Concurrent group activity classification with context modeling

Wei Fu
The 54th Institute of China
Electronics Technology Group
Corporation
Shijiazhuang, Hebei, 050081
weifu.1986@gmail.com

Chaoyang Zhao
National Laboratory of Pattern
Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, 100190
chaoyang.zhao@nlpr.ia.ac.cn

Jinqiao Wang
National Laboratory of Pattern
Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, 100190
jqwang@nlpr.ia.ac.cn

Jing Liu
National Laboratory of Pattern
Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, 100190
jliu@nlpr.ia.ac.cn

Jian Cheng
National Laboratory of Pattern
Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, 100190
jcheng@nlpr.ia.ac.cn

Hanqing Lu
National Laboratory of Pattern
Recognition
Institute of Automation,
Chinese Academy of Science
Beijing, 100190
luhq@nlpr.ia.ac.cn

ABSTRACT

Group activity classification is the task to identify activities with multiple person participation, which often involves in the usage of the context information like person relationships and person interactions. In this paper, we propose a novel approach to jointly model three co-existing cues including the activity duration time, individual action feature and the context information shared between person interactions. Our approach infers group activity labels of all the persons together with their activity durations, especially for the situation with multiple group activities co-existing. Experimental results show that our approach outperform state-of-the-art by 10%.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding - video analysis

Keywords

group activity, context information

1. INTRODUCTION

Recognizing human activities from videos has been a challenging task in the past few years. Most of traditional vision-based activity recognition works have been focused on single-person activities. However, realistic scenes of human activity often involve multiple, inter-related actions at the same time, analysis of a single individual cannot yield reliable results. Take for example those persons in Fig.1(a), we may easily know that the woman is standing and the man is walking by analyzing their low level visual features. But once take in account the context and interaction information among them, it is clear that the woman is "*waiting*" while the man is "*crossing*" the street. Therefore, context modeling is necessary for recognizing these kinds of activities.

To this end, many researchers turn to exploring the context information for analyzing a group of persons' behaviors with interactions among each other , referred as "group activity" or "collective activity" recognition [4, 3, 2, 9, 11]. Lan

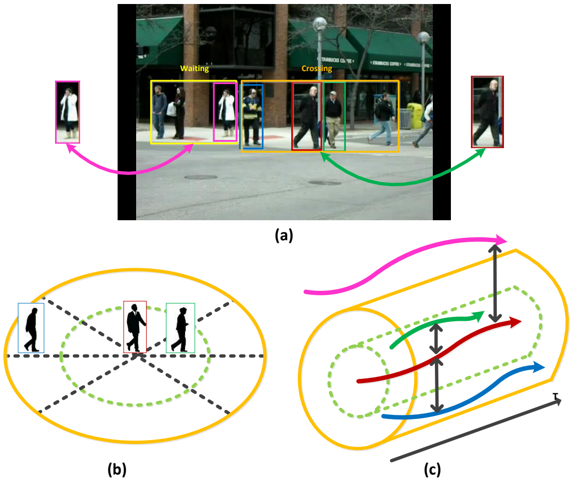

Figure 1: An illustration of concurrent group activities. (a)person interactions with each other; (b)the focal person's activity is influenced by the region context; (c)interactions between trajectories serve as context information.

ICIMCS '15, August 19-21, 2015, Zhangjiajie, Hunan, China
© 2015 ACM. ISBN 978-1-4503-3528-7/15/08...$15.00
DOI: http://dx.doi.org/10.1145/2808492.2808501

et al. used a high level latent discriminative model[7, 9, 8] to explore the group-person interaction and person-person interaction context. Choi *et al.* modeled the crowd context [3] to establish the activities performed by individuals within a crowd. Based on our observation, the existence of the group activity involves the individual action, the context information shared between persons, and the concurrence of the activity duration as well. However, few of former works make use of these kinds of information.

Additionally, persons with different group activity labels may also have influence on each other. The man "*crossing*" the street within a red bounding box in Fig.1(a) is supported both by the fact that persons nearby are all "walking" towards the same direction and the others far away from him are "*waiting*" towards another direction.

