Multimedia Systems (2017) 23:667–677
DOI 10.1007/s00530-016-0514-9

SPECIAL ISSUE PAPER

Automatic group activity annotation for mobile videos

Chaoyang Zhao[1] · Jinqiao Wang[1] · Jianqiang Li[2] · Hanqing Lu[1]

Published online: 20 April 2016

Abstract Due to the rapid growth of modern mobile devices, users can capture a variety of videos at anytime and anywhere. The explosive growth of mobile videos brings about the difficulty and challenge on categorization and management. In this paper, we propose a novel approach to annotate group activities for mobile videos, which helps tag each person with an activity label, thus helping users efficiently manage the uploaded videos. To extract rich context information, we jointly model three co-existing cues including the activity duration time, individual action feature and the context information shared between person interactions. Then these appearances and context cues are modeled with a structure learning framework, which can be solved by inference with a greedy forward search. Moreover, we can infer group activity labels of all the persons together with their activity durations, especially for the situation with multiple group activities co-existing. Experimental results on mobile video dataset show that the proposed approach achieves outstanding results for group activity classification and annotation.

✉ Jinqiao Wang
jqwang@nlpr.ia.ac.cn

Chaoyang Zhao
chaoyang.zhao@nlpr.ia.ac.cn

Jianqiang Li
lijianqiang@bjut.edu.cn

Hanqing Lu
luhq@nlpr.ia.ac.cn

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2] School of Software Engineering, Beijing University of Technology, Beijing, China

Keywords Activity annotation · Group activity · Context learning

1 Introduction

The current growth and tremendous proliferation of mobile devices have paved the way for capturing a variety of multimedia contents. Mobile devices have become a sharing platform for diversified photos and videos about the activities in the daily life. Due to the diversity of interests, users of mobile devices always upload numbers of videos to the cloud server for their personalized social activities. There is an urgent need to automatically analyze and tag these personalized mobile videos for effective storage and management.

An efficient mobile video analyzing system solves the problem of individual activity annotation by predicting and classifying the activity classes of each person, which is to make it easy to search for interesting activities and actions. Traditional works mainly focus on face or person annotation for photos and videos. However, the tagging of complex behaviors or activities is also critical for content search and browsing for mobile videos. The increase of mobile videos shared on the internet and the growth of mobile devices motivate us to conduct a detailed research for automatic activity annotation for mobile videos.

Analyzing human activities from videos has been a challenging task in the past few years. Most of the traditional vision-based activity recognition works have been focusing on single-person activities. However, videos recorded by mobile devices often involve in realistic scenes of human activities, which contain multiple, inter-related actions at the same time, and the analysis of a single individual cannot yield reliable results. The context information inside

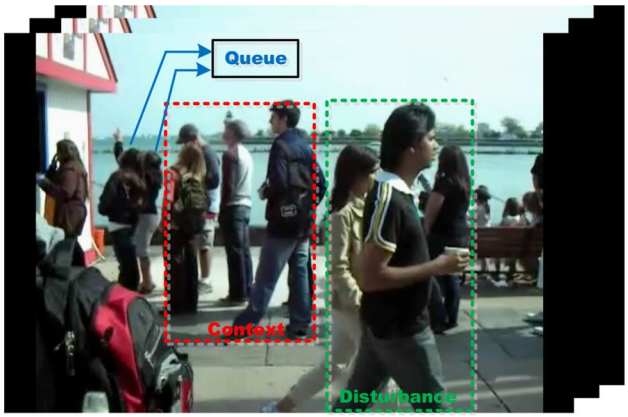

Fig. 1 An illustration of tagging the persons' activity for a mobile video. By successfully classifying that the two girls are queueing, we need not only the context information that some people are standing behind them in a *line*, but also excluding the disturbance that two persons are walking by. Source image comes from the Collective Activity Dataset [7]

these mobile videos provides lots of auxiliary information for the analysis of human activities. As illustrated in Fig. 1 a girl recorded a video that two of her friends were queueing to buy something. To annotate the persons inside the video with correct labels, we need not only their own appearance features, but also the context information that a group of persons were standing in a line facing the same direction. In addition, the disturbance that two persons were passing by needs to be carefully excluded. The analysis of these kinds of mobile videos in real-world scenario faces great challenges.

To this end, many researchers turn to analyze a group of persons' behaviors with interactions among each other within the videos. They refer this problem as "group activity" or "collective activity" recognition problem [6, 28]. Researchers have explored the group–person interactions and person–person interaction context for better classifying the activities. Based on our observation, the analysis of group activities from mobile videos often involves not only the individual action, the context information shared between persons, but also the concurrence of the activity durations. However, few of former works have made use of these kinds of information.

In this paper, we propose a framework for automatic activity annotation for mobile video. A novel concurrent group activity annotation approach is proposed in this paper. The activity duration time, the individual action feature and the context information that encodes person interactions are modeled jointly. Moreover, the interactions between persons belonging to different group activities are also considered in. By introducing carefully designed context descriptors, our approach provides strong cues that the context information like trajectory interactions and person relationships improves the group activity classification performance significantly.

The main contributions are summarized as below:

- We propose a context structure learning framework to model multiple context cues of person interactions for group activity annotation.
- Two context models including individual trajectories context model along with the activity duration model are proposed to improve the performance of group activity annotation.
- Through encoding interactions among persons with different activities, our approach can handle the situation with multiple group activities co-existing in the videos.
- The rest of this paper is organized as follows. Section 2 introduces the related work with activity recognition. The automatic group activity annotation framework is introduced in Sect. 3. The unified discriminative model of group activity recognition is proposed in Sect. 4. Section 5 introduces the model learning techniques. Experimental results are reported in Sect. 6. Finally, Sect. 7 concludes our work.

