Consistent Collective Activity Recognition with Fully Connected CRFs

Takuhiro Kaneko, Masamichi Shimosaka, Shigeyuki Odashima, Rui Fukui and Tomomasa Sato
Department of Mechano-Informatics, the University of Tokyo, Japan
{*kaneko, simosaka, odashima, fukui, tsato*}*@ics.t.u-tokyo.ac.jp*

Abstract

Recognizing collective human activities has gained attention. Collective activities are such as queueing in a line, talking together and waiting by an intersection. It is often hard to differentiate between these activities only by the appearance of the individual. Hence, recent works exploit the contextual information of other people nearby. However, these works do not take enough care of the spacial and temporal consistency in a group (e.g. considering the consistency in only adjacent area). To solve the problem, this paper describes a method to integrate individual recognition result via fully connected CRFs, which assume the relationships among all the people. Unlike previous methods that determine the range of human relations by heuristics, our method describes the "multi-scale" relationships in position, size, movement and time sequence as flexible potentials, so as to handle various types, sizes and shapes of groups. Experimental results show that our method outperforms state-of-the art methods.

1. Introduction

Collective activity recognition is one of the most challenging problems in computer vision, and actively studied [1, 2, 3, 7, 8]. Collective activities are activities performed by multiple persons: crossing, waiting, queueing, walking and talking. Since there are human interactions in collective activities, it is often hard to differentiate between these activities only by the appearance of the individual (see Fig. 1 in [3]). Hence, recent works exploit the contextual information of the others.

The ways of encoding the contextual information are categorized into the following three approaches: feature description approach, grid based approach, and graph structure approach. Feature description approaches include the contextual information in the feature descriptors [2, 7]. In these approaches, activity of each person is independently recognized, therefore, the spacial and temporal consistency in a group is not always ensured. To obtain the consistency, it is required to answer the following question: "Who is in the same group?" In reply to the question, grid based approach optimizes activities around each deformable grid [1], while graph structure approaches describe the relationship between each person in the graph structures [3, 8]. However, these works cannot describe the "multi-scale" relationships in various features such as position, size, movement and time sequence, although there exist various types, sizes and shapes of groups as shown in Fig. 1. The grid based approach [1] depends on the density and position of grids, therefore, it is difficult to exploit the long range relationships. One of the graph structure approaches [8] assumes that there is only one activity in a single image, therefore, cannot handle the scene where multiple groups exist. Another graph structure approach based on MRF [3] is intractable to include various features, and unable to handle a complicated graph structure such as fully connected model.

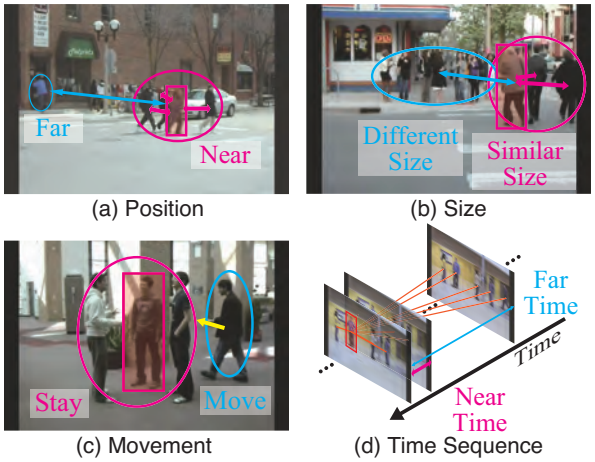

Figure 1. Who is in the same group? In dividing people into groups, there are various criteria such as (a) position, (b) size, (c) movement and (d) time sequence.

By contrast, our proposed method handles the "multi-scale" relationships in various features: position,

size, movement and time sequence. In particular, our method uses fully connected CRFs and describes human relationships as variable potentials. This approach is able to represent the various features over "multi-scale" in a single unified model. The calculation cost of fully connected model is intractable when estimating strictly, however, the cost is reduced to linear in the number of detected persons by describing the pairwise potentials with a Gaussian kernel [5].

In summary, the contributions of this paper are 1) to exploit various features to describe human relationships: position, size, movement and time sequence; 2) to describe the range of human relations not as constant values but as variable potentials; 3) to use fully connected CRFs to obtain the consistency over the "multi-scale" relationships. The experimental results show that our "multi-scale" model outperforms not only the unary only model but also state-of-the art models [2, 3, 7].