In this paper, we propose a concurrent group activity classification approach in which the activity duration time, the individual action feature and the context information between persons are modeled jointly. We also consider the interactions between persons belonging to different group activities. By introducing carefully designed context descriptors, our approach provides strong cues for context information like trajectory interactions and person relationships.

2. CONCURRENT GROUP ACTIVITY REPRESENTATION

Our approach enables analyzing human group activities by looking at context information extracted from the all the persons and their relationships in a video sequence. Given a video sequence, we assume it has been preprocessed so that persons' bounding boxes and their local trajectories can be used directly.

Assuming there are M classes of collective activities in the scene, and the label $y \in \{1, 2, ..., M\}$ denotes the activity class of a person. Let $Y = \{y_i : i = 1, 2, ..., N\}$ be the activity label set for all N persons in an video sequence and $T = \{t_i : i = 1, 2, ..., N\}$ stand for the auxiliary time duration set, where t_i is the group activity duration for the *ith* person. The task is to find the optimal hypothesis label set (Y, T) for all the persons in the scene. Based on this, we now encode the activity context information and develop four potentials to represent the group activity distributions.

Activity-duration potential $w_t^T \Psi_t(y_i, t_i)$. To measure the compatibility between the group activity label y_i and its duration t_i for the *ith* person, the activity-duration potential is parameterized as

$$w_t^T \Psi_t(y_i, d_i) = t_i w_t^T I(t_i) \tag{1}$$

where the indicator $I(t_i)$ is a $t_{max} \times 1$ vector with $(t_i)^{th}$ element marked as one and zeros for others and t_{max} is the maximum duration of activity y_i.

Unary action-activity potential $w_u^T \Psi_u(x_i, y_i, t_i)$. This potential function models the compatibility between the *ith* person's action and its activity label. Features that encode the action information are represented by the individual's pose and average velocity. For each activity label, based on the average HOG [5] feature, we train a 8-class SVM classifier which contains eight pose categories: *right, front-right, front, front-left, left, back-left, back and back-right.* Then the unary action feature is obtained as

$$x_i = (s_{max,i}, pos_i, v_i) \tag{2}$$

where $K = 8$ is the number of pose categories within a activity, s_{max} is the maximum pose classification score, v is the average velocity of the person. pos_i is the pose indicator for the *ith* person in the subregion, which generate a 8×1 vector with one for the $(pos_i)^{th}$ element and zeros otherwise. Then the action-activity potential is parameterized as:

$$w_u^T \Psi_u(x_i, y_i, t_i) = t_i w_u^T \cdot x_i \tag{3}$$

Region context potential $w_c^T \Psi_c(y_i, t_i)$. This potential measures the compatibility between the group activity label of the *ith* person and its relationships with the surrounding persons within the context regions. The context information that capture relationships of the persons within a region is defined as region context feature. Given the *ith* person as the focal person, the defined context regions shows in Fig.1(b), the feature is computed from the persons inside the context region belonging to the same activity group. As illustrated in Fig.1(c), for a video sequence, the context region is extended in time, the activity of the focal person (the red trajectory) is influenced by the persons nearby (the blue and green trajectories). Here we only consider the influences caused by the the persons inside the context region with the same activity label with the focal person. For a person j inside the context region of the focal person, we have the pose and velocity scores $(s_{1j}, ..., s_{Kj}, v_j)$. Supposing that the context region contains M sub-regions, the region context feature is represented as a $2 \times (2K + 1)$ dimensional vector with persons' pose scores, pose histogram and velocity score:

$$\begin{aligned} fc_i =(&\max_{j \in \mathcal{N}_1(i)} s_{kj}, ..., \sum_{j \in \mathcal{N}_1(i)} pos_j, \max_{j \in \mathcal{N}_1(i)} v_j, \\ &\max_{j \in \mathcal{N}_2(i)} s_{kj}, ..., \sum_{j \in \mathcal{N}_2(i)} pos_j, \max_{j \in \mathcal{N}_2(i)} v_j) \end{aligned} \tag{4}$$

