Collective Activity Localization with Contextual Spatial Pyramid

Shigeyuki Odashima, Masamichi Shimosaka, Takuhiro Kaneko, Rui Fukui, and Tomomasa Sato

The University of Tokyo, Tokyo, Japan
{odashima,simosaka,kaneko,fukui,tsato}@ics.t.u-tokyo.ac.jp

Abstract. In this paper, we propose an activity localization method with contextual information of person relationships. Activity localization is a task to determine "who participates to an activity group", such as detecting "walking in a group" or "talking in a group". Usage of contextual information has been providing promising results in the previous activity recognition methods, however, the contextual information has been limited to the local information extracted from one person or only two people relationship. We propose a new context descriptor named "contextual spatial pyramid model (CSPM)", which represents the global relationships extracted from the whole of activities in single images. CSPM encodes useful relationships for activity localization, such as "facing each other". The experimental result shows CSPM improve activity localization performance, therefore CSPM provides strong contextual cues for activity recognition in complex scenes.

1 Introduction

Recognizing human activities from images has been a challenging task. Since the most of traditional vision-based human activity recognition works have been focused on single-person activities (e.g. [1, 2]) , several recent works [3–7] are tackling for activities with multiple-people interactions (called "collective activity"). The collective activities are such as "crossing the road", "queuing" and so on. The most of the former works have focused on image/video sequence classification task [4, 7], which determines the particular image or video sequence contains the activity or not, and several works focus on activity classification task on each person [3–6]. However, the former works do not handle "who participates in the same activity group", such as "talking in two groups" (Figure 1(a)). The task of "detecting and localizing collective activities" is able to be formulated as a form similar to the object localization tasks [8, 9]. In this paper, we focus on the collective activity localization.

Collective activity recognition (including collective activity classification and localization) is difficult because sometimes people in different activity have similar appearance (Figure 1(b)). Therefore, the most of former approaches use contextual information. The former works can be categorized into following two approaches: by unary relationships [3,4,6,7] and by pairwise relationships [4,5,10]

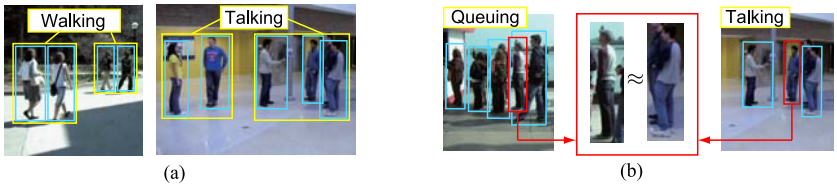

Fig. 1. Collective activity localization. (a) Examples of activity groups. Though individual activity are the same, people belong to different groups (left: two walking groups, right: two talking groups). Activity localization is a task to determine these activity groups. (b) Difficulty of collective activity recognition. Though "queuing" activity and "talking" activity are different activity, people in these activities look similar if you see individual people.

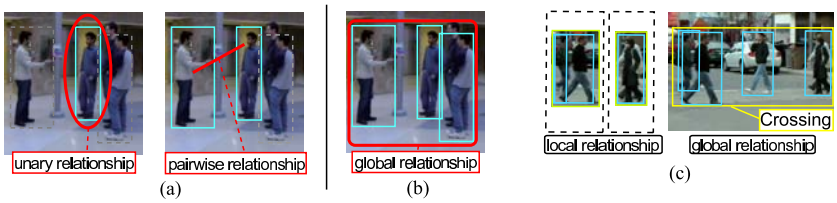

Fig. 2. Relationships in collective activities. (a) Local relationships and (b) global relationship in the collective activity. Though the former works mainly use local relationships, the proposed method focuses global relationship extracted from the whole of single activities. (c) An example of global relationship in collective activities. Since "crossing" activity is locally seen as "walking" activity, distinguishing crossing from walking only by local relationship is hard because of this ambiguity. On the other hand, global relationship provides several important cues which is difficult to acquire by local relationships (e.g. in "crossing" activity, people in "walking" to opposite direction exist on the both side of the activity group).