2 Related work

Since activity classification is the core problem for mobile video tagging, here we review the state-of-the-arts for activity recognition and collective activity recognition. Human visual recognition using context information has received much attention recently in computer vision community. Researchers have done many works on exploiting context information between scenes and objects [23], objects and objects [10, 16, 24], or human and objects [29]. These works focused on recognizing single person actions while neglecting the interactions that might exist in the scene.

Compared to single person actions, recognizing collective activities is a more complicated problem. It often involves the identification of multiple human actions and recognition of human interactions with each other. Ryoo and Aggarwal [25] modeled the pairwise interactions between people to recognize complex human activities. Gupta et al. [12] recognized group activities in sport events by human roles in the scene through learning a storyline model with AND–OR graphs. Choi et al. [7, 8] designed a "crowd context" descriptor to describe the activities performed by individuals in a crowd. Lan et al. [18, 20] used "action context" to describe the influence caused by other actions near a focal person and then explored the group–person interaction and person–person interaction context with a high-level latent discriminative model. Choi et al. [6, 28] also formulated the collectivity recognition and multiple target tracking into a unified framework to acquire

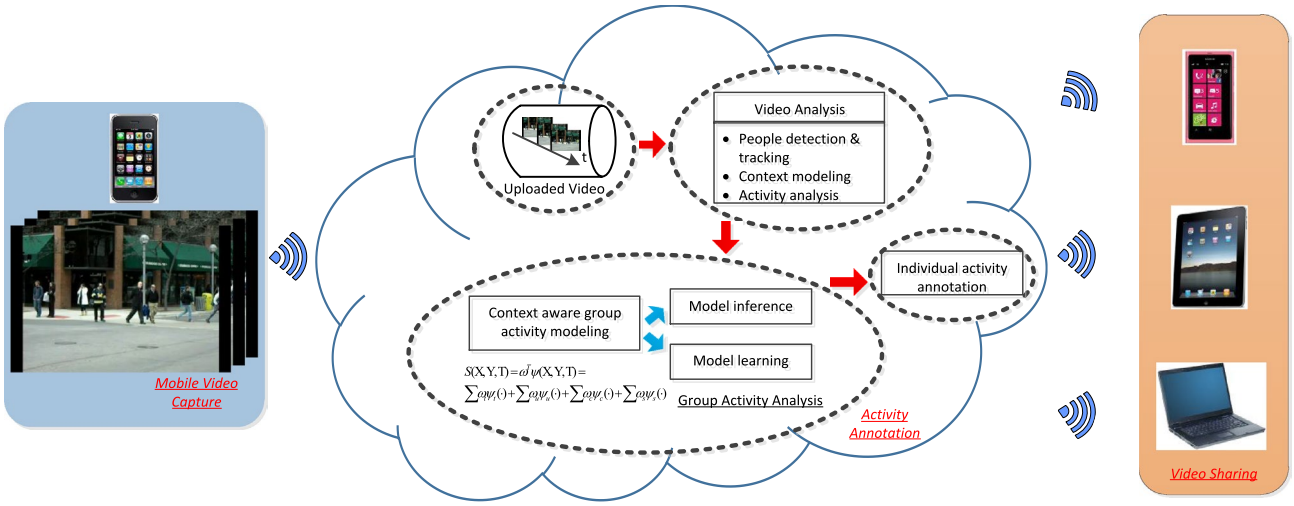

Fig. 2 Illustration of the framework of activity annotation and sharing for mobile videos

the performance gain. These works have achieved some improvements for recognizing group activities. Antic and Ommer [2] used multiple-instance learning to model the group activities. However, most of the previous approaches have made a strong hypothesis for activity modeling, i.e., there exists a dominating collective activity in the scene, interactions for persons with different activity labels are not taken into consideration. Therefore, this assumption often leads to miss-classification for scene with multiple activities co-occurring.

Choi et al. [8] recognized scene with multiple activities with an MRF model. They set one predominant activity in a scene and considered individuals with other activity labels as anomalous. Amer et al. [1] detected and localized a wide range of activities in an open scene, they used a three-layered AND–OR graph to jointly model group activities, individual actions and the objects. This approach noticed the co-occurrence of multiple activities, but did not explore the interactions existing between persons belonging to different activities. Li et al. [21] modeled group interactions in social clutter. Zhu et al. [31] defined context information between activities to improve the activity recognition rate. They used a structural model to integrate motion features and context features in and between activities. The definition of activity was based on one person interacting with the surroundings and could hardly solve the problem of collective activity recognition with many persons interacting with each other. To the best of our knowledge, few approaches has been done to explore the interactions existing between different activities. Our approach is dedicated to study the influence caused by different activities co-occurring in a scene. We refer this influence as inter-class context information and model it in a unified framework to improve the performance of collective activity recognition.

Context information is often modeled by a graph model [14, 15, 29, 30]. Additionally, structural framework also is employed for its strong ability to model low-level features and middle-level features jointly [10, 19, 20, 31]. The inference method on a graph model or a structural model often needs to search through the graphical structure to find the one that maximizes the potential function. This kind of solution is often very time consuming. Desai et al. [10] discarded the false detection results using a greedy search strategy to solve the inference problem. The same strategy was adopted to infer human activities in videos in [31]. These approaches could reduce the computational complexity of the inference problems while maintaining a considerable result.

