Examen de l'option Interprétation d'images

Master AIC

7 février 2017 - 9h-11h30 - Seul document autorisé : deux feuilles A4 recto-verso de notes de cours

A la suite de la lecture de l'article joint, répondez aux questions suivantes en argumentant le plus possible, et avec une formulation mathématique lorsque c'est pertinent.
After reading the joint paper, answer the following questions with as much argumentation as possible, and using mathematical formulation when relevant.

1. Proposer un titre et un résumé pour l'article.
   Provide a title and an abstract for this paper.
2. Quels sont les points forts de l'article ? Parmi ces points, choisissez-en un que vous discuterez en détail.
   What are the main contributions of this paper ? Choose one of them and explain it in detail.
3. Quels sont les limites de la méthode proposée ?
   Explain the limits of the proposed method.
4. Qu´est-ce que les auteurs appellent modèle génératif ? interactions inter et intra niveaux ?
   What do the authors mean by generative model ? inter and intra level interactions ?
5. Les expériences vous semblent-elle suffisamment complètes ou y a-t-il des situations qui ne sont pas bien prises en compte ?
   Are the experiments rich enough or are there situations which are not adequately covered ?
6. Que représente le descripteur AD ? Quels autres descripteurs pourraient être utiles ?
   What does the AD descriptor represent ? Which other descriptors coud be useful ?
7. Serait-il possible d'introduire des relations spatiales entre les objets pour enrichir la description des actions et des interactions ? Comment cela pourrait-il être réalisé ?
   Would it be possible to introduce spatial relationships between objects in order to enrich action and interaction description ? How could it be done ?
8. Quel serait l'apport d'une ontologie pour modéliser les connaissances ? pour guider le raisonnement ?
   What would be the benefit of an ontology to model knowledge ? To guide reasoning ?
9. Qu'apporterait une représentation par graphe de l'image ? Préciser comment serait construit le graphe et comment il pourrait être utilisé dans le système proposé.
   What would be the benefit of a graph representation of the image ? Describe precisely how you would build the graph, and how you would use it ?
10. Quelles seraient les étapes à développer pour qu'un robot puisse imiter les actions reconnues dans une image ?
    Which would be the necessary stages for a robot to imitate the actions that are recognized in an image ?
11. Comment pourrait-on étendre la méthode à des séquences d'images ? Qu'apporterait la dimension temporelle ?
    How could the method be extended to image sequences ? What could be the benefit of the temporal dimension ?

1 Introduction

Vision-based human activity recognition has always been a valuable research field of artificial intelligence. It is not hard for humans to recognize multiple group activities that are mixed and contained in one still image, but it remains a highly challenging problem for a computer recognition system to recognize them automatically. An automatic recognition system has many great applications, such as detecting abnormal group behaviours (e.g., illegal gathering) appearing on each of unrelated photos on a social media network (e.g., Twitter) in order to investigate potential terrorist threats.

The first challenge of this problem is the use of only "still images". The features that can be extracted from a single image are a lot fewer than the ones from a video clip. With a video, sufficient low level, spatio-temporal features can be extracted to support recognition in multi-steps: discovering groups [Alameda-Pineda *et al.*, 2015], recognizing individual actions [Wang *et al.*, 2015; Zhang *et al.*, 2015] and charactering single group activities based on the discovered groups and recognized individual actions [Lan *et al.*, 2012c]. On still images, we cannot follow this multi-step approach for the following reasons. The existing image-based approaches for discovering groups usually require known actions and hence be dependent on action recognition [Choi *et al.*, 2014; Odashima *et al.*, 2012]. For action recognition on still images, the low-level spatial features, instead of spatio-temporal features used on video-based approaches, are hardly enough for recognizing individual actions. Furthermore, high level features (i.e., contextual cues) [Guo and Lai, 2014; Zhou *et al.*, 2015] for action recognition are hard to extract under a complex scenario showing multi-group activities on an image. Without well-recognized individual actions, the existing image-based methods [Lan *et al.*, 2010; 2012c; Xiong *et al.*, 2015] for recognition of a single group activity cannot be applied to recognition of multi-group activities on an image.

Another challenge of this problem is from recognizing "multiple mixed group activities". Most of the existing work on group activity recognition implicitly assumed that only one group activity existed in one image [Li and Li, 2007; Choi and Savarese, 2014; Cheng *et al.*, 2014; Wang *et al.*, 2015]. Some other work learned interactions between pairs of units (i.e., persons or groups) and divided interactions among multiple (i.e., more than two) units into interactions between pairwise units [Lan *et al.*, 2012a; Tran *et al.*, 2013; Zhao *et al.*, 2014]. There approaches limit the performance when images contain multiple groups, because there is no mathematical evidence supporting that the joint probability of three or more units is proportional to neither the product of the joint probabilities of all combinations of pairwise groups nor the product of the conditional probabilities of them.

