Viewpoint Invariant Collective Activity Recognition with Relative Action Context

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

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

Abstract. This paper presents an approach for collective activity recognition. Collective activities are activities performed by multiple persons, such as queueing in a line and talking together. To recognize them, the action context (AC) descriptor [1] encodes the "apparent" relation (e.g. a group crossing and facing "right"), however this representation is sensitive to viewpoint change. We instead propose a novel feature representation called the *relative action context (RAC) descriptor* that encodes the "relative" relation (e.g. a group crossing and facing the "same" direction). This representation is viewpoint invariant and complementary to AC; hence we employ a simplified combinational classifier. This paper also introduces two methods to accelerate performance. First, to make the contexts robust to various situations, we apply post processes. Second, to reduce local classification failures, we regularize the classification using fully connected CRFs. Experimental results show that our method is applicable to various scenes and outperforms state-of-the art methods.

1 Introduction

Collective activity recognition is one of the most challenging tasks in computer vision. Since collective activities (e.g. queueing in a line, talking together or waiting by a street intersection) are performed by multiple persons, it is often hard to differentiate them only by appearance of the individual. Hence, recent works exploit the contextual information of nearby people [1–8].

When exploiting the contextual information of nearby people, it is required to answer the following question: "How to describe human relationship?" To answer the question, the action context (AC) descriptor [1] represents "apparent" relation (e.g. two persons who are queuing and facing "left" as shown in Figure 1). Such an apparent relation descriptor is suitable when appearance is specific to the target activity. For example, a waiting group is more likely observed from an anterior view rather than from a right view in an image. However, an apparent relation descriptor is sensitive to viewpoint change. To solve the problem, we develop a novel "relative" relation descriptor called the *relative action context (RAC) descriptor*. A relative relation descriptor encodes relative relation (e.g. two persons who are queuing and facing the "same" direction as illustrated in Figure 1), therefore, it contains invariance under viewpoint change

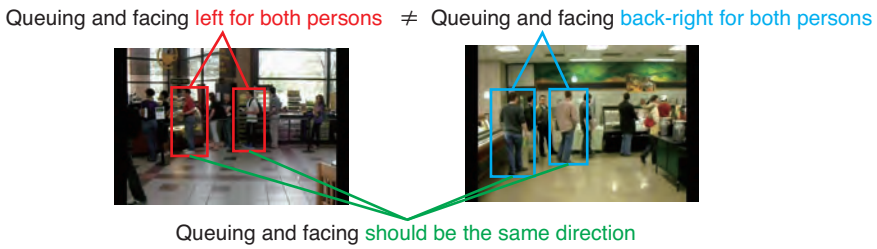

Fig. 1. How to describe human relationship? There are two approaches to describe human relationship: an apparent relation descriptor that encodes apparent relation in an image (e.g. queuing and facing "left", queuing and facing "back-right"), and a relative relation descriptor that encodes relative relation (e.g. queuing and facing the "same" direction). The former can encode appearance specific to the target activity, however, it is sensitive to viewpoint change, while the latter is robust to viewpoint change as well as consistent within the same category of collective activity.

as well as consistency within the same category of collective activity. Note that Choi *et al.* [4, 5] also propose relative relation descriptor, however, these methods exploit only poses (e.g. facing right) rather than actions (e.g. "talking" and facing right); hence they cannot encode apparent differences between activities. Furthermore, since AC and RAC descriptors represent human relation from a different standpoint and they are complementary, we employ a simplified combinational classifier, so as to obtain stable performance in various scenes.

We also introduce two methods to accelerate performance. First, to make the contexts robust to various situations, we apply the following two post processes: threshold processing and Gaussian filtering. When extracting a histogram-style context, Yang *et al.* [9] use sparse coding that not only allows the representation to capture salient properties of images but also achieves lower quantization error, so as to outperform the recent works [10, 11]. Inspired by [9], we use threshold processing to extracts salient properties from noisy contexts, and employ Gaussian filtering to relax quantization errors. Notice that recent works [9–11] use unsupervised histogram derived from the bag-of-words representation, while our work uses supervised histogram calculated by a multiclass classifier.

Second, we employ fully connected CRFs [7, 12, 13] to obtain the consistency in a group. Unlike recent works that optimize collective activity recognition via graph structures [2–6], our model assumes that all the persons in a frame are related, and describes their relationship as potentials that vary depending on the scale of features, in order to handle various group shapes. Our model has similarities to [7], however differs from it in exploiting the data in only the current frame rather than those in the entire video, so as to apply online applications.

