Discriminative Context Models for Collective Activity Recognition

Chaoyang Zhao[1*], Wei Fu[1*], Jinqiao Wang[1], Xiao Bai[2], Qingshan Liu[3] and Hanqing Lu[1],
[1]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China.
[2] School of Computer Science and Engineering, Beihang University.
[3]School of Information and Control, Nanjing University of Information Science and Technology.
{chaoyang.zhao,wfu,jqwang,luhq}@nlpr.ia.ac.cn, baixiao.buaa@gmail.com, qsliu@nuist.edu.cn

***Abstract*—Context information has been widely studied for recognizing collective activities. Most existing works assume that all individuals in a single image share the same activity label. However, in many cases, multiple activities can be coexisted and serve as the context for each other in real-world scenarios. Based on this observation, we propose a novel approach to model both the intra-class and inter-class behavior interactions among persons in the scenario. By introducing the intra-class and inter-class context descriptors, we propose a unified discriminative model to jointly capture the individual appearance information and the context patterns around the focal person in a max-margin framework. Finally, a greedy forward search method is utilized to optimally label the activities in the testing scene. Experimental results demonstrate the superiority of our approach in activity recognition.**

***Keywords*-collective activity; context information; structure modeling;**

I. Introduction

Human activity recognition is of great scientific and practical importance and has received more and more attentions in computer vision community in the past few years. Many works focus on single person action recognition. However, in some complex scenarios, analysis of a single individual could not yield reliable recognition results because human activities often involve both the action of the single person and the interaction information among all the persons in the scene. For a scene under unrestrictive conditions such as dynamic cluttered background, variations in illumination and viewpoint, intra class variability in the human appearance and non-static cameras, recognizing the action of a single person seems to be unachievable without the support of context information.

To this end, rather than recognizing an individual action in isolation, some researchers turn to analyzing the behavior for a group of persons with interactions among each other. These works are referred to as "collective activity analyzing" where actions of individuals are often interdependent and some coherency between these actions may exist. In the past few years, many works [5], [3], [2] have dedicated to employing these coherency to reinforce the recognition of each person's action and have achieved significant improvement on the performance of the activity recognition task. However, most existing works focus on modeling the interactions in the same collective activity. We argue that there can exist multiple collective activities in a scene and contextual information can also be shared among persons with different activity labels. In this paper, we focus on developing approach for recognizing collective activities by modeling both the intra-class and inter-class contextual information.

Our approach is based on the observation that there often exist more than one collective activities in a scene. Take the crossroad scene in Fig.1 as an example, there are two collective activities: *waiting* and *crossing*. Here we regard the woman within the red solid bounding box as a focal person. Traditional methods model the interactions between the focal person and the persons around her with the same activity label "*waiting*", i.e., persons within the yellow dashed box. Only the intra-class context information among persons is utilized in their methods. However, imagine the focal person is waiting because the traffic light turns red, and at the same time the other persons in the blue dashed box can cross the road. Therefore, persons with "crossing" activity could also provide an important cue for recognizing the action of the focal person.

As to the analysis above, we can benefit from both the intra-class context information and the behavior interactions among persons with different activity labels (we call them as inter-class context information) for recognizing collective activities in a single image. In practice, not all the context information from inter-class offers help, we would like to select only the important context to disambiguate the activities label of an individual. To this end, we use a discriminative model to formulate these kinds of contextual information, and automatically decide whether the interactions of two persons should be considered. By introducing carefully designed context activity descriptors, our framework jointly captures the individual appearance information, the intra-class interactions among persons within the same activity, and also the inter-class relationships among different persons in different activity groups. Experimental results on the challenging real world dataset [1] show the superiority of our approach.

Our contributions are listed as follows:

1) Different from the methods considering the whole image with the same collective activity label, our approach can discover two or more collective activity class labels in an image, which is more reasonable in real-world scenarios.
2) A unified discriminative framework of collective activity recognition is proposed to jointly model individual appearance feature, intra-class and inter-class contextual

*These two authors contributed equally to this work.