where the sub-context region $\mathcal{N}_1(i)$ and $\mathcal{N}_2(i)$ are circles of $0.5h$ and $2h$ (h is the average height of the focal person i) respectively. Then the potential is parameterized as:

$$w_c^T \Psi_c(y_i, t_i) = t_i w_c^T \cdot fc_i \tag{5}$$

Trajectory context potential $w_s^T \Psi_s(x_i, x_j, y_i, y_j, t_i, t_j)$. This potential models compatibility between the group activity labels of *ith* and *jth* person and their spatial and temporal interactions, which can also model the interactions between persons with different group activity labels. These interactions are presented by pairwise interaction features extracted from related trajectories. For two persons i and j (the red and the pink trajectories as shown in Fig.1(c)), we use dynamic time warping (DTW [1]) to measure the distance between two trajectories due to their different start points or time durations. Together with the pose information, the pairwise interaction feature is defined as:

$$fi_{i,j} = [bin(dist_{ij}), pose(i, j))] \tag{6}$$

where $dist_{ij}$ is the DTW distance between two trajectories and is further divided into 3 bins defined as *connected, near* and *far.* And $pose(i, j)$ is defined as $max(\Psi_u(x_i), \Psi_u(x_j))$. Here allow that the interaction features can be extracted from persons with different group activity labels $(y_i \neq y_j)$. Then the trajectory context potential is parameterized as:

$$w_s^T \Psi_s(x_i, x_j, y_i, y_j, t_i, t_j) = (t_i \cap t_j) w_s^T \cdot fi_{i,j} \tag{7}$$

where $(t_i \cap t_j)$ stands for the overlapped time duration.

3. STRUCTURAL MODEL LEARNING

By combining the four potentials with a structural framework, we can measure the compatibility between the label set (Y,T) and all the N persons in a video sequence as:

$$\begin{aligned} &S(X,Y,T) = \omega^T \Psi(X,Y,T) = \\ &\sum_i w_t^T \Psi_t(\cdot) + \sum_i w_u^T \Psi_u(\cdot) + \sum_i w_c^T \Psi_c(\cdot) + \sum_{i,j} w_s^T \Psi_s(\cdot) \end{aligned} \tag{8}$$

Parameters Learning: Given the activity sequences and their structural labels, our goal is to learn the parameter ω in Equ.8, for which we have

$$\begin{aligned} w &= \left[w_t^T, w_u^T, w_c^T, w_s^T\right]^T \\ \Psi(X,Y,T) &= \left[\sum_i \Psi_t(\cdot), \sum_i \Psi_u(\cdot), \sum_i \Psi_c(\cdot), \sum_{i,j} \Psi_s(\cdot)\right]^T \end{aligned} \tag{9}$$

With the training video sequence X_i, and the corresponding label set Y_i and T_i, learning ω can be converted to a regularized learning problem as follows:

$$\begin{aligned} &\arg\min_{w,\xi_i \geq 0} w^T w + C \sum_i \xi_i \\ &s.t. \forall i \quad w^T \Delta\Psi(X_i, Y_i, H_i, T_i, HT_i) \\ &\qquad \geq l(Y_i, H_i, T_i, HT_i) - \xi_i \end{aligned} \tag{10}$$

where $\Delta\Psi(\cdot) = \Psi(X_i, Y_i, T_i) - \Psi(X_i, H_i, HT_i)$, $l(\cdot)$ is the loss function to measure the difference between ground truth and the hypothetical activity label H_i and duration HT_i, and C and ξ_i are the penalty factor and the slack variable respectively. We use the cutting plane optimization algorithm proposed in [6] to solve this problem.