In summary, the contributions of this paper are 1) to develop a novel relative relation descriptor called the *relative action context (RAC) descriptor* that is invariant under viewpoint change; 2) to employ a simplified combinational classifier of AC and RAC descriptors to obtain stable performance in various scenes;

3) to make the contexts robust to various situations using simple post processes; 4) to obtain robustness to local classification failures using fully connected CRFs that assume all the relationships among the people in a frame. Experimental results show that our proposed method not only applies to various scenes but also outperforms state-of-the art methods [1, 4, 5, 7].

2 Group Context Descriptor

This section first explains an apparent relation descriptor (section 2.1), then presents a novel relative relation descriptor called the *relative action context (RAC) descriptor* (section 2.2), and finally examines how to combine the apparent relation descriptor and the relative relation descriptor (section 2.3).

2.1 Apparent Relation Descriptor

An apparent relation descriptor encodes apparent human relationship on images. For example, there exists a queuing group on an image, the group is more likely seen from a right view rather than from an anterior view. In this case, it is important to describe the appearance that persons queuing and facing to the right extend transversally. Such apparent relations are useful when the apparent relations are specific to the target activity. Our model uses the action context (AC) descriptor [1] to represent apparent relation.

AC descriptor is per-person descriptor, and each descriptor is calculated by concatenating the following two feature descriptors: one is the action descriptor that represents the action of the focal person, and the other is the context descriptor that captures the behavior of nearby people, as illustrated in Figure 2.

The action descriptor has a bag-of-words style. Instead of using raw person descriptors (e.g. HOG [14]), we describe the action descriptor generated by outputs of a multiclass SVM classifier associated with action labels. Using the score returned by the SVM classifier, the i-th person is represented as the following K-dimensional vector: $F_i = [S_{1i}, S_{2i}, ..., S_{Ki}]$, where K is the number of action classes, and S_{ki} is the score of classifying the i-th person to the k-th action class.

After the action descriptor is computed for each person, the context descriptor is calculated by integrating the action descriptor of nearby people in the "context region", as illustrated in Figure 2. The context region is further divided into M regions (called "sub-context regions") in space and time, then the context descriptor is represented as the following $M \times K$ dimensional vector:

$$\begin{aligned} C_i &= [D_{1i}, ..., D_{Mi}] \\ &= \left[\max_{j \in \mathcal{N}_1(i)} S_{1j}, ..., \max_{j \in \mathcal{N}_1(i)} S_{Kj}, ..., \max_{j \in \mathcal{N}_M(i)} S_{1j}, ..., \max_{j \in \mathcal{N}_M(i)} S_{Kj} \right], \end{aligned} \quad (1)$$

where D_{mi} is called the "sub-context descriptor" representing the context in the m-th sub-context region of the i-th person, and $\mathcal{N}_m(i)$ indicates the indices of people in the sub-context region.

Finally, the AC descriptor of i-th person A_i is computed by concatenating its action descriptor F_i and its context descriptor C_i: $A_i = [F_i, C_i]$.

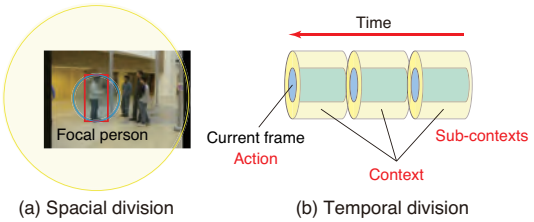

Fig. 2. Illustration of AC descriptor. AC descriptor is calculated by concatenating its action descriptor and its context descriptor. The context region is further divided into sub-context regions in (a) space and (b) time.

2.2 Relative Relation Descriptor

A relative relation descriptor encodes relative relationship between the focal person and others. For example, when the focal person is facing right and another person is facing left, the relative relation is defined as facing the "opposite" direction. This descriptor cannot represent apparent relations specific to the target activity, however, it contains invariance under viewpoint change (e.g, camera rotation), and consistency within the same collective activity.