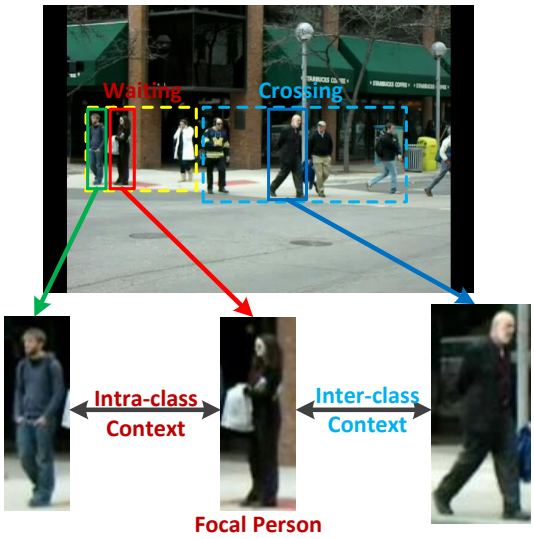

Figure 1. An illustration of context information shared by multiple collective activities within the same image. The action of this focal person is influenced both by the persons next to her with the same activity label and the persons with a different activity label far from her.

information.

3) An inter-class context descriptor is introduced to model the behavior interactions among persons with different activities, which could improve the recognition performance significantly.

II. Related work

Human visual recognition using context information has received much attention recently in computer vision community. Many works has been done on exploiting context information between scenes and objects [6], objects and objects [4], [7], [8]. Human activity recognition involves identification of human actions and recognition of an ensemble of collective actions. Many previous works exploited context provided by scenes [9] or objects [10], [11] to improve the classification performance. Object-action context [12], [13], [14] was another popular type of context used for human-object interaction modeling. Choi *et al.* modeled the crowd context [5] to establish the activities performed by individuals in a crowd. Lan *et al.* used a high level latent discriminative model [3], [15] to explore the group-person interaction and person-person interaction context. Most of the previous works focused on modeling human interactions with the same activity label, none of them considered the relationship across different activities. Xiang *et al.* used Markov Random Fields to model the intra-class context information [20]. Zhu *et al.* defined contextual information between different activities to improve the recognition rate [15]. They use a structural model to integrate motion features and context features in and between activities. However, their definition of activity is based on one person interacting with the surroundings and can hardly solve the problem of collective activity recognition with many persons interacting with each other, while our approach is dedicated to model this contextual information among different activities for improving the performance of collective activity recognition.

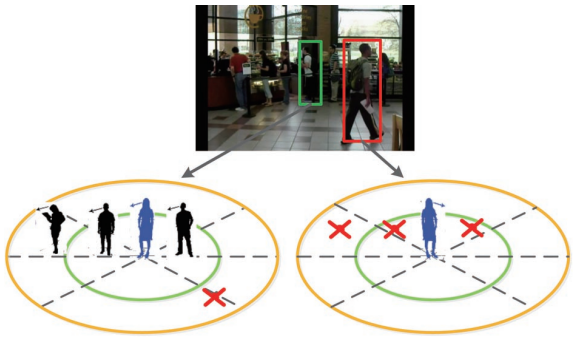

Figure 2. Intra-class context information extraction from the "sub-context region" of the focal person. The highlighted focal person is related to the people nearby with the same activity label. "X" means we exclude the people nearby with different activity labels.

The work in [3], [4], [15] used structural models for the strong ability to model low-level features and middle level features jointly. The inference method on a structural model often needs to search through the graphical structural in order to find the one that maximizes the potential function [3]. This kind of solution is very time consuming. A greedy search strategy was first proposed in [4] and extended to represent related activities in video [15], with which computational complex could be reduced with a considerable performance. Inspired by this, in this paper we also use the structural model and solve the inference problem with a modified greedy search strategy.

III. Our approach

Most existing works on collective activity recognition are based on the assumption that there is only one activity existing in a single image [2], [3], which often leads to misclassification for persons with different activity labels. Our approach is to recognize different collective activities and assign the corresponding activity label for each person in a single image. A discriminative context model is proposed for collective activity recognition. Our model includes not only the context information among persons belonging to the same activity group, but also the context information of persons with different activity labels.