2. Consistent Collective Activity Recognition with Fully Connected CRFs

2.1. Model Overview

Our goal is to ensure the spacial and temporal consistency of activities in a group. For this purpose, our method uses conditional random fields (CRFs) [6]. CRF is a probabilistic framework for labeling and segmenting structured data and able to deal with various features in a single unified model. Specifically, in order to handle the "multi-scale" relationships, our method uses fully connected CRFs. Instead of specifying the range of human relations heuristically, our approach describes human relationships in position, size, movement and time sequence as flexible potentials, so as to deal with various types, sizes and shapes of groups.

A brief overview of our model is illustrated in Fig. 2. In the preprocessing, the persons in the images have been found. Next, features (e.g. histogram of oriented gradients (HOG) and optical flow) are extracted from the detected bounding box. Unary potential and pairwise potential are calculated using these features, and integrate them via fully connected CRFs. The technical details follow in Sec. 2.2.

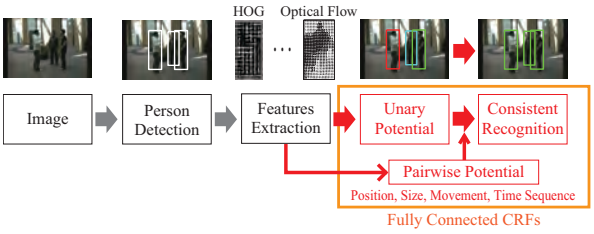

Figure 2. Overall process of our model.

2.2. Model Formulation

Fully Connected CRFs Model: Given a video, our method first detects persons by an efficient human detector [4]. The observed data from the detected persons are defined as $\boldsymbol{x} = \{x_1, ..., x_N\}$, where x_i is the observed data from the i-th person and N is the number of detected persons in the video. Let the corresponding activity labels be given by $\boldsymbol{y} = \{y_1, ..., y_N\}$. The domain of each variable y_i is a set of labels $\mathcal{L} = \{l_1, ..., l_K\}$, where K is the classes of labels. A conditional random field $(\boldsymbol{x}, \boldsymbol{y})$ is characterized by a Gibbs distribution:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp(-E(\boldsymbol{y})), \tag{1}$$

where $Z(\boldsymbol{x})$ is the partition function which normalizes the distribution, and $E(\boldsymbol{y})$ is the Gibbs energy:

$$E(\boldsymbol{y}) = \sum_i \psi_u(y_i) + \sum_i \sum_{j>i} \psi_p(y_i, y_j), \tag{2}$$

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

Unary Term: The unary potential is computed independently for each person, and encodes a distribution over the activity label y_i. The unary potential used in our implementation is described in Sec. 3.

Pairwise Term: The pairwise potential represents the relationship between each person. In the fully connected CRFs, the pairwise potential is computed for all the sets of persons as shown in Fig. 3 (c). In our model, the pairwise potential is defined as

$$\psi_p(y_i, y_j) = \mu(y_i, y_j) k(\boldsymbol{f}_i, \boldsymbol{f}_j), \tag{3}$$

where $\mu(y_i, y_j)$ is the label compatibility function given by Potts model: $\mu(y_i, y_j) = [y_i \neq y_j]$. It introduces a penalty for similar persons that are assigned different labels. The vector $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ are feature vectors for i-th and j-th persons, and $k(\boldsymbol{f}_i, \boldsymbol{f}_j)$ is the Gaussian kernel defined by the positions p_i and p_j, sizes s_i and s_j, movements m_i and m_j, times t_i and t_j, and weight w:

$$k(\boldsymbol{f}_i, \boldsymbol{f}_j) = w \exp\left(-\frac{|p_i - p_j|^2}{2\theta_1^2} - \frac{|s_i - s_j|^2}{2\theta_2^2} - \frac{|m_i - m_j|^2}{2\theta_3^2} - \frac{|t_i - t_j|^2}{2\theta_4^2} \right). \tag{4}$$

Note that we normalize positions and sizes by the median size of all the persons to describe the relationships as relative value rather than absolute value. Movement is calculated by subtracting the median optical flow without the bounding boxes from the mean optical flow within the bounding box. The former optical flow represents the camera movement, while the latter optical flow represents the person movement in the image. Optical flow is computed by the approach of Sun *et al.* [9].