3 Overview

As illustrated in Fig. 2, we propose a framework for efficiently activity annotation and sharing for mobile videos. The recorded mobile videos are uploaded to the multimedia server. Then with people detection and tracking, we extract appearance features and contextual information for activity annotation. Finally, the processed videos are uploaded to video web site for content browsing and sharing. With the annotated results, users are able to manage and share these videos based on their personal preference more easily.

Since there are already many great works for people detection, here we use the DPM detector [11] for human localization. Group activity classification is the main step for video tagging, for which we need to tag each person inside the video a belonged activity label based on their individual action features and context information. Besides, for concurrent group activities, we additionally

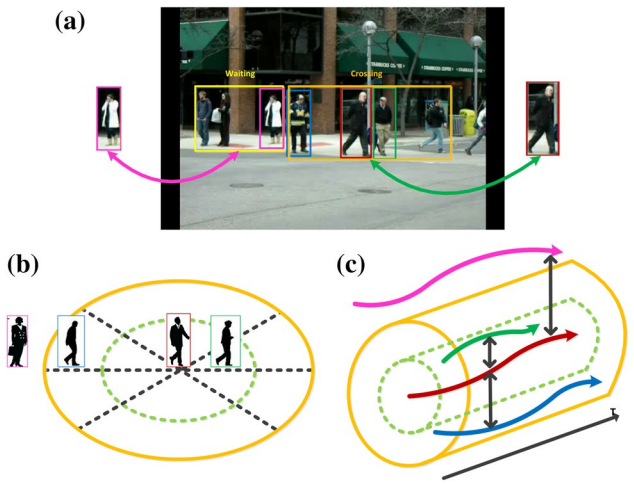

Fig. 3 An illustration of concurrent group activities. **a** Person interactions with each other with a mobile video; **b** the focal person's activity is influenced by the region context; **c** interactions between trajectories serve as context information (color figure online)

take into consideration of the influence among persons with different group activity labels. Figure 3 shows an example of a mobile video with concurrent group activities. Take for example those persons in Fig. 3a, we may easily know that the woman is standing and the man is walking by analyzing their low-level visual features. But once taking into 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.

4 Concurrent group activity modeling

In this section, we detail the context modeling procedure of our approach for efficient group activity classification. Our approach enables analyzing human group activities by looking at context information extracted from all the persons and their relationships in a video sequence. Given a video sequence, the human detection and tracking procedure ensure that the persons' bounding boxes and their local trajectories can be used directly. To evaluate the effectiveness of different context information extracted from collective activities, we establish multiple context models to encode the individual appearance feature, the activity duration time and the trajectory interactions among persons within the same scene. Whenever there are multiple group activities co-existing in a scene, our context descriptors show the ability of jointly modeling all the interactions among all the persons with variety of group activity labels.

4.1 Problem formulation

Detecting people in the video frames is task specific, here we assume that the images have been preprocessed and the locations of the persons have been found. We focus on the task of tagging each individual with an activity label. Assuming there are C classes of collective activities in the scene, where the label $y_i \in \{1, 2, \ldots, C\}$ denotes the activity class of a person. Let $Y = \{y_i : i = 1, 2, \ldots, N\}$ be the label set for all N persons in a scene and $T = \{t_i : i = 1, 2, \ldots, N\}$ stand for the auxiliary time duration set, where t_i is the group activity duration for the ith person. The task is then converted into finding the optimal hypothesis label set (Y, T) for all the N persons in the scene. We extract features $X = \{x_i : i = 1, 2, \ldots, N\}$ from the scene for all the persons. The low-level feature x_i represents different action h_i within the group activity label y_i.

We use $S(X, Y, T)$ to represent the compatibility of the low-level feature X, the group activity label set Y and the duration time set T. Since $S(X, Y, T)$ is modeled by multiple context information, here we introduce each context model in detail.

4.2 Activity–duration potential

We start from modeling the activity–duration of the person for a video clip. To measure the compatibility between the group activity label y_i and its duration t_i for the ith person, we define the activity–duration potential $w_t^T \Psi_t(y_i, t_i)$ as

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

where w_t^T stands for the parameter needs to be learnt and Ψ_t means our defined feature descriptor that encodes the activity–duration relationship. $I(t_i)$ is the indicator for the ith person that with activity label y_i. For the scene with N possible group activities, $I(t_i)$ is a $N \times 1$ vector with ith element marked as 1 when the activity label of t_i is y_i and all the other elements are set to be 0.

4.3 Unary action–activity potential

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 [9] 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

$$x_i = (s_{\max,i}, \mathrm{pos}_i, v_i) \tag{2}$$

where $K = 8$ is the number of pose categories within a activity, $s_{\max,i}$ is the maximum pose classification score

with the activity label y_i, v is the average velocity of the person. pos_i is the pose indicator for the ith person in the subregion, which generates a $K \times 1$ vector with one for the (k)th element and zeros otherwise. The (k)th element represents for the pose label that $s_{\text{max,i}}$ belongs to. Then the action–activity potential is parameterized as:

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

where w_u^T represents the parameter needs to be learnt and Ψ_u is our defined feature descriptor that encodes the action–activity relationship together with the activity duration t_i. Here the index "u" means "unary".