A. Problem formulation

Assuming there are M classes of collective activities in the scene, where the label $y_i \in \{1, 2, ..., M\}$ denotes the activity class of a person. Let $Y = \{y_i : i = 1, 2, ..., N\}$ be the label set for all N persons in an image. The task is then converted into finding the optimal hypothesis label set Y for all the persons $X = \{x_i : i = 1, 2, ..., N\}$ in the image. By considering all the context interactions among the persons, the proposed approach is formulated as follows:

$$S(X,Y) = \sum_{i=1}^{N} w_u^T \Phi_u(x_i, y_i) + \sum_{i=1}^{N} w_c^T \Phi_c(x_i, y_i) + \sum_{i,j=1, y_i \neq y_j}^{N} w_s^T \Phi_s(x_i, x_j, y_i, y_j) \quad (1)$$

where $\Phi_u(x_i, y_i)$ denotes the potential of the intra-activity appearance feature, $\Phi_c(x_i, y_i)$ denotes the potential of the intra-class contextual feature and $\Phi_s(x_i, x_j, y_i, y_j)$ denotes the potential of the inter-class contextual feature. The subscript y_i, y_j means different activity class labels, and i, j stands for different persons in an image.

We first extract appearance features for each person, then compute the contextual features by considering interactions together with other persons nearby. We assume there are connections among persons with different activity labels and model this connections with discriminative models.

B. Intra-activity appearance descriptor

The potential function $\Phi_u(x_i, y_i)$ describes the unary feature of the ith person with the $y_i th$ activity category. Rather than directly using certain raw features (e.g. HOG[17]), here we use the person pose classification scores as the feature vector. We train a 8-class SVM classifier based on the HOG descriptor for each activity label, which contains eight pose categories: *right, front-right, front, front-left, left, back-left, back* and *back-right*. Together with a bias term, we have the intra-class appearance potential for the ith person belonging to the $y_i th$ activity as follows:

$$\Phi_u(x_i, y_i)^T = (S_{1i}, ..., S_{Ki}, 1) \quad (2)$$

where $K = 8$ is the number of pose categories within a activity and 1 is the bias term.

C. Intra-class context descriptor

$\Phi_c(x_i, y_i)$ models the contextual information for all the persons sharing the same activity label within a region area, which we referred to as intra-class context feature. The method is inspired by "context region" used in [3], which indicates that one person's activity is related to others around him both in space and time. Given the ith person as the focal person, the intra-class context descriptor is computed from the pose descriptors of persons in the context region we define. For a person j inside the context region of the focal person, we have the pose classification results $(S_{1j}, ..., S_{Kj})$, here K is the number of poses. Similarly as [2], here we define "sub-context regions" around the focal person in space. Supposing that the context region contains M sub-regions, the context descriptor is represented as a $M \times K$ dimensional vector:

$$\Phi_c(x_i, y_i)^T = (\max_{j \in \mathcal{N}_1(i)} S_{1j}, ..., \max_{j \in \mathcal{N}_1(i)} S_{Kj}, ... \max_{j \in \mathcal{N}_M(i)} S_{1j}, ..., \max_{j \in \mathcal{N}_M(i)} S_{Kj}) \quad (3)$$

where $\mathcal{N}_m(i)$ indicates the index of a person in the mth "sub-context region" of the ith person. Here we set K=8, M=2. $\mathcal{N}_1(i)$ and $\mathcal{N}_2(i)$ are circles of $0.5h$ and $2h$ (h is the height of the focal person i) respectively. An illustration of the process is shown in Fig. 2. This descriptor captures the contextual information of persons nearby as well as the focal person's posture information. Instead of considering all the persons inside the context region, here we only consider the persons with the same activity labels. We believe that this intra-class context descriptor represents the global relationships inside an activity group, person with different activity labels inside the context region may bring confusion when computing $\Phi_c(x_i, y_i)$.

