. Li, "Realistic action recognition via sparsely-constructed Gaussian processes," *Pattern Recognit.*, vol. 47, no. 12, pp. 3819–3827, Dec. 2014.
[29] Y. Ma and P. Cisar, "Event detection using local binary pattern based dynamic textures," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2009, pp. 38–44.
[30] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven crowd analysis in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1235–1242.
[31] L. Shao, S. Jones, and X. Li, "Efficient search and localization of human actions in video databases," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 3, pp. 504–512, Mar. 2014.
[32] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal Laplacian pyramid coding for action recognition," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 817–827, Jun. 2014.
[33] S. Todorovic, "Human activities as stochastic Kronecker graphs," in *Proc. 12th Eur. Conf. Comput. Vis.*, vol. 2. 2012, pp. 130–143.
[34] A. Vedaldi and B. Fulkerson. (2008). *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. [Online]. Available: http://www.vlfeat.org/

[35] J. Wan, V. Athitsos, P. Jangyodsuk, H. J. Escalante, Q. Ruan, and I. Guyon, "CSMMI: Class-specific maximization of mutual information for action and gesture recognition," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3152–3165, Jul. 2014.
[36] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3169–3176.
[37] H. Wang, C. Yuan, W. Hu, H. Ling, W. Yang, and C. Sun, "Action recognition using nonnegative action component representation and sparse basis selection," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 570–581, Feb. 2014.
[38] Z. Wang, Q. Shi, C. Shen, and A. van den Hengel, "Bilinear programming for human activity recognition with unknown MRF graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1690–1697.
[39] J.-S. Wu, W.-S. Zheng, and J.-H. Lai, "Approximate kernel competitive learning," *Neural Netw.*, vol. 63, pp. 117–132, Mar. 2015.
[40] C.-N. J. Yu and T. Joachims, "Learning structural SVMs with latent variables," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1169–1176.
[41] J. Zhang and S. Gong, "Action categorization by structural probabilistic latent semantic analysis," *Comput. Vis. Image Understand.*, vol. 114, no. 8, pp. 857–864, 2010.
[42] J. Zhang and S. Gong, "Action categorization with modified hidden conditional random field," *Pattern Recognit.*, vol. 43, no. 1, pp. 197–203, 2010.
[43] X. Zhen, L. Shao, and X. Li, "Action recognition by spatio-temporal oriented energies," *Inf. Sci.*, vol. 281, pp. 295–309, Oct. 2014.

Xiaobin Chang received the B.Sc. degree from Sun Yat-sen University, in 2012, where he is currently pursuing the M.Sc. degree under the supervision of Dr. W.-S. Zheng. He is interested in image and video caption generator and related machine learning methods. His current research interests are in human activity recognition and visual surveillance.

Wei-Shi Zheng joined university under the one-hundred-people program in 2011. He is currently an Associate Professor with Sun Yat-sen University. His research direction is machine vision and intelligence learning. His current interests are in object association and activity analysis in computer vision, in particular, focusing on person re-identification and activity recognition for visual surveillance. He has authored over 60 papers, including over 40 publications in main journals (the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON NEURAL NETWORKS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B, and *Pattern Recognition*) and top conferences (ICCV, CVPR, IJCAI, and AAAI). He received the New Star of Science and Technology of Guangzhou in 2012, and the Guangdong Natural Science Funds for Distinguished Young Scholars in 2013. He joined the organization of four tutorial presentations in ACCV 2012, ICPR 2012, ICCV 2013, and CVPR 2015 along with other colleagues.

Jianguo Zhang received the D.Phil. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2002. He was a Lecturer with the Electronics, Electrical Engineering and Computer Science Department, Queen's University Belfast, U.K., from 2007 to 2010, and a Researcher with the Department of Computer Science, Queen Mary University of London, from 2005 to 2007, the Lear Group, INRIA Rhône-Alpes, France, from 2003 to 2005, and the School of Electrical and Electronic Engineering, Nanyang Technological University of Singapore, from 2002 to 2003. He is currently a Senior Lecturer (an Associate Professor) of Visual Computation with the School of Computing, University of Dundee. His research interests are visual surveillance, object recognition, image processing, medical imaging, and machine learning. He received the Best Cancer Paper Award at the Medical Image Understanding and Analysis Conference in 2014, and the best paper award at the International Machine Vision and Image Processing Conference in 2008. He won both tasks of the first international contest on Performance Evaluation of Indirect Immunofluorescence Image Analysis Systems at the International Conference on Pattern Recognition in 2014, the Best Performing Running Up Award on Brain Tumor Digital Pathology Segmentation Challenge at MICCAI 2014, and won the International Pascal Visual Object Classification Challenge twice in 2005 and 2006. He also received the President Prize of the Chinese Academy of Sciences. He was the Founding Co-Chair of the International Workshop Series on Video Event Categorization, Tagging and Retrieval (2009–2014), and has served as an Area Chair of the British Machine Vision Conference annually since 2011. He was an Editor of a book entitled *Intelligent Video Event Analysis and Understanding* (Springer-Verlag series), and a Guest Editor of *Pattern Recognition*.