The activity of one particular person is influenced heavily by the context information. As shown in Fig. 4, for an open scenario, persons often have varied activity labels, our intuition is that persons closer to each other may have strong interactions with each other, even if they have varied activity labels. We model the context information for all the persons within the scene, including those with different activity labels. In the following, we introduce two different context potentials for context information modeling.

4.4 Region context potential

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 subregions are shown in Fig. 3b, the feature is computed from persons inside the context subregions. For each context subregion, we divide the subregion into D small bins. Each bin may cover many persons at the same time. For all the persons within the same bin, the dth bin context feature is defined as

$$f_{\mathcal{N}_{\text{sub}}}(d) = \left[\max_{j \in \mathcal{N}_{\text{sub}}(d)} (S_{\text{pose},j}), \sum_{j \in \mathcal{N}_{\text{sub}}(d)} \text{pos}_j, \max_{j \in \mathcal{N}_{\text{sub}}(d)} (v_j) \right] \tag{4}$$

where $\text{sub} = 1, 2$ represents the subregions, $d = 1, \ldots, D$ means the dth bin in that subregion. Hence, the region context feature encodes the pose score $S_{\text{pose},j} = [s_{1j}, \ldots, s_{Kj}]$, pose histogram $\sum_{j \in \mathcal{N}_{\text{sub}}(d)} \text{pos}_j$ and the velocity score $\max(v_j)$ for the jth person falling in the dth bin. $f_{\mathcal{N}_{\text{sub}}}(d)$ finds the person who shows the most influence in the dth bin. For all D bins within a subregion, the region context feature is defined as D context descriptor concatenating together:

$$f_{\mathcal{N}_{\text{sub}}} = (f_{\mathcal{N}_{\text{sub}}}(1), \ldots, f_{\mathcal{N}_{\text{sub}}}(D)) \tag{5}$$

Supposing that the context region contains two subregions, the region context feature is defined as

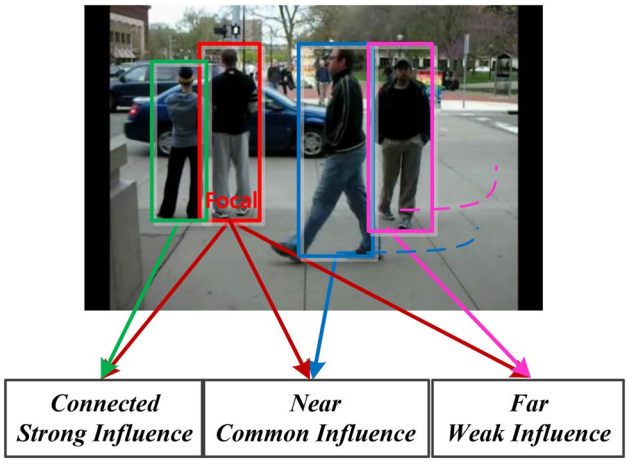

Fig. 4 A demonstration of how people influence each other within a certain distance. The focal person is influenced by the trajectories within a distance. Here we divide the trajectory distance into three bins: *connected, near, far*

$$\begin{aligned} fc_i &= [f_{\mathcal{N}_1}, f_{\mathcal{N}_2}] \\ &= [f_{\mathcal{N}_1}(1), \ldots, f_{\mathcal{N}_1}(D), f_{\mathcal{N}_2}(1), \ldots, f_{\mathcal{N}_2}(D)] \end{aligned} \tag{6}$$

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

$$w_c^T \Psi_c(y_i, t_i) = t_i w_c^T \cdot fc_i \tag{7}$$

here w_c^T is the parameter needs to be learn and Ψ_c is our defined region context descriptor that combines the activity duration t_i and the region context feature fc_i for the ith focal person.

4.5 Trajectory context potential

This potential models the compatibility between the ith and jth person for their spatial and temporal interactions. These interactions are presented by pairwise interaction features extracted from related trajectories. As illustrated in Fig. 3c, for a video sequence, when 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). For two persons i and j (the red and the pink trajectories in Fig. 3c), we use dynamic time warping (DTW [3]) to measure the distance between the two trajectories due to their different start points or time durations. Together with the pose information, the pairwise interaction feature is defined as:

$$fi_{i,j} = [\text{bin}(\text{dist}_{ij}), \text{pose}(i, j)] \tag{8}$$

where dist_{ij} is the DTW distance between two trajectories. It is further divided into three bins defined as *connected, near* and *far* as illustrated in Fig. 4. *Connected, near and*

far are defined based on the statistical spatial relationship of the different trajectories. For the two trajectories, their relationship is set to be *connected* when the distances of their most frames are within $0.5h$; the relationship is set to be *near* when distances are between $0.5h - h$, others are set to be *far*. Here h is also the focal person height.

pose(i, j) is defined as $\max(\Psi_u(x_i), \Psi_u(x_j))$. Then the trajectory context potential is parameterized as:

$$w_s^T \Psi_s(x_i, x_j, y_i, y_j, t_i, t_j) = (t_i \cap t_j) w_s^T \cdot fi_{i,j} \tag{9}$$

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

It is common for a scene to contain two or more group activities simultaneously. Hence, analyzing the interactions existing among persons with different group activity labels has the potential of bringing further performance gain. Here for the designed context potentials, we allow the context descriptors to encode the interactions of persons with different group activity labels ($y_i \neq y_j$).

For the context potentials, all the parameters need to be tune are listed as follows.