Inspired by the co-occurrence of scenes and activities, we design a four-level generative model consisting of scene/event, group activity, standard pose and visible pose levels. This model represents the intra-level interactions among groups (that are hard to be represented mathematically) by the inter-level interactions between groups and scenes/events. Similarly, this model represents the intra-level interactions among persons within a group by the inter-level

*Corresponding Author.

interactions between the standard poses and visible poses. The inter-level interactions are computed using the generative relationships between the corresponding levels. These representations of interactions also make our model be more robust against occlusion and overlap of poses. When we label an unknown image using a stochastic methodology, different from the above-mentioned multi-step approaches, the scenes/events, group activities and standard poses will be determined simultaneously. The generative relationships between adjacent levels will be considered synthetically, and thus will allow the model to exploit the limited information in a still image effectively. To compare with other approaches, we conduct experiments on the popular Collective Activity Classification (CAC) dataset [Choi *et al.*, 2009], on which many state-of-the-art methods have been performed. The experimental results demonstrate that our model produces good interpretations for mixed group activities and outperforms all of the state-of-the-art methods on this dataset.

We highlight our contributions as follows.

- We propose a generative model to provide an interpretation of multiple mixed group activities captured in a still image. This model is not a multi-step approach and it does not rely on individual human action recognition.
- Our model can well model interactions among multiple units (groups or persons), and it is more robust against occlusion and overlap of poses in an image.
- A new set of labels that annotate the positions of human's body parts on all images in the CAC dataset is produced and used as the ground truths of human poses. Furthermore, a *Special dataset* containing all of the CAC images which contain multiple group activities is also created. These two datasets are released for others to use when working on the same areas for recognition of mixed group activities.

2 Related Work

Group activity recognition has received much research attention for it is an intriguing scientific question and a highly useful engineering application. Many video-based methods have been proposed. They perform well based on spatio-temporal features, for example, Wang et al. employed a three-layered AND-OR graph [Wang *et al.*, 2015] to form a multi-level representation for activities. Choi et al. modelled the activity performed by a crowd with crowd context [Choi and Savarese, 2014]. However, their performance will significantly decrease if the spatio-temporal features are replaced with spatial features that are available in images. With a shorter history, very few image-based methods have been proposed, for example, Xiong et al. recognized complex events from static images through fusing deep channels [Xiong *et al.*, 2015]. However, both image-based and video-based methods above implicitly assumed that only one group activity exists in one image or video clip. They laid more emphases on modelling the interaction or context within a group.

There have been even fewer papers considering multiple groups. Some of them focused on discovering groups [Odashima *et al.*, 2012; Choi *et al.*, 2014; Sun *et al.*, 2014; Alameda-Pineda *et al.*, 2015]. For activity recognition, the existing approaches tended to approximately divide interactions among multiple units into interactions between pairwise units. For example, Zhu et al. proposed a structural model for videos and integrated motions and context features in and between activities [Zhu *et al.*, 2013]. Zhao et al. designed a discriminative model to analyse the inter-class and intra-class context between pairwise persons [Zhao *et al.*, 2014]. Differently, we propose a generative model to provide exact interpretations for interactions among multiple units.

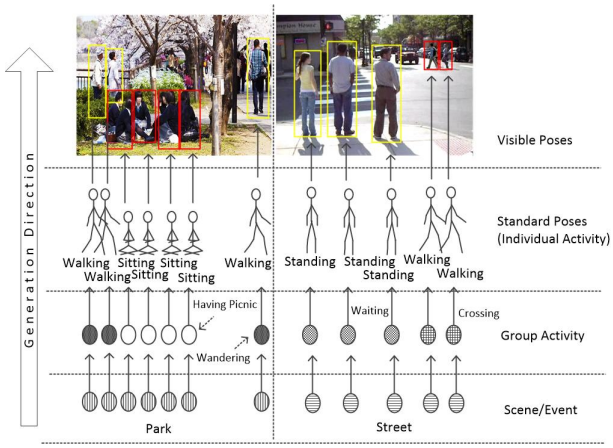

Figure 1: Overview of the generative model. Each person has four variables, corresponding the scene or event that a person is in, the group activity that the person is performing, the standard pose that the person is in, and the visible pose that the person is observed respectively.