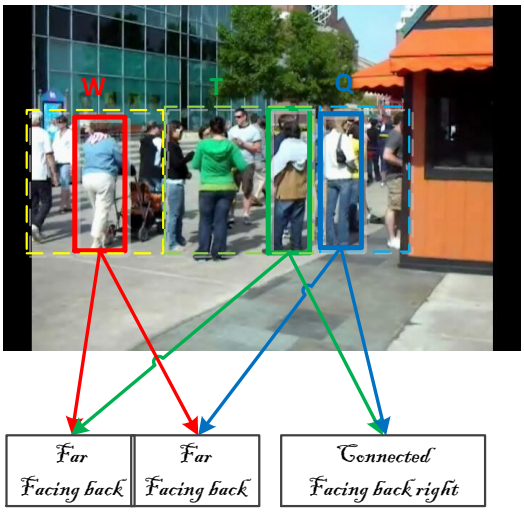

Figure 3. Intr-class context information extraction from persons with different activity labels.

D. Inter-class context descriptor

In order to take into account the inter-class interactions, we model the spatial contextual information between co-existing persons with different activity labels. Notice that the intra-class context descriptor already modeled the spatial context information between persons with same activity label, here we only consider persons with different activity labels as shown in Fig 3. For two persons i and j with different activity labels y_i and y_j, their spatial context information are modeled with $\Phi_s(x_i, x_j, y_i, y_j)$:

$$\Phi_s(x_i, x_j, y_i, y_j)^T = [bin(\frac{d_{ij}^2}{h_i \cdot h_j}, 3), pose(i, j)] \quad (4)$$

where d_{ij} is the distance between bounding box i and j, h_i and h_j are the related height, respectively. The spatial relationships are further divided into 3 by using $bin(r, 3)$, where the corresponding bins are defined as *connected, near* and *far*. And $pose(i, j)$ is defined as $max(\Phi_u(x_i, y_i), \Phi_u(x_j, y_j))$.

IV. Optimization

In this section, we will describe the optimization of the proposed discriminative model from two aspects: model learning and inference.

A. Model learning

In the procedure of model learning, the goal is to estimate the optimal parameters of appearance model w_u, the intra-class context model w_c and the inter-class context model w_s. Equ. 1 can be rewritten as follows:

$$S(X,Y) = w^T \Phi(X,Y) \tag{5}$$

$$w = \begin{bmatrix} w_u \\ w_c \\ w_s \end{bmatrix}, \quad \Phi(X,Y) = \begin{bmatrix} \sum_i \Phi_u(x_i, y_i) \\ \sum_i \Phi_c(x_i, y_i) \\ \sum_{i,j} \Phi_s(x_i, x_j, y_i, y_j) \end{bmatrix}$$

where w is the model parameter we need to learn. Assume that we have the training images X_i and their corresponding label set Y_i, we want to train a model w that, with a new image X_j, tends to produce the true label vector $Y_j^* \simeq Y_j$. The objective function can be converted to a regularized learning problem as follows:

$$\begin{aligned} &\arg \min_{w, \xi_i \geq 0} w^T w + C \sum_i \xi_i \\ &s.t. \quad \forall i, H_i \quad w^T \Delta\Phi(X_i, Y_i, H_i) \geq l(Y_i, H_i) - \xi_i \end{aligned} \tag{6}$$

where $\Delta\Phi(X_i, Y_i, H_i) = \Phi(X_i, Y_i) - \Phi(X_i, H_i)$ and $l(Y_i, H_i)$ is the loss function to measure the difference between Y_i and H_i. In our approach, the loss function is defined as the Hamming loss of ground truth label set Y_i and H_i. We use the cutting plane optimization algorithm proposed in [4] to solve our problem.

B. Inference

With the model parameter w, the inference procedure is to find the best label set Y^* for an input image X. The task is to solve the following optimization problem:

$$Y^* = \arg \max_Y S(X, Y) \tag{7}$$

A greedy forward search strategy [4], [15] is proposed to find the optimum labels and durations of the targeted activities. While this greedy search algorithm cannot guarantee a globally optimum solution, in practice it works well to find good solutions. Inspired by their method, we identify our optimum label vector Y^* with a similar greedy search strategy. To be specific, we define a function to measure the score change by adding the person-class pair (i, a) as follows:

$$\Delta(i,a) = S(X, Y(I \cup (i,a))) - S(X, Y(I)) \tag{8}$$

Firstly, we initialize the label vector Y to be 0 for all persons. Then we greedily select the ith single person that, when labeled as a particular activity class a, increases the score S by the largest amount. After that we have $y_i = a$, and the ith person is added to the labeled set I. We repeat this procedure until all the N persons are re-assigned new labels. The whole computation can be very efficient by tracking the potential gain of adding label assign incrementally. Alg.1 gives the overview of the inference process.