Similar to [6, 7, 10], in this paper, instead of using pixel values after camera calibration or using the pixel values directly, our context descriptors encode these pixel measures to statistical features to model the interactions and relative spatial relationships among persons within the same scene. All these context descriptors are encoded into the context potentials for structural model learning, which further guarantees the robustness of these descriptors. In this way, our feature encoding method is not strictly scene related and can guarantee certain performance with varied training/testing data or similar scenes (Table 1).

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

$$\begin{aligned} S(X, Y, T) &= \omega^T \Psi(X, Y, T) \\ &= \sum_i w_t^T \Psi_t(\cdot) + \sum_i w_u^T \Psi_u(\cdot) \\ &\quad + \sum_i w_c^T \Psi_c(\cdot) + \sum_{i,j} w_s^T \Psi_s(\cdot) \end{aligned} \tag{10}$$

Table 1 Detailed parameters for context descriptors

Description	Parameter	Value/heuristics
Focal person's height	h	Pixels in the image
Pose category number	K	8
Region context subarea	$\mathcal{N}_1(\cdot)$	$0.5h$
radius (Fig. 3b)	$\mathcal{N}_2(\cdot)$	$0.5h - 2h$
Region context section number		
for each subarea (Fig. 3b)	D	8
trajectory relationship		Near: distance $< 0.5h$
Connected, near, far	bin(dist$_{ij}$)	Connected: $0.5h <$ distance $< h$
(Eq. 8) (Fig. 4)		Far: otherwise

Here all the individual and context information are modeled jointly within one framework.

5 Structural model learning

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

5.1 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 [31]. Although this greedy search algorithm cannot guarantee a globally optimum solution, in practice it works well to find good solutions.

Firstly, we initialize the label vector Y to be 0 for all persons. Then we greedily select the ith single person that, when labeled as a particular activity class a, increases the score S by the largest amount. After that we have $y_i = a$, and its belonged duration t_i is acquired by the tracing procedure of the ith person. The ith person is added to the labeled set I. We

Algorithm 1 The Greedy Forward Search Inference

Input:
A test image with N total persons.
Output:
The optimal re-assigned label vector Y^* and T^*.
1: **Initialization:**
$I = \emptyset$, $S = 0$
$\Delta(i,a) = \omega_t^T \Psi_t(\cdot) + \omega_u^T \Psi_u(\cdot) + \omega_c^T \Psi_c(\cdot) + \omega_s^T \Psi_s(\cdot)$
2: **Repeat:**
$(i,a)^{opt} = arg\max_{(i,a)\notin I} \Delta(i,a)$;
$I = (i,a)^{opt} \cap I$;
$S = S + \Delta(i,a)^{opt}$;
$Y^* = Y(I)$, $T^* = T(Y^*)$;
$\Delta(i,a) = \Delta(i,a) + \omega_s^T \Psi_s(x_i, x_{i^*}, a, a^*)$
3: **Until** $\Delta(i,a)^{opt} < 0$ or all N persons are labeled.

repeat this procedure until all the N persons are re-assigned with new labels. The whole computation can be very efficient by tracking the potential gain of adding label assign incrementally. Algorithm 1 describes the inference process.

5.2 Model learning

In the procedure of model learning, given a set of N training samples $\{(X_i, Y_i, T_i)\}$ $(i = 1, 2, \ldots, N)$, the goal is to estimate the optimal parameters of duration model w_t, unary action model w_u, the region context model w_c and the trajectory context model w_s that tend to produce the correct group activity label set Y and T for a new test video X. Equation 10 can be rewritten as follows under the structural SVM framework:

$$S(X,Y) = w^T \Psi(X,Y) \tag{12}$$

$$w = \begin{bmatrix} w_t \\ w_u \\ w_c \\ w_s \end{bmatrix}, \quad \Psi(X,Y) = \begin{bmatrix} \sum_i \Psi_t(x_i, y_i) \\ \sum_i \Psi_u(x_i, y_i) \\ \sum_i \Psi_c(y_i) \\ \sum_{i,j} \Psi_s(x_i, x_j, y_i, y_j) \end{bmatrix}$$

where w is the model parameter we need to learn and our inference procedure solves Eq. 11.

Assuming that we have the training video X_i and their corresponding label set Y_i, we want to train a model w that, with a new video X_j, tends to produce the true label vectors $Y_j^* \simeq Y_j$ and $T_j^* \simeq T_j$. The objective function can be converted to a regularized learning problem as follows:

$$\begin{aligned} &\arg \min_{w,\xi_i \geq 0} w^T w + C \sum_i \xi_i \\ &\text{s.t.} \, \forall i \quad w^T \Delta\Psi(X_i, Y_i, H_i, T_i, HT_i) \\ &\qquad \geq l(Y_i, H_i, T_i, HT_i) - \xi_i \end{aligned} \tag{13}$$

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.

The problem in Eq. 13 can be converted to an unconstrained convex optimization problem [10, 27]. It can iteratively search for the increasingly tight quadratic upper and lower cutting planes of the objective function until the gap between the two bounds reaches certain thresholds. We adopt the cutting plane optimization algorithm [10] to solve our problem and set all weights related to background activities to be zeros.

6 Experiments

6.1 Group activity classification for mobile videos

There are some standard benchmarks such as the KTH [26] and Weizmann [4] for human action understanding. However, the videos in those datasets were recorded in controlled settings and seldom involved in complex human group activities. In this paper, to evaluate the performance of the proposed approach, especially for the group activity classification algorithm, we carry out our experiments on Collective Activity Dataset [7]. This dataset contains 44 video clips acquired using low-resolution handled cameras, which serve as perfect simulation for videos shot by mobile devices. All the people in every 10th frame of the videos are assigned with one of the following collective activity labels: *waiting, queuing, walking,* and *talking*, together with one of the eight pose categories: *right, front-right, front, front-left, left, back-left, back* and *back-right*.