The most relevant work to ours is a discriminative model proposed by Lan et al. [Lan *et al.*, 2012a]. It also has a four-level structure, but our usage of the levels and way of connections between levels go totally differently. The levels in their model are used to define different kinds of labels to be recognized, while our levels are used to convert intra-level interactions into mathematically calculable inter-level interactions. It is not necessary to assign meanings of reality for the values of the scene/event and standard pose levels of our model. Besides, the way of connections of their model considers only the interactions between pairwise roles. In our model, however, the way of connections and the generative relationships between levels enable us to model both interactions among multiple groups and interactions among multiple persons within a group.

3 Generative Model

3.1 Overview

Given an image containing several human groups, we aim to predict the group activities for every person. As shown in Figure 1, each person is associated with four variables at four levels respectively. They correspond to the group activity that the person is performing, the standard pose that the person is in, and the visible pose that the person is observed.

The scene/event variables could have either specific definitions or meanings, such as "in a park" and "on a street", or

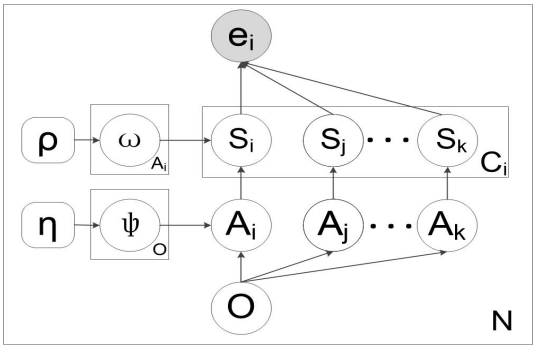

Figure 2: The graphical model. The model consists of four types of nodes, corresponding to the scene/event (O), the group activity (A), the standard pose (S) and the visible pose (e) levels, where i and j are the indexes for person i and j respectively, N is the total number of persons in the image and C_i donates the group that person i is from.

no explicit meaning or category name. They are just used to denote a certain type of images in a training dataset. Persons in one image have the same scene/event value, so we let the persons in one image share the same scene/event variable and the generation will start from this variable. The activity variable indicates the group activity performed by a person, and persons with the same activity label belong to one group naturally. The standard pose variable describes a person's most possible standard pose, and it is selected from the candidate standard poses corresponding to a certain activity. The visible pose of a person is simply the pose observed in the image and it is captured by pose estimation algorithms in practice. We use an action descriptor AD to describe both visible and standard poses. It is a vector formed by the absolute angle of the torso and the relative angles between other body parts and the torso. This form of AD connects the model and the image smoothly and contributes to the robustness against rotation.

The graphical model shown in Figure 2 illustrates the connections between the adjacent levels. The global interactions (GI) among groups, which originally exist within level A, are now converted into the combination of the co-occurrences between the scene/event and the group activities. Similarly, we introduce the standard pose level S to model the local interactions (LI) among persons within a group by taking into account the inter-level interactions between levels e and S. This level actually expands the group activities and refines the connection between a group activity and a visible pose. When designing the generative relationship between levels e and S, we consider the following four related factors. First, in practice, a human does not always perform a standard pose and his real pose will be more or less different from a standard pose. Second and third, overlaps and occlusions affect the precision of pose detection. Fourth, the LI causes deviation of visible poses from standard poses. Accordingly, we define the generative relationships as follows.

3.2 Generation Process

With the graphical model shown in Figure 2, we process the generation and give the distributions for the generative relationships between levels.

1. For an image I, the generation begins with a discrete random variable O standing for the event/scene category. We assume a fixed uniform prior distribution $p(O)$ for O like [Li and Li, 2007] and select $O \sim p(O)$.

2. For person i, given the scene/event class O, the activity A_i is chosen according to

$$A_i \sim p(A_i|O,\psi) = Mult(A_i|O,\psi), \tag{1}$$

where A_i is a discrete variable denoting the class label. A multinomial parameter ψ governs the distribution of A_i given O. ψ is a matrix of size $totalO \times totalA$, whereas η is a $totalA$ dimensional vector acting as a Dirichlet prior. Now, the GI ($p(A)$) can be represented as $\prod_i p(A_i|O)$.

3. Given the activity label A_i, the standard pose S_i of person i is chosen according to

$$S_i \sim p(S_i|A_i,\omega) = Mult(S_i|A_i,\omega), \tag{2}$$

where S_i is a discrete variable denoting a standard pose, which is learned by a SVM trained with a group of typical poses in the training dataset. The ω and ρ act as the multinomial parameter and the Dirichlet prior respectively as above. They both have $totalA$ lines corresponding to $totalA$ potential activity classes.