(Figure 2(a)). Unary relationships are information extracted from the near region of a person or image point, such as how people near a focused person look like. Pairwise relationships are information of two focused people, such as in "queuing" activity, people are facing in same direction. However, these approaches use only local information, i.e. one-person or two-people relationships, so global information from whole participants of the activity has been ignored (Figure 2(b)). The global information extracted from whole activity participants would be useful for activity localization, for example, "talking" activity needs "facing each other". At the same time, the global relationships are useful for recognizing activities: though a "crossing" activity is locally seen as "walking" activity, "crossing" and "walking" can be distinguished by using person layout of the whole of an activity (Figure 2(c)). This paper introduces a new context descriptor named "contextual spatial pyramid model (CSPM)", which extracts global relationships from whole participants in single activity. Thanks to its representation similar to spatial pyramids [11], CSPM is able to encode global people layout in an activity (e.g. "facing each other"), therefore CSPM provides a strong cue for activity localization. We show how CSPM improves activity localization performance with the experimental results.

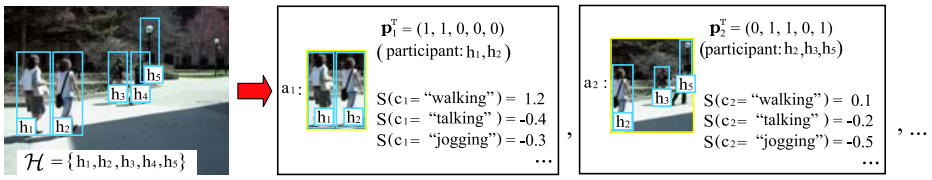

Fig. 3. Activity localization model. An activity window is defined by participants of the activity. The proposed method enumerates possible activity windows composed by subsets of detected people, and assigns score to the activity windows.

1.1 Related work

Contextual information is widely used in object detection [12], human-object interaction recognition [13–15] and collective activity recognition [3, 4, 6, 7, 10]. Though the contextual information is useful for recognition, the most of previous works have focused only on local relationships. The methods with whole activity appearance [14, 16] are close to our approach. However, though the methods use only whole appearance information as global relationship, our method is explicitly able to include much richer contextual information, such as "facing each other".

Though pairwise relationship is widely used as an contextual information (e.g. [4, 10, 12, 13, 15]), a recent work [4] reports that not all possible pairwise relationships are useful for activity recognition. Therefore, several recent works estimate the hidden structure of person relationships to improve activity classification accuracy [4, 5]. These hidden structure can be interpreted as "who are related in the scene", i.e. "who participates in the same activity" or "who are related in the target activity". Our method estimates the structure by detecting activity groups and by extracting sufficient relationships from the whole of an activity group, therefore the proposed method can be regarded as an extension of these structure-inference based methods.

Here we summarize the main contributions of this paper. (1) Collective activity *localization*: the most of works in collective activity recognition have focused on image-level or single-person activity classification. Since several works [6, 7] mention activity localization in collective activity recognition tasks, they do not give localization performance evaluations. We present an activity localization model from multiple person detections, also we give performance evaluation of activity localization methods. (2) A new context descriptor extracted from the whole of an activity: we present a new context descriptor named contextual spatial pyramid model (CSPM). CSPM provides rich contextual cue for activity localization (e.g. person layout in the whole of an activity).

2 Modeling activity localization

In this section, we introduce our activity localization model. Inspired by sliding window classifiers, the proposed method enumerates possible person groups (activity windows), and assigns scores to them (Figure 3).

The method first detects people in the image (Felzenszwalb's object detectors [9] are employed in our experiment). Assume N_p people are detected in the image. We write the k^{th} detected person as $h_k \in \mathcal{H}$, where $\mathcal{H}$ represents set of detected people. The location l_{h_k} of h_k is given by a rectangle on the image, i.e. its left, top, right, bottom $(x_{l,h_k}, y_{t,h_k}, x_{r,h_k}, y_{b,h_k})$.

The i^{th} activity window a_i is defined by selecting activity participants from $\mathcal{H}$. The activity participants $\mathbf{p}_i$ of the activity window a_i can be written as $\mathbf{p_i}^\mathsf{T} = (p_{i1}, p_{i2}, ..., p_{iN_p})$. $p_{ik} \in \{1, 0\}$ is the indicator variable, which represents the k^{th} person h_k is the participant of activity window a_i or not. The location l_{a_i} of i^{th} activity window a_i is defined as the rectangle which surrounds the all participants of a_i.