Similarly to [2], we define actions by concatenating poses and activities (e.g. talking and facing right). This means that the action descriptor and the sub-context descriptor are $K(=U\times V)$ dimensional vectors, where U is the number of activity classes and V is the number of pose classes. Using U and V, we redefine the action descriptor F_i, and the sub-context descriptor D_{mi} in Section 2.1:

$$\begin{aligned} F_i &= [S_{1i}, S_{2i}, ..., S_{Ki}] \\ &= [S_{11i}, S_{12i}, ..., S_{uvi}, ..., S_{UVi}], \end{aligned} \tag{2}$$

$$\begin{aligned} D_{mi} &= \left[\max_{j\in\mathcal{N}_m(i)} S_{1j}, ..., \max_{j\in\mathcal{N}_m(i)} S_{Kj} \right] \\ &= \left[\max_{j\in\mathcal{N}_m(i)} S_{11j}, \max_{j\in\mathcal{N}_m(i)} S_{12j}, ..., \max_{j\in\mathcal{N}_m(i)} S_{uvj}, ..., \max_{j\in\mathcal{N}_m(i)} S_{UVj} \right]. \end{aligned} \tag{3}$$

Our proposed descriptor (the *relative action context (RAC) descriptor*) is calculated by shifting AC descriptor based on the pose of the focal person, as shown in Figure 3. First, the pose of the i-th person $\hat{v}_i$ is calculated from the person descriptor (e.g. HOG [14]) using a multiclass classifier. In terms of the pose $\hat{v}_i$, the i-th person's relative action descriptor $\hat{F}_i$ and the relative sub-context descriptor $\hat{D}_{mi}$ are defined as

$$\begin{aligned} \hat{F}_i = \big[& S_{1\hat{v}_i i}, ..., S_{1Vi}, S_{11i}, ..., S_{1(\hat{v}_i-1)i}, ..., \\ & S_{U\hat{v}_i i}, ..., S_{UVi}, S_{U1i}, ..., S_{U(\hat{v}_i-1)i} \big], \end{aligned} \tag{4}$$

$$\begin{aligned} \hat{D}_{mi} = \Bigg[& \max_{j\in\mathcal{N}_m(i)} S_{1\hat{v}_i j}, ..., \max_{j\in\mathcal{N}_m(i)} S_{1Vj}, \max_{j\in\mathcal{N}_m(i)} S_{11j}, \max_{j\in\mathcal{N}_m(i)} S_{1(\hat{v}_i-1)j}, ..., \\ & \max_{j\in\mathcal{N}_m(i)} S_{U\hat{v}_i j}, ..., \max_{j\in\mathcal{N}_m(i)} S_{UVj}, \max_{j\in\mathcal{N}_m(i)} S_{U1j}, \max_{j\in\mathcal{N}_m(i)} S_{U(\hat{v}_i-1)j} \Bigg]. \end{aligned} \tag{5}$$

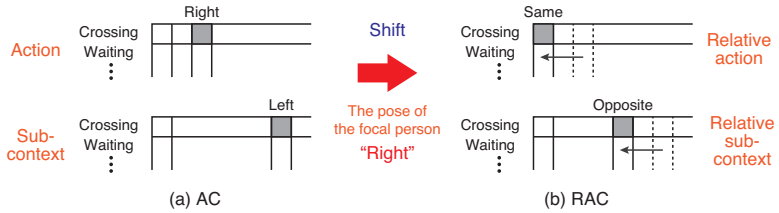

Fig. 3. Illustration of a method for constructing RAC descriptor from AC descriptor. RAC descriptor is calculating by shifting AC descriptor based on the pose of the focal person. When the focal person is facing right and another person is facing left in AC descriptor, another person is defined as facing the "opposite" direction in RAC descriptor.

The relative context descriptor of i-th person $\hat{C}_i$ is computed by concatenating its relative sub-context descriptor: $\hat{C}_i = [\hat{D}_{1i}, ..., \hat{D}_{Mi}]$. Finally, the RAC descriptor of i-th person R_i is computed by concatenating its relative action descriptor $\hat{F}_i$ and its relative context descriptor $\hat{C}_i$: $R_i = [\hat{F}_i, \hat{C}_i]$.

2.3 Combination of Descriptors

After extracting the apparent relation descriptor (AC descriptor) and the relative relation descriptor (RAC descriptor), we transform them into probabilities via softmax transformation, and combine them via the MAX rule [15]:

$$\hat{y}_i = \arg\max_{y_i} P_i(y_i) \quad \text{s.t.} \quad P_i(y_i) = \max_k P_i(y_i|d_k), \tag{6}$$

where $P_i(y_i)$ is the probability that the activity of the i-th person is y_i, $P_i(y_i|d_1)$ is the probability calculated from the apparent relation descriptor, $P_i(y_i|d_2)$ is the probability computed from the relative relation descriptor.