4. Given the S and A of all persons, the visible pose e_i of i is chosen according to

$$e_i \sim p(e_i|S_i, S_{all\ j\in G_i, j\neq i}), \tag{3}$$

where e_i is a discrete variable taking the form of an AD and G_i donates the persons in the same group as i. We use a matrix $\delta = [\delta_{ij}]$ to record the groups, where $\delta_{ij} = 1$ if and only if persons i and j are in the same group. The δ will change automatically during the inference process. In our method, given the visible pose, we decide the standard pose independently, and hence we decompose Eq. 3 as

$$\begin{aligned} p(e_i|S_i, S_{all\ j\in G_i, j\neq i}) &\propto p(e_i|S_i) \prod_{j}^{j\in G_i, j\neq i} p(e_i|S_j) \\ &= p(e_i|S_i) \prod_{j=1}^{N} p(e_i|S_j)^{\delta_{ij}}. \end{aligned} \tag{4}$$

where

$$p(e_i|S_i) = \frac{1}{Z_\alpha} exp(-\alpha \cdot \varphi(e_i, S_i)), \tag{5}$$

$$p(e_i|S_j) = \frac{1}{Z_\beta} exp(-\beta \cdot \varphi(e_i, S_j) \cdot \vartheta(e_i, e_j))(j \neq i). \tag{6}$$

α and β are scaling parameters. Z_α and Z_β are the normalization constants. $\varphi(e_i, S_i)$ and $\varphi(e_i, S_j)$ are scores returned by the SVMs corresponding to S_i and S_j, and represent the distances of the e_i from the separating hyperplanes of the SVMs. $\vartheta(e_i, e_j)$ is the distance between persons i and j.

Eq. 4 reflects the LI among persons in the same group. The SVM scores in Eq. 5 and Eq. 6 relate to the overlaps and occlusions. When we try to maximize the likelihood of Eq. 4, the SVM which has the lowest score will be selected

to provide the score, so that we find the most similar standard pose to the visible pose with the consideration of overlaps and occlusions. Therefore, we implement the four factors mentioned in Section 3.1.

Putting everything together, we arrive at the following joint distribution for the scene/event, group activities, standard poses and visible poses,

$$p(O, A, S, e|\psi, \omega) = p(O) \cdot \prod_{i=1}^{N} [p(A_i|O, \psi) p(S_i|A_i, \omega) p(e_i|S_i) \cdot \prod_{j=1}^{N} p(e_i|S_j)^{\delta_{ij}}]. \quad (7)$$

With these distributions (Eqs. 1-6), once we learn the parameters of the distributions, we can use the model to label unknown images through maximum likelihood estimation. Given the Dirichlet hyperparameters η and ρ, we use the same method as [Li and Li, 2007] to learn the parameters from the dataset. Moreover, we train SVMs in Eqs. 4-6 based on the training data. We first run a clustering algorithm on the visible poses of an activity in the dataset and then use every cluster to train a SVM.

3.3 Labelling an Unknown Image

Given an unknown image with only visible poses, we classify the group activity of each person by calculating the maximum likelihood at the group activity and scene/event levels. From Eqs. 4-7, we have the likelihood of the image given the activity labels and scene label,

$$p(e|A, O, \omega, \psi) = \frac{\sum_S p(e|S) p(S|A, \omega) p(A|O, \psi)}{\prod_k p(A_k|O, \psi)} = \frac{\sum_S \prod_k [p(e_k|S_k) \prod_j p(e_k|S_j)^{\delta_{kj}} p(S_k|A_k, \omega) p(A_k|O, \psi)]}{\prod_k p(A_k|O, \psi)}. \quad (8)$$

Then, the decision of the activity labels together with the scene class label can be made based on the maximum likelihood estimation of the image given the activity labels and the scene class label, which is

$$O, A = \arg\max_{O,A} p(e|O, A, \omega, \psi). \quad (9)$$

The maximum likelihood estimation is not tractable computationally and we use the Variational Message Passing algorithm (VMP) [Winn and Bishop, 2005] for approximation.

4 Experiments

4.1 Dataset and Experiment Settings

There are very few public datasets containing images, of which each captures multiple group activities. To compare more approaches, we select the popular Collective Activity Classification (CAC) dataset, on which many excellent approaches have been tested [Choi *et al.*, 2009]. It contains 44 video clips and each person in a video is assigned to one of the five activity categories: crossing, waiting, queuing, walking and talking. We add a scene label to each clip to train our model and there are 3 scene classes in this dataset. The dataset also provides pose category labels, but we try to begin from a low feature level. We use the multi-scale deformable part based model [Felzenszwalb *et al.*, 2008] to detect every body part in the images. By removing meaningless detections, we get the ground truth of the body joints [Zhou *et al.*, 2016]. To better test the performance on multiple group activities, we form a *Special dataset* by selecting 24 of the 44 clips, which contains at least two kinds of activities. On both original CAC dataset and the Special dataset, we select every 10-th frame of the clips to form the image datasets.