Activity localization is the task to compute score that an activity window belongs to an activity category c. Our method computes scores by features extracted from each person (unary features) and features extracted from the whole of the activity (group features). The score $S(c_i = c)$ when the activity window a_i belongs to an activity category c is computed as follows:

$$S(c_i = c) = \sum_{k \in \mathcal{P}_i} \mathbf{w}_u(c)^\mathsf{T} \phi_u(h_k) + \mathbf{w}_g(c)^\mathsf{T} \phi_g(a_i) \tag{1}$$

where $\phi_u(h_k)$, and $\phi_g(a_i)$ depict the unary features of the person h_k and the group features of the activity window a_i, respectively. $\mathcal{P}_i \in \{1, ..., N_p\}$ is the set of indices where $p_{ik} = 1$. Roughly speaking, the first term of Eq.(1) represents the appearances of people in an activity group, and the second term represents the global relationships in an activity group. To detect activities on a image, the method enumerates the activity windows with scores over a threshold in each activity category.

Eq.(1) can be rewritten as a linear SVM form (e.g. [17]):

$$S(c_i = c) = \mathbf{w}(c)^\mathsf{T} \phi(a_i) \tag{2}$$

where $\mathbf{w}(c)^\mathsf{T} = (\mathbf{w}_u(c)^\mathsf{T}, \mathbf{w}_g(c)^\mathsf{T})$ and $\phi(c)^\mathsf{T} = (\sum_{k \in \mathcal{P}_i} \phi_u(h_k)^\mathsf{T}, \phi_g(a_i)^\mathsf{T})$.

Implementation: By the scoring procedure, we usually get multiple overlapping detections for each instance of an activity. We apply greedy nonmaximum suppression procedure [9] for activity windows in the same activity category with 50% over overlap.

Our method needs to compute $N_c(2^{N_p} - 1)$ scores when N_c activity categories are defined. This computation is NP-hard. However, the person detector is reliable and N_p is not so large in our application, so we can enumerate all activities in our study. Also, we empirically find that the effect of the maximum N_p is low if the value is not too small, because detecting all people in the crowded situation is infeasible due to occlusions. We set $\max N_p = 10$ for computational efficiency. Note that search techniques such as branch-and-bound or $\mathbf{A}^*$ would be able to applied for general cases.

3 Contextual feature descriptors

In this section, we introduce contextual feature descriptors of the proposed method. To compute scores, we extract unary features and group features from activity participants. Rather than directly using certain raw features (e.g. HOG features [8]), we use contextual features, i.e. action classification scores of each person, etc. Action denotes a simple, atomic posture performed by a single person (e.g. standing and facing right, etc.). Action classification scores are computed by pre-trained SVM classifier based on HOG features extracted from detected people's bounding boxes.

Unary features: $\phi_u(h_k)^{\mathsf{T}} = (\phi_u^{\mathrm{a}}(h_k)^{\mathsf{T}}, \phi_u^{\mathrm{pd}}(h_k)^{\mathsf{T}}, 1)$. $\phi_u^{\mathrm{a}}(h_k)$ is a feature generated by action scores, $\phi_u^{\mathrm{pd}}(h_k) \in \mathbb{R}$ is person detection score of the person detector [9], 1 is bias term.

In this work, we employ 2 types of features as $\phi_u^{\mathrm{a}}(h_k)$. The first feature is bag-of-word style feature $\phi_u^{\mathrm{bow}}(h_k) \in \mathbb{R}^K$ [4], where K is the number of action categories. $\phi_u^{\mathrm{bow}}(h_k)$ of the person h_k is computed as follows:

$$\phi_u^{\mathrm{bow}}(h_k)^{\mathsf{T}} = (S_{1k}, ..., S_{Kk}) \tag{3}$$

where $S_{ik} \in \mathbb{R}$ represents person h_k's classification score of i^{th} action. $\phi_u^{\mathrm{bow}}(h_k)$ represents the focal person's posture information by histogram representation.