Algorithm 1 The Greedy Forward Search Algorithm

Input:
A testing image with N total persons
Output:
The optimal reassigned label vector Y^*
1: **Initialization:**
$I = \emptyset$, $S = 0$
$\Delta(i,a) = \omega_u^T \Phi_u(x_i, a) + \omega_c^T \Phi_c(x_i, a)$
2: **Repeat:**
$(i,a)^{opt} = arg \max_{(i,a) \notin I} \Delta(i,a)$;
$I = (i,a)^{opt} \cap I$;
$S = S + \Delta(i,a)^{opt}$;
$Y^* = Y(I)$;
$\Delta(i,a) = \Delta(i,a) + \omega_s^T \Phi_s(x_i, x_{i^*}, a, a^*)$
3: **Until** $\Delta(i,a)^{opt} < 0$ or all N persons are labeled

V. Experiments

In order to evaluate the performance of the proposed approach, we carry our experiments on the challenging real world dataset [1] and compare with state-of-the-art approaches.

Dataset: The collective activity dataset [1] contains 44 video clips acquired using low-resolution handhold cameras. All the persons in every 10th frame of the videos are assigned one of the following five collective activity categories: *crossing, waiting, queuing, walking* and *talking*, and one of the following eight pose categories: *right, front-right, front, front-left, left, back-left, back* and *back-right*. More than $1/5$ of the clips contain two or more activities. We select $1/4$ of the video clips to form the testing set, and the rest are used for training.

A. Comparisons using different feature fusion approaches.

The confusion matrix of classification accuracy is used to evaluate the effectiveness of our approach. And we present the results of our methods with different feature fusion strategies as below:

1) A baseline classifier only using appearance features for each activity category.
2) A classifier using appearance features and simple spatial context information. Here the spatial context encodes the relative spatial relationship among persons by a spatial histogram feature [4].
3) A classifier using appearance features, intra-class context described in section III-C, and the spatial context from different activities.
4) A classifier using appearance features, intra-class context features and inter-class context features.

The confusion matrices of different feature fusion approaches are shown in Fig. 4, in which we can see a significate improvement in "cross", "wait", "walk" and "talk" from Fig. 4(a) to Fig. 4(b). This is consistent to the fact that contextual information is important in collective activity recognition. Spatial context in [4] describes the spatial relationships among different persons, which makes use of the advantage that some activities are sensitive to the spatial relationship with other persons nearby, such as "walking" and "crossing" in Fig. 4(b). Without considering the inter-class information, "appearance feature + spatial context" often leads to poor

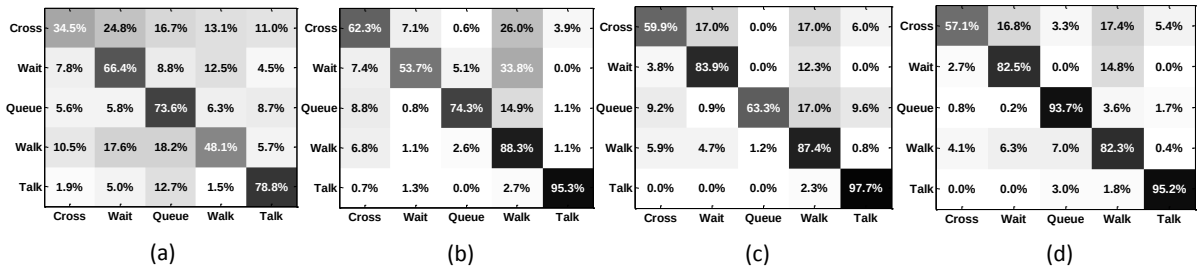

Figure 4. Confusion matrices for activity classification accuracy on the collective activity dataset with different feature fusion strategies: (a) appearance feature. (b) appearance feature and spatial context. (c) appearance feature, intra-class context and spatial context. (d) appearance feature, intra-class context and inter-class context. Rows are ground truths, and columns are predictions. Each row is normalized to sum to 1.