We apply 5-fold cross validation to evaluate our model on the formed dataset. The distribution parameters and the SVMs are first learned based on the training data. We use the method in [Pan *et al.*, 2015] to capture the 3-D locations of the persons and use the 3-D distance between two persons to calculate $\vartheta(e_i, e_j)$ in Eq. 6.

To compare as many as possible the existing methods on CAC dataset, we conduct two groups of experiments. First, on the Special dataset, we figure out the activity of each person. Then, on the original CAC dataset, we recognize the majority activity (i.e., the activity performed by most of persons in the image). In these experiments, we learn 6 SVMs of standard poses for each activity and set the ratio of $\alpha : \beta$ to be 1. We also show the experimental results about the effects of the number of SVMs for each activity and the ratio of $\alpha : \beta$.

4.2 Results and Performance Analysis

The overall accuracy and variance and the per-class ones are reported in Table 1. Our model achieves an overall accuracy of 82.07% and a variance lower than 1%. In terms of the accuracies of each class, we have achieved no accuracies lower than 74% and the ones for activities "Queue" and "Talk" have been outstanding. "Queue" and "Talk" get higher accuracies because the features of them are clearer than those of the other three activities. Some methods, e.g., [Lan *et al.*, 2012c], achieve high accuracies of queuing and talking activities by using the facing-direction annotations of the original CAC dataset and the fact that queuing persons usually face to the same direction and talking persons usually face to each other. Comparatively, without these annotations, our model has also successfully learned these patterns. Most of the incorrect classifications happen in the activities "Cross" and "Wait", because they occur in street scenes, which are more complex and have more confusing interference terms, such as "Cross" against "Walk" and "Waiting" against "Talk" or "Walk". Furthermore, none of the classification outcomes shown in Table 1 has variances exceeding 1.6%, so the proposed model presents good robustness for mixed group activity recognition.

The confusion matrix of our model is shown in Table 2. Most incorrect classifications of "Cross" locate at "Walk" (9%), while most incorrect classifications of "Wait" are against 'Walk" (11%) and "Talk" (8%). Since these confusions can also happen in human cognition, our model makes acceptable interpretation for mixed group activities.

Table 1: The overall and per-class accuracies and variances.

Activity	Accuracy and Var
Overall	$82.07 \pm 0.85\%$
Cross	$76.83 \pm 0.22\%$
Wait	$74.36 \pm 1.51\%$
Queue	$93.76 \pm 0.79\%$
Walk	$87.63 \pm 0.59\%$
Talk	$98.16 \pm 1.06\%$
Scene/Event	$94.21 \pm 0.98\%$

Table 2: The confusion matrix of our generative model.

	Cross	Wait	Queue	Walk	Talk
Cross	0.77	0.05	0.02	0.09	0.07
Wait	0.05	0.74	0.02	0.11	0.08
Queue	0.00	0.00	0.95	0.02	0.03
Walk	0.03	0.03	0.01	0.88	0.05
Talk	0.01	0.00	0.00	0.01	0.98

We visualize the classification results of our model in Figure 5. The rightmost image in the second row is a good example that shows the effect of GI and LI. It is hard to tell whether the man in black is walking or standing by looking at only his pose. The pose is more like standing, but the old walking man next to him increases the probability of his walking. Meantime, the girls waiting there suggest higher probability of the two men's performing crossing than walking, because this decision allows the joint probability of their activities and a scene label "Street" to reach a maximum. Examples of mistaken classifications are shown in the third row. In the first image, the two persons at far-end who are walking are classified as "Cross", because their poses cannot provide clear enough information for walking, and so does the walking person in the fourth image. In the second one, the leftmost man is labelled with "Talk" like the two persons who are standing further away, because depth estimation method does not provide a right depth. In the third image, the walking woman is labelled as "Queue" because the location and pose of the lady do not differ much from the other queuing persons. In our future work, we will try to fix them by adding more features.

Figure 3 shows how the number of standard poses corresponding to each activity affects the overall and per-class accuracies. More standard poses provide more selections for an optimum, and hence increase the accuracies to their maximums. When the number of the standard poses reaches a certain extent, very little difference exists between the poses, so adding more standard poses provide no more improvement. Besides, the more complex an activity is, the more standard poses are needed to achieve the optimal accuracy of the activity recognition. For example, "Queue" needs only 3 poses while "Walk" needs 6 ones to achieve the maximum accuracies. All accuracies almost reach to their maximums before or near 6 standard poses and after that, the average time cost to recognize one image increases dramatically and exponentially. Therefore, we learn and use 6 standard poses for each activity in this dataset.