Inference: The inference procedure is to find the best label set Y^* together with time duration T^* for each labeled activity with an input video X. The task is to solve the following optimization problem:

$$(Y,T)^* = \arg\max_{Y,T} S(X,Y,T) \tag{11}$$

The optimum label vectors Y^* and T^* are obtained by a greed search approach as [12]. Although this greedy search algorithm cannot guarantee a globally optimum solution, in practice it works well to find good solutions.

4. EXPERIMENT

Dataset and Settings: We carry out our experiments on the challenging real world dataset [3]. The dataset contains 44 video sequences with all the persons in every 10th frame of the videos are assigned one of the five collective activity categories: *crossing, waiting, queuing, walking.* More than 1/5 of the videos contain two or more activities in the same scene. We use the trajectory labels after our own corrections.

To compare our model with the state-of-the-art approaches, we count the activity labels assigned for each person in every 10th frame and measure the performance by the classification accuracy. 33 video sequences are randomly selected to train the model and the rest are used as the testing set. We repeat the process 10 times and report the average results.

Evaluation of different feature fusion strategies. We first evaluate the performance of several feature fusion strategies. The confusion matrices of the group activity classification accuracy are shown in Fig.2. We can see that the approach without context information yields the worst result as shown in Fig.2(a). By adding the "*action context*"[9], the performance shown in (b) improves more than 30% on average precision, which indicates the positive effect of the context information. Fig.2 (c) presents the further improvement with our region context. From Fig.2 (d) and (f), we can see the importance of *pose*(i,j) in Equ.7. Our trajectory context potential in (f) also outperforms the spatial context [6] in Fig.2 (e). Compared with Fig.2 (g), the classification accuracy in Fig.2 (h) benefits from modeling the inter-group interactions especially when multiple activities co-exist in a scene. Our final model in Fig.2 (h) archives the best result over all the approaches.

Approaches	Average Accuracy (%)
ActionContext model [10]	68.2
RandomForest model [4]	70.9
Latent Model [9]	79.1
Our approach without multiple activities	82.6
Our approach	**89.9**

Table 1: Comparison results with state-of-the-arts.

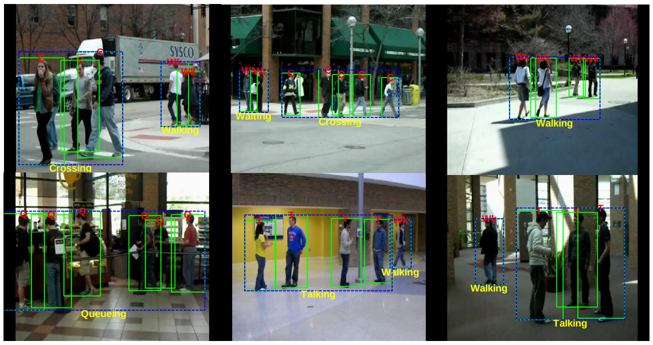

Figure 3: Illustration of group activity recognition results.

Comparison with state-of-the-arts. The comparison results with state-of-the-arts are presented in Tab.1. The "ActionContext model" [10] used the action context feature. The "RandomForest model" [4] used a random forest classifier to model the spatial-temporal information. The "Latent Model" [9] used a hierarchical latent model to formulate the group activity. The "Our approach without multiple activities" stands for the model described in Fig.2(g). Its result already outperforms the state-of-the-art, which suggests the effectiveness of our designed concurrent context descriptors. By considering the situation of multiple group activities co-existing, our approach outperforms state-of-the-art approaches by 10%. Fig.3 illustrates some intuitional results, in which persons are labeled by their group activities.