performance in some activities like "waiting" and "queueing". Comparing Fig. 4(b) with Fig. 4(c), we can observe that "intra-class context" brings great improvement for those activities in which persons stay close to each other. Using "spatial context" as inter-class context brings a performance degradation in "queueing". Our approach combines appearance feature, intra-class context feature and the inter-class context feature as describes in section III all together. As the result shows in Fig. 4(d), though the accuracy drops about 2% for activities "crossing", "waiting", "walking" and "talking" compared to Fig. 4(c), our approach still yields most reasonable result. The average accuracy improve about 4.5% compared to Fig. 4(c).

B. Comparisons with state-of-the-art approaches

We compare our approach with state-of-the-art results for activity classification. As shown in Table. I, the top three rows are the experimental results with different features as mentioned before. The action context baseline method [2] defined a descriptor to encodes information about action of an individual person and behaviors of all the other persons nearby. This approach can be regarded as a special case of our appearance feature and intra-class context feature fusion strategy. The random forest method [5] constructed a random forest structure to learn the context for collective activity recognition. And the latent model method [3] used a latent variable framework to explore the group-person and person-person interactions. As we can see, due to allowing the co-existence of multi-class activities in a scenario, our approach achieves the best classification performance comparing with state-of-the-arts by taking into account both the intra-class and inter-class context information.

Some examples of collective activity recognition results can be found in Fig.5. Notice that our approach can classify multiple group activities existing in the same image. Each person in the image is assigned with a activity label, and the persons with the same label are also used to form the collective activity results. We can see that images with collective activity labels can be recognized correctly. Besides, some false results together with their ground truth labels are shown in the second row in Fig.5. The wrong label assignments for some persons in the scene may be due to the fact that some activities need additional information to be correctly recognized.

Approaches	Average Accuracy (%)
Appearance Features	60.6
Spatial Context	76.6
Intra-class Context	78.7
Action Context [2]	68.2
Random Forest [5]	70.9
Latent Model [3]	79.1
Our approach	**83.2**

Table I
COMPARISON RESULTS WITH STATE-OF-THE-ARTS.

VI. CONCLUSION

In this paper we present a novel approach to recognize collective activities. The approach combines the individual appearance feature, intra-class and inter-class context information into a discriminative structure optimization framework. Different from the methods considering the whole image with the same collective activity label, our approach can discover two or more collective activity class labels existing in a scenario. Experimental and comparison results demonstrate that jointly modeling the individual appearance feature and the activity context features can significantly improve the recognition accuracy of collective activities.

VII. ACKNOWLEDGEMENT

This work was supported by 973 Program (2010CB327905) and National Natural Science Foundation of China (61273034, 61070104, and 61202325).

REFERENCES

[1] Wongun Choi and Shahid, K. and Savarese, S., *What are they doing? : Collective activity classification using spatio-temporal relationship among people*, in Computer Vision Workshops (ICCV Workshops), 2009.

[2] Tian Lan, Yang Wang, Greg Mori and Stephen N. Robinovitch, *Retrieving actions in group contexts*, International Workshop on Sign Gesture Activity, 2010.