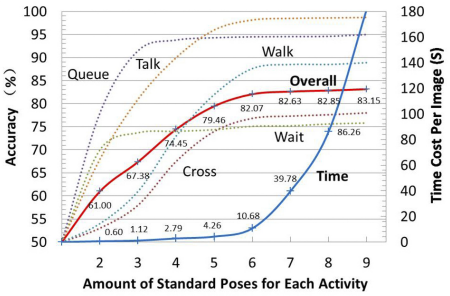

Figure 3: The accuracies and time cost per image on different amounts of standard poses contained by one activity. The blue line illustrates the time consumed and other lines illustrate accuracies.

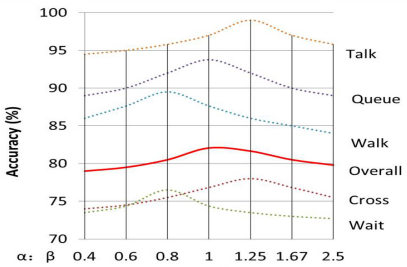

Figure 4: The accuracies with different ratios of $\alpha : \beta$. The higher this ratio is, the higher weights the interactions within a group will have.

The ratio of $\alpha : \beta$ describes a tradeoff between the generation from a person's own standard pose and the effect of the LI. The higher this ratio is, the higher weight the LI will have. As shown in Figure 4, different activities reach the peaks with different ratios. The peak at a bigger ratio means the activity is performed more collaboratively than individually. For example, "Talk" and "Cross" more highly rely on the interactions than "Wait" and "Walk". The influence of the parameters shown in Figures 3 and 4 also provides evidence that our model makes reasonable interpretation for mixed group activities.

4.3 Comparisons with state-of-the-art Methods

There have been some papers working on the CAC dataset and we include 11 methods in total for comparison. For some video-based methods [Choi *et al.*, 2009; 2011; London *et al.*, 2013], we replace their spatio-temporal features with spatial features and run them together with image-based methods on the Special dataset. For the other video-based methods [Choi and Savarese, 2014; Tran *et al.*, 2013], which can not be applied on images, we directly list their accuracies at the top part of Table 3. Their high accuracies benefit from the spatio-temporal features.

The comparisons on the Special dataset are shown in the middle part of Table 3. "Appearance Features" and "Spatial Context" employ different low level features for recognition.