5. CONCLUSION

In this paper, we have presented a novel approach to recognize group activities. By formulating the activity time durations, the individual action features and the trajectory interactions jointly, our concurrent activity model exploits the effective context information, especially for the situations of multiple activities co-existing scenes. Experimental

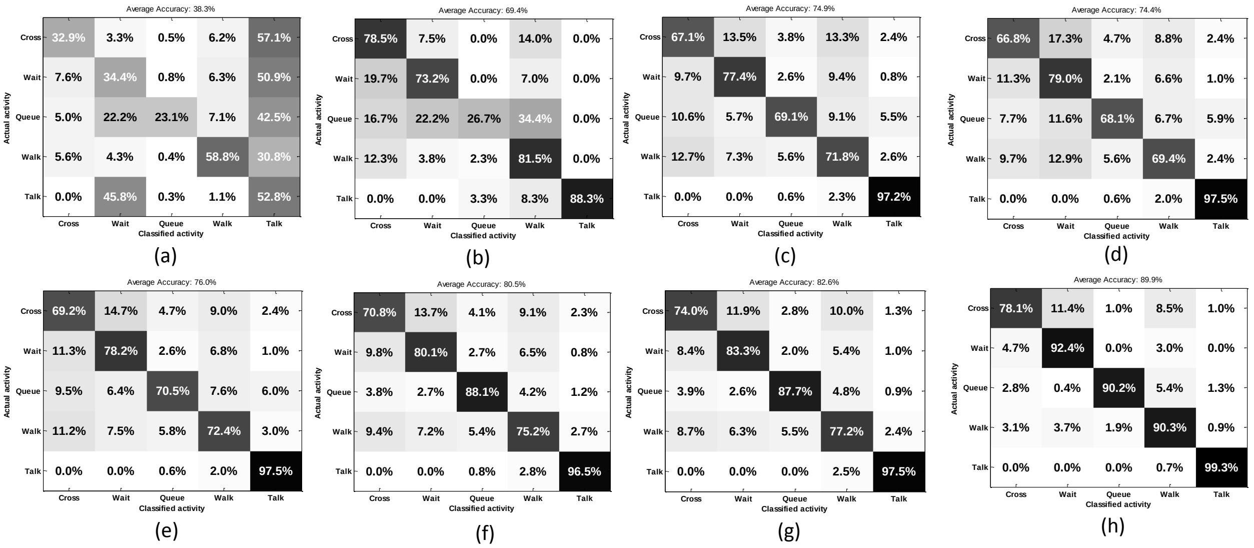

Figure 2: Confusion matrices for activity classification accuracy with different feature fusion strategies: (a) duration and unary feature; (b) duration, unary feature and action context in [10]; (c) duration, unary feature and region context; (d) duration, unary feature and trajectory context with $pose(i,j)$ in Equ.7; (e) duration, unary feature and spatial context in [6]; (f) duration, unary feature and trajectory context; (g) duration, unary feature, region and trajectory context without multiple activities co-existing; (h) duration, unary feature, region and trajectory context with multiple activities co-existing.

results demonstrate that our proposed model improves the performance significantly.

6. ACKNOWLEDGEMENT

This work was supported by 863 Program 2014AA015104, and National Natural Science Foundation of China 61273034, and 61332016.

7. REFERENCES

[1] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370, 1994.

[2] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision (ECCV)*, 2012.

[3] W. Choi, K. Shahid, and S. Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1282–1289, 2009.

[4] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3273–3280, 2011.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[6] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *International journal of computer vision*, 95(1):1–12, 2011.

[7] T. Lan, L. Sigal, and G. Mori. Social roles in hierarchical models for human activity recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1354–1361, 2012.

[8] T. Lan, Y. Wang, G. Mori, and S. N. Robinovitch. Retrieving actions in group contexts. In *European Conference on Computer Vision (ECCV), 2010*, 2010.

[9] T. Lan, Y. Wang, W. Yang, S. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(8):1549–1562, 2012.

[10] T. Lan, W. Yang, Y. Wang, and G. Mori. Beyond actions: Discriminative models for contextual group activities. In *In Advances in Neural Information Processing Systems*, 2010.

[11] S. Odashima, M. Shimosaka, and T. Kaneko. Collective activity localization with contextual spatial pyramid. In *European Conference on Computer Vision (ECCV)*, 2012.

[12] Y. Zhu, N. M. Nayak, and A. K. Roy-Chowdhury. Context-aware modeling and recognition of activities in video. *CVPR*, 2013.