The second feature is action context (AC) descriptor [4] $\phi_u^{\mathrm{ac}}(h_k) \in \mathbb{R}^{3K}$ in an image. The original AC descriptor encodes both of spatial information and temporal information, we employ spatial information only, to detect activities in each image independently. $\phi_u^{\mathrm{ac}}(h_k)$ of a person h_k is computed as follows:

$$\begin{aligned}\phi_u^{\mathrm{ac}}(h_k)^{\mathsf{T}} = (S_{1k}, ..., S_{Kk}, &\max_{m \in \mathcal{N}_1(h_k)} S_{1m}, ..., \max_{m \in \mathcal{N}_1(h_k)} S_{Km}, \\ &\max_{m \in \mathcal{N}_2(h_k)} S_{1m}, ..., \max_{m \in \mathcal{N}_2(h_k)} S_{Km})\end{aligned} \tag{4}$$

where $\mathcal{N}_1(h_k)$ and $\mathcal{N}_2(h_k)$ are "sub-context regions" of k^{th} person (in this work, we define $\mathcal{N}_1(h_k)$ and $\mathcal{N}_2(h_k)$ as circles of 0.5h and 2h respectively (h is person h_k's height), according to Lan's parameter [4]). $\phi_u^{\mathrm{ac}}(h_k)$ can capture the information of people nearby as well as the focal person's posture information.

Group features: $\phi_g(a_i)^{\mathsf{T}} = (\phi_g^{\mathrm{cspm}}(a_i)^{\mathsf{T}}, 1)$. $\phi_g^{\mathrm{cspm}}(a_i) \in \mathbb{R}^{KN_{\mathrm{cd}}}$ represents features in activity window a_i extracted by contextual spatial pyramid model. N_{cd} is the number of subregions of CSPM. Figure 4(a) represents an overview of CSPM. To handle people layouts, $\phi_g^{\mathrm{cspm}}(a_i)$ has representations similar to spatial pyramids [11]. $\phi_g^{\mathrm{cspm}}(a_i)$ represents action layout in the activity window a_i by computing bag-of-words like features in the subregions (e.g. in "talking" activity, right-facing persons are on the left side and left-facing persons are on the right side). $\phi_g^{\mathrm{cspm}}(a_i)$ is computed as the following average-pooling representation:

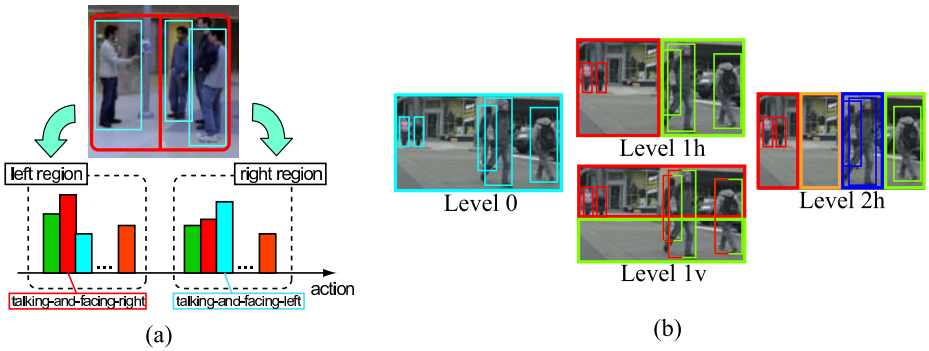

Fig. 4. Contextual spatial pyramid model (CSPM). (a) An overview of CSPM. CSPM encodes global relationships of an activity, by extracting action scores of participants in the subregions. For example, if "talking-and-facing-right" score is high in the left region and "talking-and-facing-left" score is high in the right region, the overall feature represents "talking and facing each other". (b) Spatial pyramid representations. In this work, CSPM takes regions with different separation level (total $N_{\mathrm{cd}} = 9$ subregions).

$$\phi_g^{\mathrm{cspm}}(a_i)^{\mathsf{T}} = (\frac{1}{M_{\mathcal{R}_1}} \sum_{m \in \mathcal{R}_1(a_i)} S_{1m}, ..., \frac{1}{M_{\mathcal{R}_1}} \sum_{m \in \mathcal{R}_1(a_i)} S_{Km}, ..., \frac{1}{M_{\mathcal{R}_{N_{\mathrm{cd}}}}} \sum_{m \in \mathcal{R}_{N_{\mathrm{cd}}}(a_i)} S_{1m}, ..., \frac{1}{M_{\mathcal{R}_{N_{\mathrm{cd}}}}} \sum_{m \in \mathcal{R}_{N_{\mathrm{cd}}}(a_i)} S_{Km}) \quad (5)$$

where $\mathcal{R}_j(a_i)$ is the j^{th} subregion in the spatial pyramid and $M_{\mathcal{R}_j}$ is the number of people in the j^{th} subregion. The proposed method regards the person h_k is in the subregion $\mathcal{R}_j(a_i)$ if h_k participates the activity window a_i (i.e. $p_{ik} = 1$) and if h_k's bounding box intersects subregion $\mathcal{R}_j(a_i)$. If the subregion $\mathcal{R}_j(a_i)$ contains no people, the bin values of the subregion are set to zero.