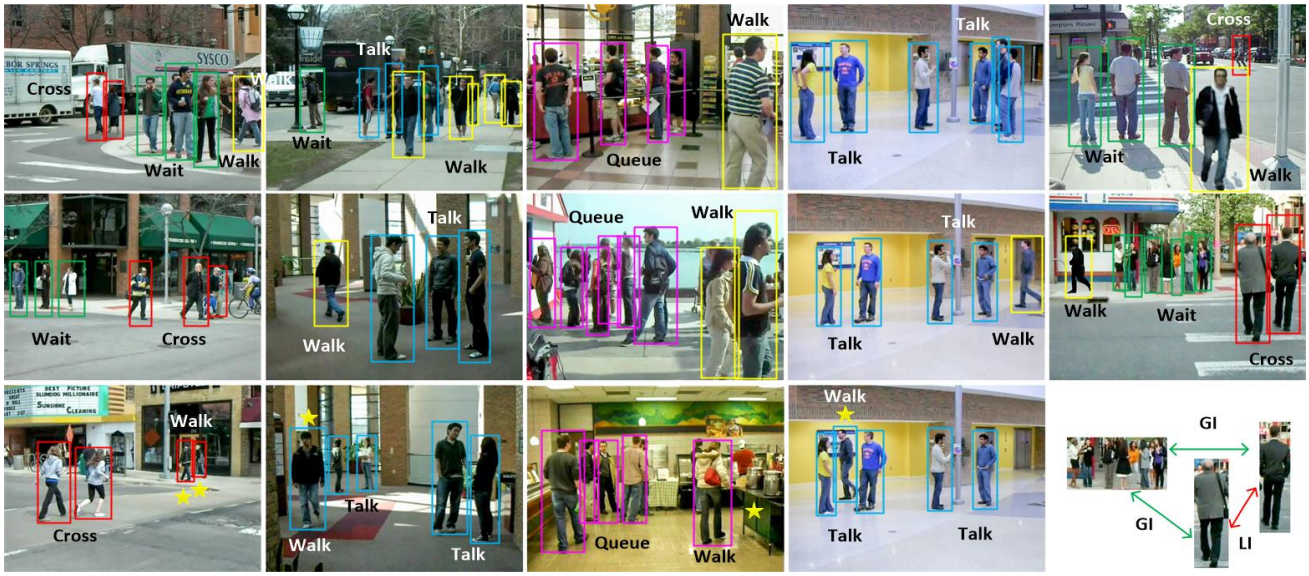

Figure 5: The visualization of the results of our model. Boxes in different colors are used to denote different activities: red for "Cross", green for "Wait", purple for "Queue", yellow for "Walk" and blue for "Talk". Real activity labels are noted by the words in the images. Some incorrect classifications (indicated by little yellow stars) are also shown in the third row.

"STV+MC" uses a SVM classifier to classify STV descriptors. "RSTV+MRF" constructs a random forest structure to learn context for recognition. "HL-MRL+ACD" uses Hinge-loss Markov Random Fields for classification. "Discriminative Context" is a newly proposed method and it utilizes the original action category labels. Even if we train a classifier with an accuracy of 90% to predict these labels with the annotations in our dataset, the accuracy of this method drops from 83.0% to 74.7%. Therefore, these results prove that our model outperforms the state-of-the-art methods for the mixed activity recognition.

We also conduct another group of experiments on the original dataset to compare the methods that only provide one collective activity label for each image. The results are shown in the bottom part of Table 3. "Action Context" uses action features together with some other features, and "Global bag-of-words" is the baseline method mentioned in [Lan *et al.*, 2012c]. "Latent Model" is a discriminative model which performs pretty well under the assumption of only one activity presented in one image. Our model achieves the highest accuracy again. Compared with the result of ours on Special dataset, the higher accuracy of ours here is due to the fact that counting the majority activity actually ignores some mistaken activity labels of individuals.

Table 3: Comparisons with the state-of-the-art methods.

Method	Accuracy
Video-Based Method	
Social Cues[Tran *et al.*, 2013]	78.7%
Inter-Class Context[Choi and Savarese, 2014]	79.0%
Special Dataset	
Appearance Features[Zhao *et al.*, 2014]	60.6%
STV+MC[Choi *et al.*, 2009]	65.9%
HL-MRF+ACD[London *et al.*, 2013]	69.2%
RSTV+MRF[Choi *et al.*, 2011]	70.9%
Discriminative Context[Zhao *et al.*, 2014]	74.7%
Spatial Context[Zhao *et al.*, 2014]	76.6%
Our Method	**82.4%**
Original CAC Dateset	
Action Context[Lan *et al.*, 2012b]	68.2%
Global Bag-of-words[Lan *et al.*, 2010]	70.9%
Latent Model[Lan *et al.*, 2012c]	79.7%
Our Method	**86.2%**

5 Conclusions

In this paper, we have presented a generative model to recognize mixed group activities in still images. This model well interprets both interactions among multiple groups and interactions among multiple persons within the same group. It does not need to work in a multi-step way or rely on basic techniques for individual action recognition. Experimental results and a comprehensive comparison have demonstrated that our model provides good interpretations for multiple mixed group activities and outperforms the state-of-the-art methods on the CAC dataset.

Acknowledgments

This work is supported by National High Technology Research and Development Program of China (No.2013CB329605) (973 Program) and International Graduate Exchange Program of BIT. It is also supported by UTS IRS and Top-up scholarships, and UTS-FEIT scholarship.

References

[Alameda-Pineda *et al.*, 2015] Xavier Alameda-Pineda, Yan Yan, Elisa Ricci, Oswald Lanz, and Nicu Sebe. Analyzing free-standing conversational groups: A multimodal approach. In *Proceedings of ACM Conference on Multimedia*, pages 5–14. ACM, 2015.

[Cheng *et al.*, 2014] Zhongwei Cheng, Lei Qin, Qingming Huang, Shuicheng Yan, and Qi Tian. Recognizing human group action by layered model with multiple cues. *Neurocomputing*, 136:124–135, 2014.

[Choi and Savarese, 2014] Wongun Choi and Silvio Savarese. Understanding collective activitiesof people from videos. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(6):1242–1257, 2014.

[Choi *et al.*, 2009] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Computer Vision-ICCV Workshops*, pages 1282–1289. IEEE, 2009.

[Choi *et al.*, 2011] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *Proceedings of the IEEE Conference on CVPR*, pages 3273–3280. IEEE, 2011.