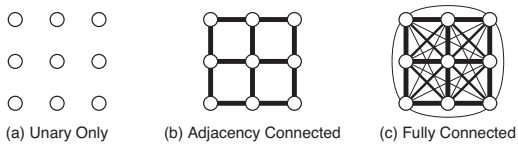

Figure 3. Node relationships in each graph structure.

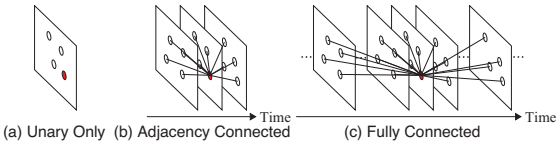

Figure 4. Connected range in each graph structure.

2.3. Inference and Learning

In inference, the maximum a posteriori (MAP) labeling of the random field is estimated:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}\in\mathcal{L}^N} P(\boldsymbol{y}|\boldsymbol{x}). \tag{5}$$

Since it is intractable to compute the exact distribution $P(\boldsymbol{y}|\boldsymbol{x})$ for all the sets of labels $\mathcal{L}^N$, our model uses a mean field approximation. The mean field approximation finds $Q(\boldsymbol{y}) = \prod_i Q_i(y_i)$ close to $P(\boldsymbol{y})$ in terms of minimizing the KL-divergence $\mathbf{D}(Q||P)$. A naive implementation of this approximation has quadratic complexity in the number of variables N. However, the pairwise potential in our model is defined by the Gaussian kernel, therefore, it is possible to use a highly efficient approximated inference algorithm via high-dimensional filtering [5]. This reduces the calculation complexity from quadratic to linear in the number of variables N.

In learning, the kernel parameters w, θ_1, θ_2, θ_3 and θ_4 are estimated. Due to non convexity of kernel width $\theta_1, \theta_2, \theta_3, \theta_4$ on log-loss criterion, it is hard to optimize them globally, therefore, we use grid search from the training set with cross-validation.

3. Experiments

We evaluate our model on the collective activity dataset [2]. 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. Some videos include multiple groups or activity transition. All the persons in every 10th frame are labeled with the ground truth: pose, activity and bounding box information.

Implementation: To evaluate our graph structure model for unary only model, we use action context (AC) descriptor [7] as the baseline. AC descriptor is one of state-of-the art methods based on contextual feature description. Actions are defined by combining poses and activities [8]. In our implementation, the unary potential in (2) is defined as $\psi_u(y_i) = -\log(\mathrm{prob}(y_i))$, where $\mathrm{prob}(y_i)$ represents the probability that the activity of i-th person is y_i. $\mathrm{prob}(y_i)$ is calculated by normalizing the score of multi-class SVM classifier on AC descriptor. To convert SVM score into probability, sigmoid function is used.

Quantitative Results: We use the same leave-one-out scheme described in [2, 3, 7] to compare fairly. When one video is used as the testing set, the other videos are used as the training set. The confusion matrix of our method using fully connected CRFs and the baseline using only unary potential are shown in Fig. 5. In the confusion matrix, rows represent ground truth and columns represent prediction. Each row is normalized to sum to 1. These confusion matrices show that our model achieves a significant improvement over the unary only model. Note that walking vs crossing is still ambiguous in our model, because whether walking or crossing often depends on not human relationships but environmental settings: a sidewalk or a pedestrian crossing.

We also compare our method with recent works in Table 1. The first row to the fifth row show the results using unary only models such as Fig. 4 (a), while the sixth row to the ninth row show the results using graph structure models such as Fig. 4 (b) (c). The first, fourth, seventh and ninth rows show the results by our implementation. The first row shows the result using HOG without the contextual information. The fourth row shows the result using AC descriptor with the contextual information. To evaluate fully connected CRFs model, we compare adjacency connected CRFs (AC-CRF) model in the seventh row, with fully connected CRFs (FC-CRF) model in the ninth row. AC-CRF model considers human relationships in the adjacency frames as shown in Fig. 4 (b). Note that state-of-the art method (RSTV + MRF) [3] needs the trajectory data of each person to obtain the consistency via 3D MRF, however, our method does not need the surplus data.