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 ten times and report the average results.

For the Collective Activity Dataset, more than 31 % of the labeled images contain two or more collective activities in the scene. Hence, the interactions existing among persons with different group activity labels provide important cues for activity annotation. However, most existing approaches ignored this kind of information. In this paper, we consider the situation with multiple group activities co-existing. In the experiments, we will analyze the

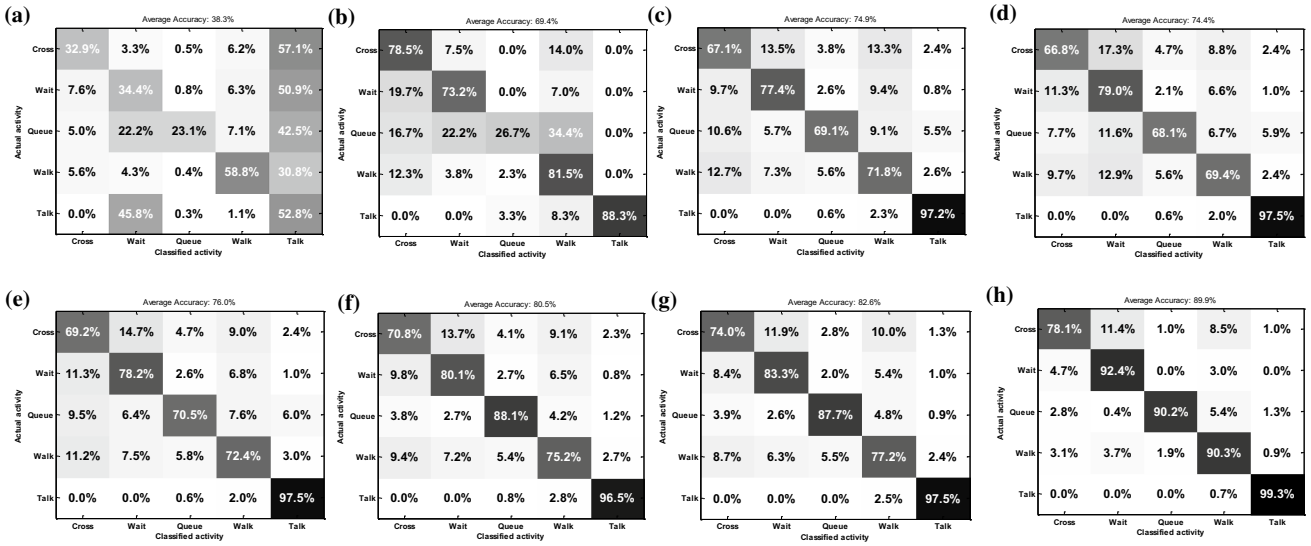

Fig. 5 Confusion matrices for activity classification accuracy with different feature fusion strategies: **a** duration and unary feature; **b** duration, unary feature and action context in [18]; **c** duration, unary feature and region context; **d** duration, unary feature and trajectory context with *pose*(i, j) in Eq. 9; **e** duration, unary feature and spatial context in [10]; **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

Table 2 Comparison results with state-of-the-arts

Approaches	Average accuracy (%)
ActionContext model [18]	68.2
RandomForest model [8]	70.9
Latent model [20]	79.1
Our approach without multiple activities	82.6
Person–person interaction model [5]	83.3
Our approach (with multiple activities)	**89.9**

Best performance result is in bold

interactions between the co-occurring activities by testing different context information.

6.1.1 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. 5. We can see that the approach without context information yields the worst result as shown in Fig. 5a. By adding the "*action context*" [18], the performance shown in (b) improves more than 30 % on average precision, which indicates the positive effect of the context information. Figure 5c presents the further improvement with our region context. From Fig. 5d, f, we can see the importance of pose(i, j) in Eq. 9. Our trajectory context potential in (f) also outperforms the spatial context [10] in Fig. 5e. Compared with "our approach without multiple activities" in Fig. 5g, the classification accuracy in Fig. 5h benefits from modeling the inter-group interactions especially when multiple activities co-exist in a scene. Our final model "our approach with multiple activities" in Fig. 5h archives the best result over all the approaches. For "our approach without multiple activities", we only consider the influence existing among persons with the same activity label and ignore the influence caused by other persons for all the potentials designed in this paper. On the other hand, for "our approach with multiple activities", we bring in the influence that may exist among all the persons in the scene. The context descriptors encode the interactions among all the persons with different activity labels.