[Choi *et al.*, 2014] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Discovering groups of people in images. In *Computer Vision–ECCV*, pages 417–433. Springer, 2014.

[Felzenszwalb *et al.*, 2008] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the IEEE Conference on CVPR*, pages 1–8. IEEE, 2008.

[Guo and Lai, 2014] Guodong Guo and Alice Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014.

[Lan *et al.*, 2010] Tian Lan, Yang Wang, Weilong Yang, and Greg Mori. Beyond actions: Discriminative models for contextual group activities. In *Advances in neural information processing systems*, pages 1216–1224, 2010.

[Lan *et al.*, 2012a] Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *Proceedings of the IEEE Conference on CVPR*, pages 1354–1361. IEEE, 2012.

[Lan *et al.*, 2012b] Tian Lan, Yang Wang, Greg Mori, and Stephen N Robinovitch. Retrieving actions in group contexts. In *Trends and Topics in Computer Vision*, pages 181–194. Springer, 2012.

[Lan *et al.*, 2012c] Tian Lan, Yang Wang, Weilong Yang, Stephen N Robinovitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(8):1549–1562, 2012.

[Li and Li, 2007] Li-Jia Li and Fei-Fei Li. What, where and who? Classifying events by scene and object recognition. In *Proceedings of the IEEE Conference on CVPR*, pages 1–8. IEEE, 2007.

[London *et al.*, 2013] Ben London, Shamsul Khamis, Stephen H Bach, Bert Huang, Lise Getoor, and Lisa Davis. Collective activity detection using hinge-loss Markov random fields. In *IEEE Conference on CVPR Worshop*, pages 566–571. IEEE, 2013.

[Odashima *et al.*, 2012] Shigeyuki Odashima, Masamichi Shimosaka, Takuhiro Kaneko, Rui Fukui, and Tomomasa Sato. Collective activity localization with contextual spatial pyramid. In *Computer Vision–ECCV Workshops*, pages 243–252. Springer, 2012.

[Pan *et al.*, 2015] Jiyan Pan, Martial Hebert, and Takeo Kanade. Inferring 3d layout of building facades from a single image. In *Proceedings of the IEEE Conference on CVPR*, pages 2918–2926, 2015.

[Sun *et al.*, 2014] Lei Sun, Haizhou Ai, and Shihong Lao. Activity group localization by modeling the relations among participants. In *Computer Vision–ECCV*, pages 741–755. Springer, 2014.

[Tran *et al.*, 2013] Khai N Tran, Apurva Bedagkar-Gala, Ioannis A Kakadiaris, and Shishir K Shah. Social cues in group formation and local interactions for collective activity analysis. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 539–548, 2013.

[Wang *et al.*, 2015] Limin Wang, Yu Qiao, and Xiaoou Tang. Mofap: A multi-level representation for action recognition. *International Journal of Computer Vision*, pages 1–18, 2015.

[Winn and Bishop, 2005] John M Winn and Christopher M Bishop. Variational message passing. In *Journal of Machine Learning Research*, pages 661–694, 2005.

[Xiong *et al.*, 2015] Yuanjun Xiong, Kai Zhu, Dahua Lin, and Xiaoou Tang. Recognize complex events from static images by fusing deep channels. In *Proceedings of the IEEE Conference on CVPR*, pages 1600–1609, 2015.

[Zhang *et al.*, 2015] Tao Zhang, Wenjing Jia, Baoqing Yang, Jie Yang, Xiangjian He, and Zhonglong Zheng. Mowld: a robust motion image descriptor for violence detection. *Multimedia Tools and Applications*, pages 1–20, 2015.

[Zhao *et al.*, 2014] Chaoyang Zhao, Wei Fu, Jinqiao Wang, Xiao Bai, Qingshan Liu, and Hanqing Lu. Discriminative context models for collective activity recognition. In *ICPR, 2014 International Conference on*, pages 648–653. IEEE, 2014.

[Zhou *et al.*, 2015] Zheng Zhou, Kan Li, and Xiangjian He. Recognizing human activity in still images by integrating group-based contextual cues. In *Proceedings of ACM Conference on Multimedia*, pages 1135–1138. ACM, 2015.

[Zhou *et al.*, 2016] Zheng Zhou, Kan Li, Xiangjian He, and Li Mengmeng. Dataset: http://cs.bit.edu.cn/szdw/jsml/js/lk/index.htm. 2016.

[Zhu *et al.*, 2013] Yingying Zhu, Nandita M Nayak, and Amit K Roy-Chowdhury. Context-aware modeling and recognition of activities in video. In *Proceedings of the IEEE Conference on CVPR*, pages 2491–2498. IEEE, 2013.