3 Methods to Accelerate Performance

3.1 Post Processes: Context Conversion

In order to make a histogram-style context (e.g. A_i, R_i) robust to various situations, we employ the following two post processes: threshold processing that allows the representation to capture the salient properties from noisy context; Gaussian processing that reduces quantization errors.

Threshold Processing: Given a histogram-style context, we execute the following threshold processing to each score s: $\hat{s} = s$ (if s $> \alpha$) or 0 (otherwise), where α is a threshold. Threshold processing makes the contexts sparse and allows the representation to be specialized. In implementation, we define $\alpha = 0$. A_i and R_i are the scores returned by the SVM classifiers, therefore, this threshold value implies that we exploit only the properties of present actions.

Gaussian Filtering: When converting continuous volume into quantized volume, quantization errors can be problematic. To relax such errors, our method executes the following Gaussian filtering to each score s: $\hat{s} = [s_l, s, s_r] \cdot [\frac{1}{4}, \frac{2}{4}, \frac{1}{4}]^T$, where s_l is the left neighbor score and s_r is the right neighbor score. For example, when s is a score of talking and facing right, s_l is a score of talking and facing back-right, and s_r is a score of talking and facing front-right.

3.2 Regularization using Fully Connected CRFs

In order to reduce local classification failures, we impose smoothness by applying fully connected CRFs [7, 12, 13]. In particular, our model not defines human relation as heuristic but assumes that all the persons in a frame are related and describes their relation as potential that vary depending on the scale of features, so as to apply various group shapes. The observed data of the detected persons are defined as $\boldsymbol{x} = \{x_1, ..., x_N\}$, where x_i is the observed data of the i-th person and N is the number of detected persons in a frame. Let the corresponding activity labels be defined as $\boldsymbol{y} = \{y_1, ..., y_N\}$. A conditional random field $(\boldsymbol{x}, \boldsymbol{y})$ is characterized by a Gibbs distribution: $P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp(-E(\boldsymbol{y}))$, where $Z(\boldsymbol{x})$ is the partition function, and $E(\boldsymbol{y})$ is the Gibbs energy:

$$E(\boldsymbol{y}) = \sum_i \psi_u(y_i) + \sum_i \sum_{j>i} \psi_p(y_i, y_j), \tag{7}$$

where $\psi_u(y_i)$ is the unary potential and $\psi_p(y_i, y_j)$ is the pairwise potential.

The unary potential $\psi_u(y_i)$ is defined as $\psi_u(y_i) = -\log(P_i(y_i))$, where $P_i(y_i)$ is the probability that the activity of the i-th person is y_i. The pairwise potential is defined in terms of the positions p_i and p_j, and weight w:

$$\psi_p(y_i, y_j) = w\mu(y_i, y_j) \exp\left(-\frac{|p_i - p_j|^2}{2\theta^2}\right), \tag{8}$$

where $\mu(y_i, y_j)$ is the label compatibility function given by Potts model [16]: $\mu(y_i, y_j) = [y_i \neq y_j]$. Note that we normalize positions by the minimum height of all the persons in a frame, and describe human relationship as relative value rather than absolute value, to obtain robustness to a difference in perspective.

Our model defines the pairwise potential as Gaussian kernel, therefore, in inference, it is possible to apply highly efficient approximated inference algorithm via mean field approximation and high-dimensional filtering [12]. This reduces the calculation cost to linear to the number of the detected persons N. In learning, the kernel parameters w, θ are estimated. Due to non-convexity of kernel width θ on log-loss criterion, it is hard to optimize it globally, therefore, we use grid search from the training set with cross-validation.

4 Experiments

Collective Activity Dataset: We evaluate our model on the collective activity dataset [4]. This dataset consists of 44 short videos of crossing, waiting, queueing, walking and talking. The videos were recorded under realistic conditions,

including camera shaking, background clutter and transient mutual occlusions of persons. All the persons in every 10th frame are labeled with the ground truth: pose, activity and bounding box information. We use the same leave-one-video-out scheme described in [1, 4, 5, 7], and report activity recognition results on a per-person basis. We use a linear SVM (e.g. LIBLINEAR [17]) as a classifier, and its parameters are set according to cross-validation in the training set.

Evaluation of the Group Context Descriptors: To evaluate our proposed group context descriptors, we demonstrate the two experiments: (1) comparison among the three group context descriptors (AC descriptor, RAC descriptor, and the combination of them), (2) comparison among the post processes (threshold processing and Gaussian filtering). Note that the focus of these experiments is evaluating each factor of group context descriptors. In this evaluation, we assume that persons are detected without errors, i.e., use ground-truth person locations.