$\phi_g^{\mathrm{cspm}}(a_i)$ is generated by extracting actions of participants in each subregion, so each bin value of CSPM represents global relationships of an activity group. For example, if the participants of an activity group are globally "facing each other", the bin values of "facing-right" in the left region and "facing-left" in the right region will be high. Therefore, CSPM descriptor can encode global interactions between people in an activity group.

Figure 4 shows the spatial pyramid representation of $\phi_g^{\mathrm{cspm}}(a_i)$. $\phi_g^{\mathrm{cspm}}(a_i)$ takes subregions from level 0, 1h, 1v and 2h ($N_{\mathrm{cd}} = 9$).

4 Experiment

We demonstrate our method on the extended version [6] of the collective activity dataset [3]. The dataset contains 72 annotated video clips acquired by low resolution hand held cameras. In the original dataset, all the people in every tenth frame of the videos are assigned one of the following seven activity categories:

Table 1. Per-class and mean average precision (AP) scores on the collective activity dataset. Left: baseline with bag-of-words style features (BoWS), right: baseline with action context (AC) descriptors. The **bold** scores represent best scores in each baseline setting, the ***italic*** scores represent lower scores than baseline scores. CSPM improves mean AP scores in both of BoWS and AC feature settings.

Class	BoWS	BoWS + CSPM
Crossing	0.090	**0.104**
Dancing	0.215	**0.697**
Jogging	0.426	**0.429**
Queuing	0.115	**0.216**
Talking	0.107	**0.381**
Waiting	0.065	**0.097**
Walking	0.020	**0.053**
Mean AP	0.148	**0.282**

Class	AC	AC + CSPM
Crossing	**0.144**	*0.099*
Dancing	0.353	**0.734**
Jogging	**0.439**	*0.430*
Queuing	0.044	**0.122**
Talking	0.046	**0.087**
Waiting	0.093	**0.125**
Walking	0.021	**0.046**
Mean AP	0.163	**0.235**

crossing, waiting, queuing, walking, talking, dancing and jogging, and one of the following eight pose categories: right, front-right, front, front-left, left, back-left, back and back-right. Following Lan's definition [4], we define 56 action labels (7 activity labels $\times$ 8 pose labels) by combining the pose and activity information, i.e. the action labels include crossing and facing right, crossing and facing front-right, etc. Note that actions are intermediate outputs: action classification scores are used only for feature descriptions. We define ground truth activities on each image, by assigning people participating to the activity and the activities' category. We select one fourth of the video clips to form the test set, and the rest of the video clips are used for training (total 2943 training images and 882 test images). Following PASCAL VOC Challenge's localization criteria [18], the detected activity is considered as a correct detection if the overlap ratio between the bounding box of detected activity and the bounding box of ground truth activity exceeds 50%.

Results: To evaluate our feature model, we compare localization accuracy with several feature settings: unary features only (with bag-of-words style features (ϕ_g^{bow}) and with AC features (ϕ_g^{ac}) : variant of [4]) and unary features and CSPM (with ϕ_g^{bow} and with ϕ_g^{ac}). We define "unary features only" feature set (i.e. local relationships only) as baseline, and evaluate effectiveness of CSPM.

We compute precision-recall curves and the average precision (AP) scores across activity classes. We show the precision-recall curves in Figure 5 and the comparison of AP scores in Table 1. As seen in Table 1, the proposed CSPM descriptor improves localization performance in all activity categories when bag-of-words style features are used as unary features (Table 1 left), and improves localization performance in 5 activity categories when action context descriptors are used as unary features (Table 1 right). Also, CSPM descriptor improve mean AP scores of activity categories in both baseline settings, therefore this result shows CSPM provides useful cues for activity localization.