6.1.2 Comparison with state-of-the-arts

The comparison results with state-of-the-arts are presented in Table 2. The "ActionContext model" [18] used the action context feature. The "RandomForest model" [8] used a random forest classifier to model the spatial–temporal information. The "Latent Model" [20] used a hierarchical latent model to formulate the group activity. The "Our approach

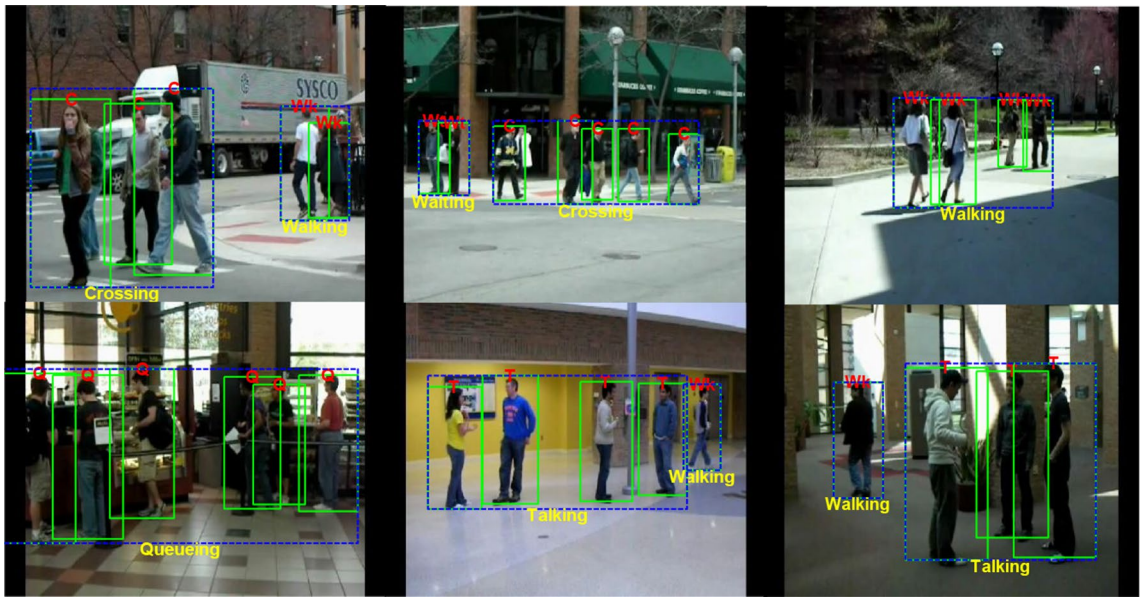

Fig. 6 Illustration of group activity tagging results. Persons are labeled with different group activity labels and their belonging groups

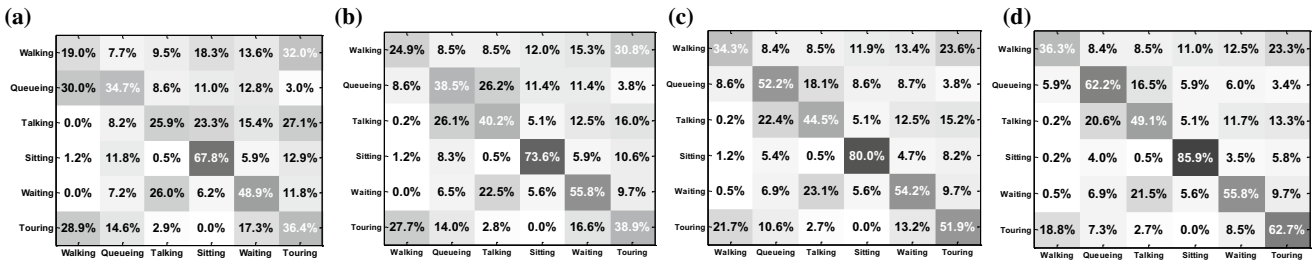

Fig. 7 Confusion matrices for activity classification accuracy in UCLA Courtyard dataset. **a** Using only unary action–activity potential; **b** combing unary and time duration potential; **c** combing unary, time duration and region context potential; **d** combining all the potentials

without multiple activities" stands for the model described in Fig. 5g. Without considering the interactions across different activities, its result already outperforms most of the approaches, which suggests the effectiveness of our designed context descriptors. By considering the situation of multiple group activities co-existing, our approach outperforms current state-of-the-art [5] by 6 % (Table 2).

Table 3 Average classification accuracy for different context information fusion strategies in UCLA Courtyard dataset

Approaches	Overall (%)	Mean per-class (%)
Unary potential	39.7	38.8
+Activity duration	46.8	45.3
+Region context	52.5	52.8
+Trajectory context	56.5	58.7

6.2 Examples of activity annotation

The group activity classification provides us relatively accurate classification results. Based on these classification results, we can then tag each person a belonged activity label, thus helping to classify the videos. Figure 6 shows some intuitional results, in which persons are labeled by their group activities. These results can help the users to automatically categorize to different classes for effective content management. The visualized activity recognition results can also be easily delivered to a variety of client terminals by the server for information sharing.

6.3 Group activity classification for surveillance videos

Though we design our approach for classifying group activities in mobile videos, we also test its generalization ability for surveillance videos. We carry out our experiment in

UCLA Courtyard Dataset [1], which consists of high-resolution videos that recorded multiple co-occurring activities which took place in a courtyard of the UCLA campus. The person together with their belonged activities are labeled in each frame. Each group activity consists of individual actions, pose and the orientations of the group participants. The annotated group activities include: *walking-together (walking), standing-in-line (standing), discussing-in-group (talking), sitting-together (sitting), waiting-in-group (waiting) and guided-tour (touring)*.

Since few of the previous works evaluated the group activity classification models on this dataset, here we test our approach on this dataset to evaluate the influence caused by different context information. We extract the labeled images every 30th frame and extract the trajectories for each person. We then split the data by 50–50 % into training and testing sets. We also report both the overall and mean per-class accuracies computed from all the testing images.