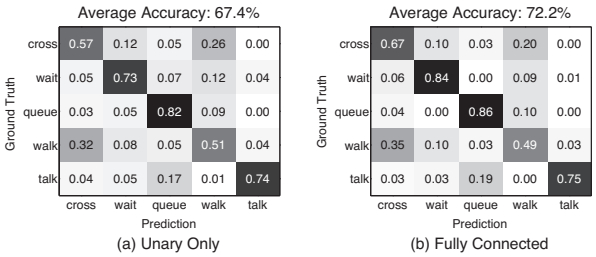

Figure 5. Confusion matrices for activity classification.

Table 1. Comparison of activity classification accuracies for different methods.

Method	Average Accuracy
HOG	50.0%
STV in [2]	64.3%
RSTV in [3]	67.2%
AC in ours	67.4%
AC in [7]	68.2%
STV + MC in [2]	65.9%
AC + AC-CRF	69.6%
RSTV + MRF in [3]	70.9%
AC + FC-CRF	**72.2%**

Qualitative Results: Example results are presented in Fig. 6-7. Fig. 6 shows success and failure examples. The labels C (magenta), S (blue), Q (cyan), W (red), T (green) and NA (white) indicate crossing, waiting, queueing, walking, talking and not assigned. Top two rows show examples of successful classification and bottom row shows examples of false classification. Fig. 7 shows example results in the scene where multiple groups exist. Top two rows show examples of being consistent in groups and bottom row shows examples of being inconsistent in groups. These results demonstrate that our method is robust for temporal false recognition and able to handle the multiple existence of groups.

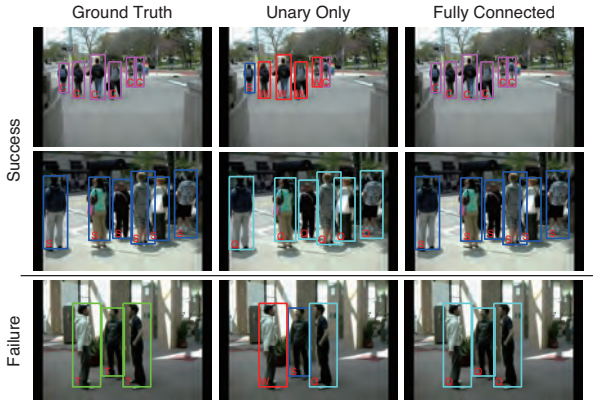

Figure 6. Example results of collective activity recognition.

4. Conclusion and Discussion

This paper has described a method for consistent collective activity recognition with fully connected CRFs, which assume the relationships among all the people. Our model leverages various features such as position, size, movement, and time sequence in a single unified model, and describes the "multi-scale" relationships in these features as flexible potentials, so as to handle various types, sizes and shapes of groups. Experimental results demonstrate that our model is robust for temporal false recognition, and able to deal with multiple existence of groups. Evaluation results on the collective activity dataset show that our method outperforms state-of-the art methods, as well as the method using unary only model.

At the present time, our method is based on batch processing. In the future, we hope to extend our method to online processing for online applications.

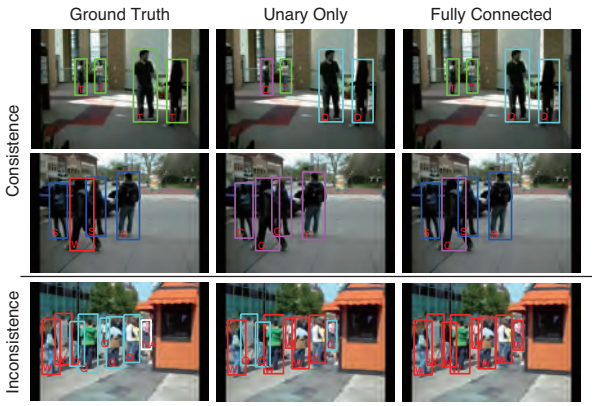

Figure 7. Example results in the scene where multiple groups exist.

References

[1] M. R. Amer and S. Todorovic. A chains model for localizing participants of group activities in videos. In *ICCV*, 2011.

[2] W. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *International Workshop on Visual Surveillance*, 2009.

[3] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *CVPR*, 2011.

[4] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[5] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *Adv. in NIPS 24*, 2011.

[6] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[7] T. Lan, Y. Wang, and G. Mori. Retrieving actions in group contexts. In *International Workshop on Sign Gesture Activity*, 2010.

[8] T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions: Discriminative models for contextual group activities. In *Adv. in NIPS 23*, 2010.

[9] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *CVPR*, 2010.