In the all feature settings, methods with CSPM record highest AP scores in 5 activity categories (with BoWS, 3 categories: queuing, talking, walking, and with

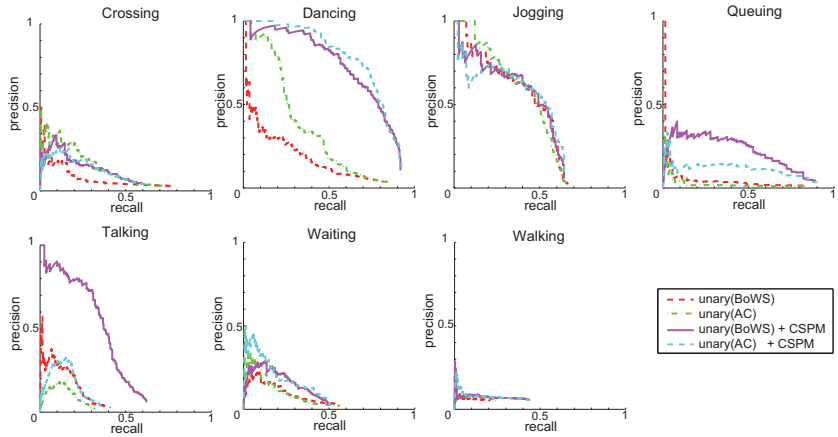

Fig. 5. Precision-recall curves (best viewed in color). Note that several results (crossing, waiting, walking) are shown in different scale for readability.

AC, 2 categories: dancing, waiting), and the method with AC descriptor only records highest AP scores in 2 activity categories (crossing, jogging). Though AP score decreases with CSPM descriptor in several activity categories when the AC descriptor is used as baseline, it is because that AC and CSPM represent different information in the activity groups. Though AC descriptor is a strong descriptor (especially people in the scene participate in single activity, such as "crossing" activity), the AC descriptor encodes person relationships regardless of the activity group, therefore the AC descriptor may encode relationships inconsistent with CSPM descriptors. More efficient feature combination is one of the future works. We visualize the localization results in Figure 6 (correct detections) and Figure 7 (false detections).

Though the proposed method of mean AP scores (roughly 26%) is lower than precision scores on the state-of-the-art result of *person*-level classification or *image*-level classification, it is because that activity localization task is much more difficult. The localization task needs determining the "people in the same activity" in addition to determining activity labels (e.g. in Figure 7, detection results of crossing, dancing, jogging, talking are treated false detections due to localization failure though the detected activities contain ground truth activities), so the activity localization task is a difficult task compared with activity *classification* task.

5 Conclusion

This paper has described a novel activity localization method with a new context descriptor named contextual spatial pyramid model (CSPM). CSPM encodes rich global relationships in an activity (such as "facing each other"), with its spatial-pyramid-like representation. The experimental result shows CSPM provides useful relationships to improve activity localization performance.

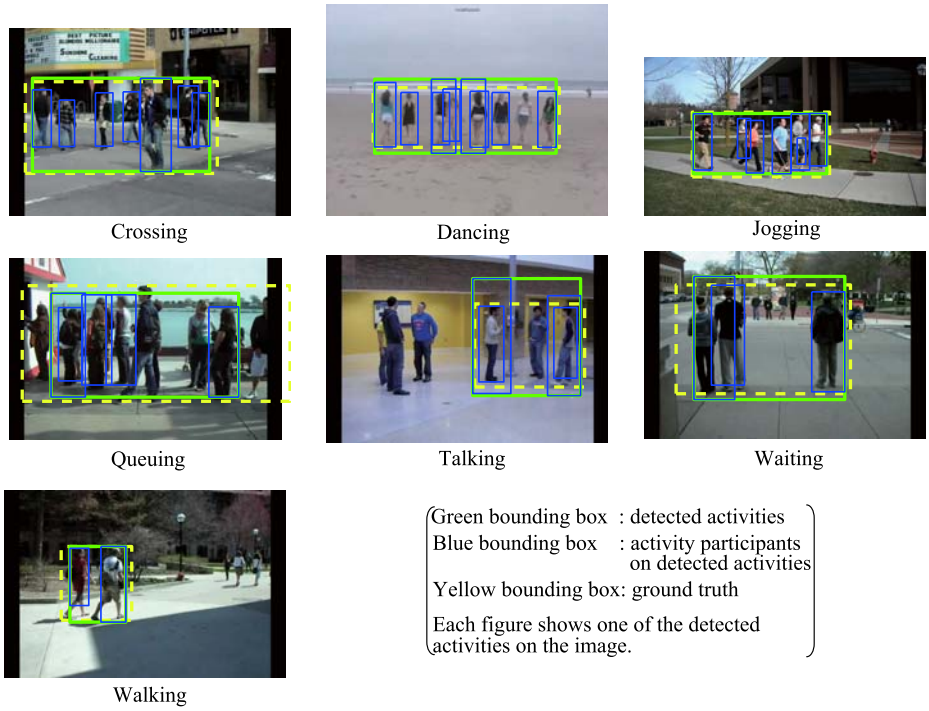

Fig. 6. Examples of correct activity localization results with BoWS + CSPM feature setting (best viewed in color).