The quantitative results are summarized in Table 1. Since the test set is imbalanced about activity classes and the difficulty of recognition is different among videos, we report overall, mean per-class and mean per-video accuracies. The quantitative results are competitive between AC descriptor and RAC descriptor in Table 1, however, the applicable scenes are different as presented in Figure 4. When apparently similar groups exist in the training data, AC descriptor is useful, however, when viewpoint change occurs, RAC descriptor is more useful. The combination of AC and RAC descriptors can handle both applicable scenes as illustrated in the top three rows of Figure 4, and exceeds AC and RAC descriptors in terms of number, as shown in Table 1. However, it is hard to handle the scene where the multiple groups are overlapping as shown in the bottom row of Figure 4. Note that we also evaluated other combination methods such as the MIN, Product, Sum rules [15], and an SVM classifier for the AC and RAC descriptors: $[A_i, R_i]$. All of them performed slightly worse than our approach.

To evaluate our proposed post processes, we compare classification accuracies with and without threshold processing or Gaussian filtering, in Table 1. Both AC and RAC descriptors have the highest accuracies using Gaussian filtering after threshold processing. This implies that relaxing quantization errors after extracting salient properties is the most effective. We also evaluated the other smoothness method such as mean filtering: $\hat{s} = [s_l, s, s_r] \cdot [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]^T$. It performed worse than our approach because it smoothens properties to excess.

Comparison with state-of-the-art Methods: We also compare our method with the recent methods [1, 4, 5, 7]. To compare fairly, we apply the pedestrian detector in [18] to detect the persons, similarly to [1, 4, 5, 7]. Experimental results are shown in Table 2 and Figure 5. Here, we report the regularizing classification result using CRFs. It exceeds the result without CRFs, since this model reduces local classification failures. Note that the approach of [7] needs the data in the entire video to obtain the spatial and temporal consistency, and the approach of [5] needs the 3D trajectory data of each person to apply 3D MRF, while our method does not need the surplus data.

Table 1. Comparison of group context descriptors: AC descriptor (the first row to the fifth row), RAC descriptor (the sixth row to the tenth row), and the combination of them (the eleventh row to the twelfth row). Comparison of post processes: threshold processing and Gaussian filtering. The characters T, G, TG, GT indicate threshold processing, Gaussian filtering, Gaussian filtering after threshold processing, and threshold processing after Gaussian filtering.

Method	Overall	Mean per-class	Mean per-video
AC	71.1	69.0	63.3
AC + T	73.0	71.2	64.7
AC + G	71.5	69.3	64.0
AC + TG	**74.0**	**72.2**	**66.3**
AC + GT	73.0	70.9	64.9
RAC	71.4	69.4	65.0
RAC + T	72.3	70.4	66.6
RAC + G	72.5	70.7	66.4
RAC + TG	**73.1**	**71.4**	**67.4**
RAC + GT	72.6	70.8	66.8
Combination (AC + RAC)	73.2	71.2	66.4
Combination (AC + TG) + (RAC + TG)	**75.1**	**73.1**	**68.6**