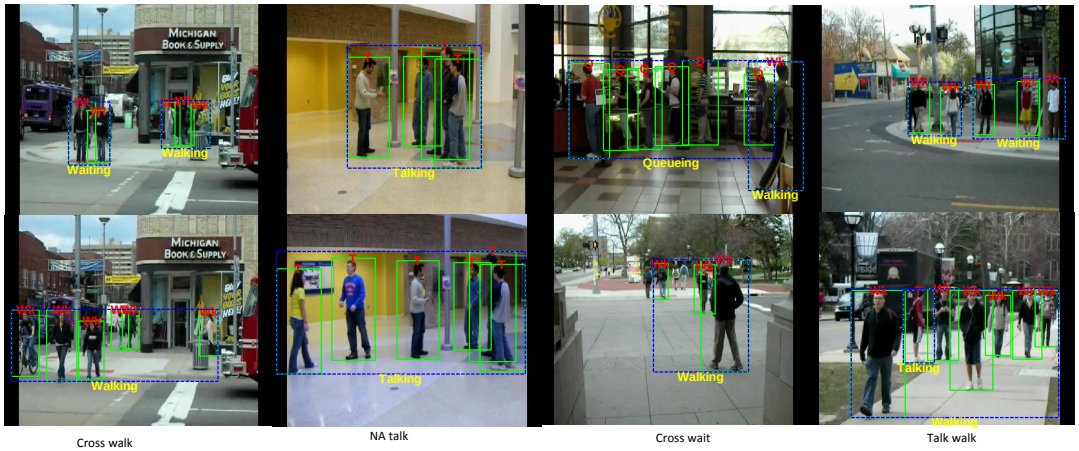

Figure 5. Visualization of collective activity recognition results on the collective activity dataset. The top row shows some correct results. The bottom row shows some false results together with their ground truth labels.

[3] Tian Lan, Yang Wang, Weilong Yang and Greg Mori, *Beyond Actions: Discriminative Models for Contextual Group Activities*, in Neural Information Processing Systems Foundation, 2010.

[4] Chaitanya Desai, Deva Ramanan and Charless C. Fowlkes, *Discriminative Models for Multi-Class Object Layout*, in International Journal on Couputer Vision, 2010.

[5] Wongun Choi, Khuram Shahid and Silvio Savarese, *Learning Context for Collective Activity Recogniton*, in IEEE International Conference on Computer Vision and Pattern Recognition, 2011.

[6] K.P. Murphy, A. Torralba and W. T. Freeman, *Using the Forest to See the Trees: A Graphical Model Relating Features, Objects, and Scenes*, in Neural Information Processing Systems Foundation, 2004.

[7] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora and S. Belongie, *Objects in Context*, in IEEE International conference on Computer Vision, 2007.

[8] A. Jain, A. Gupta and L. S. Davis, *Learning What and How of Contextual Models for Scene Labeling*, in European Conference on Computer Vision, 2010.

[9] M. Marszalek, I. Laptev and C. Schmid, *Actions in Context*, in IEEE International Conference on Computer Vision and Pattern Recognition, 2009.

[10] D. Han, L. Bo and C. Sminchisescu, *Selection and Context for Action Recognition*, in IEEE International Conference on Computer Vision, 2009.

[11] H. Kjellstrom, J. Romero, D.M. Mercade and D. Kragic, *Simulataneous Visual Recognition of Manipulation Actions and Manipulated Objects*, in European Conference on Computer Vision, 2008.

[12] B. Yao and L. Feifei, *Grouplet: A Structured Image Representation for Recognizing Human and Object Interactions*, in IEEE International Conference on Computer Vision and Pattern Recognition, 2010.

[13] C. Desai, D. Ramanan and C. Fowlkes, *Discriminative Models for Static Human-Object Interactions*, in Proc. Workshop Structured Models in Computer Vision, 2010.

[14] B. Yao and L. Feifei, *Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities*, in IEEE International Conference on Computer Vision and Pattern Recognition, 2010.

[15] Yingying Zhu, Nandita M. Nayak and Amit K. Roy-Chowdhury, *Context-Aware Modeling and Recognition of Activities in Video*, in IEEE International Conference on Computer Vision and Pattern Recognition, 2013.

[16] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan, *Object Detection with Discriminatively Trained Part Based Models,* in IEEE trans on Pattern Recognition and Machine Intellgence, 2010.

[17] Navneet Dalal and Bill Triggs, *Histograms of Oriented Gradients for Human Detection*, in IEEE International Conference on Computer Vision and Pattern Recognition, 2005.

[18] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun., *Support Vector Learning for Interdependent and Structured Output Spaces,* in IEEE International Conference on Machine Learning, 2004.

[19] Changsheng Li, Qingshan Liu, Jing Liu, and Hanqing Lu., *Learning ordinal discriminative feature for age estimation,* in IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[20] Yu Xiang, Xiangdong Zhou, Tat-Seng Chua, and Chong-Wah, *Ngo: A revisit of Generative Model for Automatic Image Annotation using Markov Random Fields,* in IEEE conference on Computer Vision and Pattern Recognition, 2009.