References

1. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. IJCV **79** (2008) 299–318
2. Wang, Y., Mori, G.: Human action recognition by semilatent topic models. IEEE Trans. on PAMI **31** (2009) 1762–1774
3. Choi, W., Shahid, K., Savaese, S.: What are they doing?: collective acitivity classification using spatio-temporal relationship among people. In: International Workshop on Visual Surveillance. (2009)
4. Lan, T., Wang, Y., Yang, W., Robinovitch, S.N., Mori, G.: Discriminative latent models for recognizing group activities. IEEE Trans. on PAMI **34** (2012) 1549–1562
5. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: CVPR. (2012)
6. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: CVPR. (2011)
7. Amer, M.R., Todorovic, S.: A chains model for localizing participants of group activities in videos. In: ICCV. (2011)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
9. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. on PAMI **32** (2010) 1627–1645

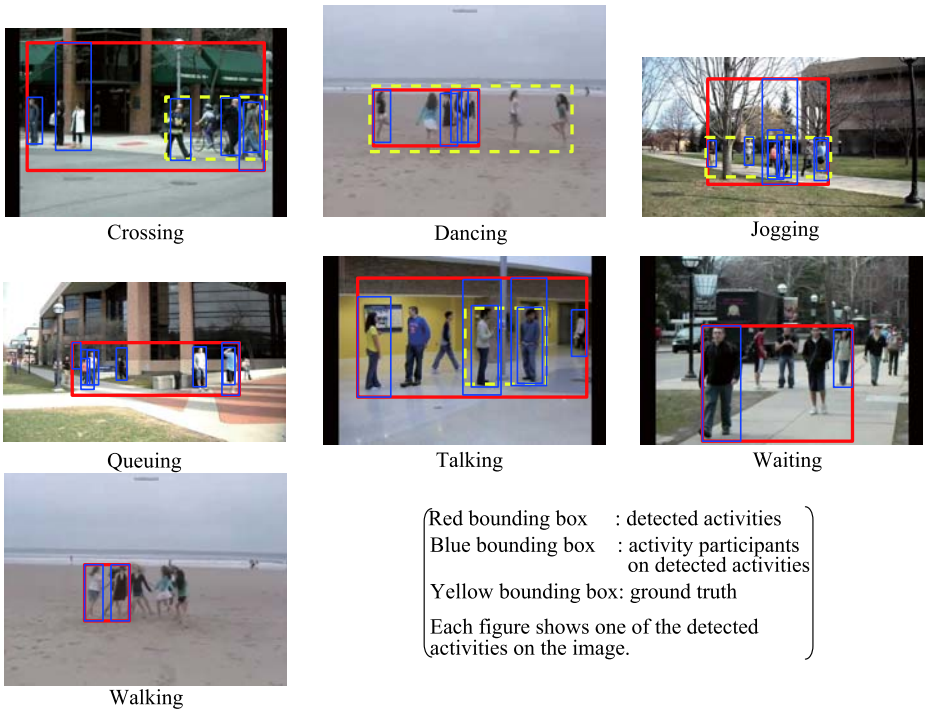

Fig. 7. Examples of failed activity localization results with BoWS + CSPM feature setting (best viewed in color). Several detection results are treated failed detections due to localization failure (crossing, dancing, jogging, talking).

10. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In: ICCV. (2009)
11. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
12. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. In: ICCV. (2009)
13. Gupta, A., Kembhavi, A., Davis, L.: Observing human-object interactions: using spatial and functional compatibility for recognition. IEEE Trans. on PAMI **31** (2009) 1775–1789
14. Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: CVPR. (2011)
15. Yao, B., Khosla, A., Fei-Fei, L.: Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses. In: ICML. (2011)
16. Amer, M.R., Todorovic, S.: Sum-product networks for modeling activities with stochastic structure. In: CVPR. (2012)
17. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. JMLR **9** (2008) 1871–1874
18. Everingham, M., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. IJCV **88** (2010) 303–338