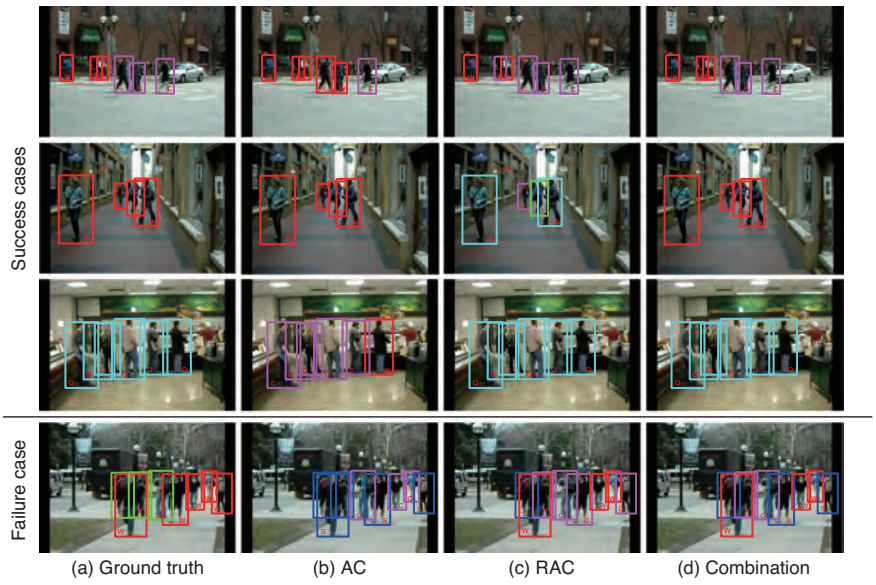

Fig. 4. Qualitative results of collective activities recognition using group context descriptors: (a) ground truth, (b) AC descriptor, (c) RAC descriptor, (d) the combination of them. The labels C (magenta), S (blue), Q (cyan), W (red), T (green) indicate crossing, waiting, queueing, walking and talking. Top three rows show examples of successful classification and a bottom row shows examples of false classification.

Table 2. Comparison of activity classification accuracies using different methods. Top eight rows show the results using only group context descriptors, and bottom four rows show the results regularized by graph structures. Our apparent relation descriptor (AC + TG) outperforms state-of-the art apparent descriptor (AC in [1]). Our relative relation descriptor (RAC + TG) also exceeds state-of-the art relative descriptors (STV in [4] and RSTV in [5]). Moreover, the combination of them surpasses them.

Method	Mean per-class
HOG	50.0
STV in [4]	64.3
RSTV in [5]	67.2
AC in [7]	67.4
AC in [1]	68.2
RAC + TG	68.5
AC + TG	71.3
Combination	**71.9**
STV + MC in [4]	65.9
RSTV + MRF in [5]	70.9
AC + FC-CRF in [7]	72.2
Combination + CRF	**73.2**

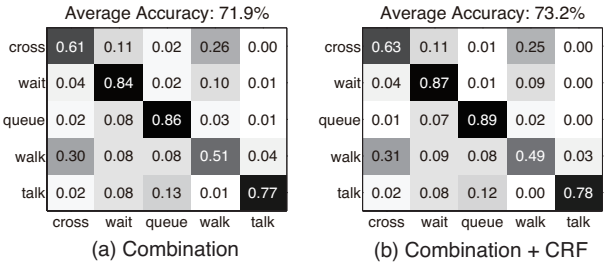

Fig. 5. Confusion matrices for activity classification with and without CRFs: (a) the result with the combination of AC + TG and RAC + TG, and (b) the result with the combination + CRF. In the confusion matrices, rows represent ground truths and columns represent predictions. Each row is normalized to sum to 1. Note that walking vs crossing is still ambiguous in our model, because these activities often depend on not human relationship but environmental settings: a sidewalk or a pedestrian crossing.

5 Conclusion

This paper has described the novel relative relation descriptor called RAC descriptor, as against an apparent relation descriptor such as AC descriptor. Owing to its "relative" relation description, the proposed RAC descriptor is viewpoint invariant and consistent within the same category of collective activity. AC and

RAC descriptors are complementary in human relation representation; hence we employ a simplified combinational classifier, so as to obtain stable performance in various scenes. We also introduce two methods to improve performance. One is post processes that make the contexts robust to various situations, and the other is regularizing classification by fully connected CRFs that assume all the relationships among the people in a frame. Finally, our experimental results on the collective activity dataset demonstrate that our method recognizes collective activity stably in various scenes and outperforms state-of-the art methods.

References

1. Lan, T., Wang, Y., Mori, G.: Retrieving actions in group contexts. In: International Workshop on Sign Gesture Activity. (2010)
2. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: Adv. in NIPS 23. (2010)
3. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: CVPR. (2012)
4. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: International Workshop on Visual Surveillance. (2009)
5. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: CVPR. (2011)
6. Amer, M.R., Todorovic, S.: A chains model for localizing participants of group activities in videos. In: ICCV. (2011)
7. Kaneko, T., Shimosaka, M., Odashima, S., Fukui, R., Sato, T.: Consistent collective activity recognition with fully connected CRFs. In: ICPR. (2012 (to appear))
8. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: ICCV. (2009)
9. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR. (2009)
10. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision. (2004)
11. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
12. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Adv. in NIPS 24. (2011)
13. Zhang, Y., Chen, T.: Efficient inference for fully-connected crfs with stationarity. In: CVPR. (2012)
14. Dalal, N., Triggs, B.: Histogram of oriented gradients for human detection. In: CVPR. (2005)
15. Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. PAMI **20** (1998) 226–239
16. Boykov, Y.Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: ICCV. (2001)
17. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9** (2008) 1871–1874
18. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR. (2008)