Figure 7 shows the confusion matrixes of activity classification with different context potentials implementation. We can see that the results show similar properties compared to the results for mobile videos in Sect. 6.1. Without the context information, we can only derive poor results as shown in Fig. 7a. "walking" and "touring" are hard to be separated from each other without the help of context information. From Fig. 7b–d, we can observe that the classification performance is improved steadily with the carefully designed context potentials involved. Table 3 shows the corresponding average classification results for the UCLA dataset, which also shows that the modeling of context information can bring remarkable performance gain for these kinds of group activity classification problems.

The approach in [1] modeled the group activity, primitive action, and object labels together with the AND–OR graph model. It used much more label information than our approach to obtain an average accuracy higher than 65 %. However, our experiments carried on the UCLA Courtyard Dataset predict the group activity labels without the help of any other auxiliary information such as action label or object label. When there is no auxiliary information, our approach shows better classification result than the AND–OR model on Collective Activity Dataset (84.8 % for first five class average classification accuracy in Table 2 in [1]).

7 Conclusion

In this paper, we present a framework for automatic activity classification and annotation for mobile videos. A novel concurrent group activity classification approach is proposed for efficient mobile video content analysis. 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 and comparison results on public mobile video dataset demonstrate that jointly modeling the individual appearance feature and the activity context features can significantly improve the recognition accuracy of collective activities. The efficient activity annotation approach for the mobile videos helps the users to easily browse and search their favorite content on the internet.

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

References

1. Amer, M.R., Xie, D., Zhao, M., Todorovic, S., Zhu, S.C.: Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In: ECCV (2012)
2. Antic, B., Ommer, B.: Learning latent constituents for recognition of group activities in video. In: European Conference on Computer Vision (ECCV) (2014)
3. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. KDD workshop, vol. 10, pp. 359–370 (1994)
4. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Tenth IEEE International Conference on Computer Vision (ICCV), vol. 2, pp. 1395–1402 (2005)
5. Chang, X., Zheng, W.-S., Zhang, J.: Learning person-person interaction in collective activity recognition. IEEE Trans. Image Process. **24**(6), 1906–1918 (2015)
6. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: European Conference on Computer Vision (ECCV) (2012)
7. Choi, W., Shahid, K., Savarese, S.: What are they doing? Collective activity classification using spatio–temporal relationship among people. In: IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1282–1289 (2009)
8. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3273–3280 (2011)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 886–893 (2005)
10. Desai, C., Ramanan, D., Fowlkes, C.C.: Discriminative models for multi-class object layout. Int. J. Comput. Vis. **95**(1), 1–12 (2011)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
12. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding videos, constructing plots learning a visually grounded storyline

model from annotated videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2012–2019 (2009)
13. Han, D., Bo, L., Sminchisescu, C.: Selection and context for action recognition. In: IEEE International Conference on Computer Vision (ICCV), vol. 9, pp. 1933–1940 (2009)
14. Han, Y., Wu, F., Lu, X., Tian, Q., Zhuang, Y., Luo, J.: Correlated attribute transfer with multi-task graph-guided fusion. In: Proceedings of the 20th ACM international conference on Multimedia, ACM, pp. 529–538 (2012)
15. Han, Y., Wei, X., Cao, X., Yang, Y., Zhou, X.: Augmenting image descriptions using structured prediction output. IEEE Trans. Multimed. **16**(6), 1665–1676 (2014)
16. Jain, A., Gupta, A., Davis. L.S.: Learning what and how of contextual models for scene labeling. In: Computer Vision—ECCV 2010. Springer, pp. 199–212 (2010)
17. Kjellström, H., Romero, J., Martínez, D., Kragić, D.: Simultaneous visual recognition of manipulation actions and manipulated objects. In: Computer Vision–ECCV 2008. Springer, pp. 336–349 (2008)
18. Lan, T., Yang, W., Wang, Y., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: Advances in Neural Information Processing Systems 23, pp. 1216–1224 (2010)
19. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1354–1361 (2012)
20. Lan, T., Wang, Y., Yang, W., et al.: Discriminative latent models for recognizing contextual group activities. IEEE Trans. Pattern Anal. Mach. Intell. **34**(8), 1549–1562 (2012)
21. Li, R., Porfilio, P., Zickler, T.: Finding group interactions in social clutter. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2722–2729 (2013)
22. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, IEEE, pp. 2929–2936 (2009)
23. Murphy, K., Torralba, A., Freeman, W.: Using the forest to see the trees: a graphical model relating features, objects and scenes. Adv. Neural Inf. Process. Syst. **16**, 1499–1506 (2003)
24. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: IEEE 11th international conference on Computer Vision, 2007. ICCV 2007, IEEE, pp. 1–8 (2007)
25. Ryoo, M.S., Aggarwal, J.K.: Spatio–temporal relationship match: video structure comparison for recognition of complex human activities. In: IEEE 12th International Conference on Computer Vision, IEEE, pp. 1593–1600 (2009)
26. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, IEEE, vol. 3, pp. 32–36 (2004)
27. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: Proceedings of the twenty-first international conference on Machine learning, ACM, p. 104 (2004)
28. Choi, W., Savarese, S.: Understanding collective activities of people from videos. IEEE Trans. Pattern Anal. Mach. Intell. **36**(6), 1242–1257 (2013)
29. Yao, B., Fei-Fei, L.: Grouplet: a structured image representation for recognizing human and object interactions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 9–16 (2010)
30. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 17–24 (2010)
31. Zhu, Y., Nayak, N., Roy-Chowdhury, A.: Context-aware modeling and recognition of activities in video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2491–2498 (2013)