Pattern Recognition Letters 43 (2014) 109–118

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A fully connected model for consistent collective activity recognition in videos ☆

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

Department of Mechano-Informatics, The University of Tokyo, Tokyo, Japan

ARTICLE INFO

Article history:
Available online 14 February 2014

Keywords:
Collective activity recognition
Fully connected model
CRFs
Spatial and temporal consistency

ABSTRACT

We propose a novel method for consistent collective activity recognition in video images. Collective activities are activities performed by multiple persons, such as queuing in a line, talking together, and waiting at an intersection. Since it is often difficult to differentiate between these activities using the appearance of only an individual person, the models proposed in recent studies exploit the contextual information of other people nearby. However, these models do not sufficiently consider the spatial and temporal consistency in a group (e.g., they consider the consistency in only the adjacent area), and therefore, they cannot effectively deal with temporary misclassification or simultaneously consider multiple collective activities in a scene. To overcome this drawback, this paper describes a method to integrate the individual recognition results via fully connected conditional random fields (CRFs), which consider all the interactions among the people in a video clip and alter the interaction strength in accordance with the degree of their similarity. Unlike previous methods that restrict the interactions among the people heuristically (e.g., within a constant area), our method describes the "multi-scale" interactions in various features, i.e., position, size, motion, and time sequence, in order to allow various types, sizes, and shapes of groups to be treated. Experimental results on two challenging video datasets indicate that our model outperforms not only other graph topologies but also state-of-the art models.

1. Introduction

Vision-based human activity recognition is of scientific and practical importance, and has been actively studied in the research field of computer vision. Many previous studies focused on recognizing actions performed by a single person in a video clip [3,23,25]. However, in real-world applications, such as surveillance monitoring, the previous methods are inapplicable, since human actions are rarely performed by a single person, but instead by multiple persons. For example, it is difficult to differentiate between the activities of the two persons shown in Fig. 1(a), by considering the appearance of the individual person. In order to recognize activities performed by multiple persons, which we call "collective activities," it is necessary to exploit the contextual information of the people nearby. When we have identified the activities of people nearby, it immediately becomes clear that the left person in Fig. 1(a) is queuing and the right person is talking, as shown in Fig. 1(b).

In some recent studies, methods have been proposed for collective activity recognition using the contextual information of people nearby. Choi et al. [6], Lan et al. [20], and Kaneko et al. [14] encoded the contextual information by exploiting the feature descriptors extracted from a focal person and his/her surrounding area. These descriptors are more effective than feature descriptors without contexts (e.g., histogram of oriented gradients (HOG) [8]). However, in the models, the activity of each person is classified independently, and therefore, the spatial and temporal consistency in a group is not always ensured.

In order to obtain this consistency, the question "Which people are in the same group?" must be answered, and an activity in each group must be optimized. To answer the question, Amer and Todorovic [2] optimized activities around deformable grids, while Lan et al. [21], Choi et al. [7], Choi and Savarese [5], and Khamis et al. [15,16] used graph structures that describe the interactions between persons. However, the models used in these studies cannot describe the "multi-scale" interactions in various features, such as position, size, motion and time sequence, although there exist various types, sizes, and shapes of groups, as shown in Fig. 2. The model proposed by Amer and Todorovic [2] depended on the density and position of the grids, and therefore, it was difficult to exploit long-range relationships. In the model proposed by

☆ This paper has been recommended for acceptance by A. Del Bimbo.

* Corresponding author. Tel.: +81 3 5841 6333.
E-mail address: kaneko@ics.t.u-tokyo.ac.jp (T. Kaneko).

http://dx.doi.org/10.1016/j.patrec.2014.02.002
0167-8655/

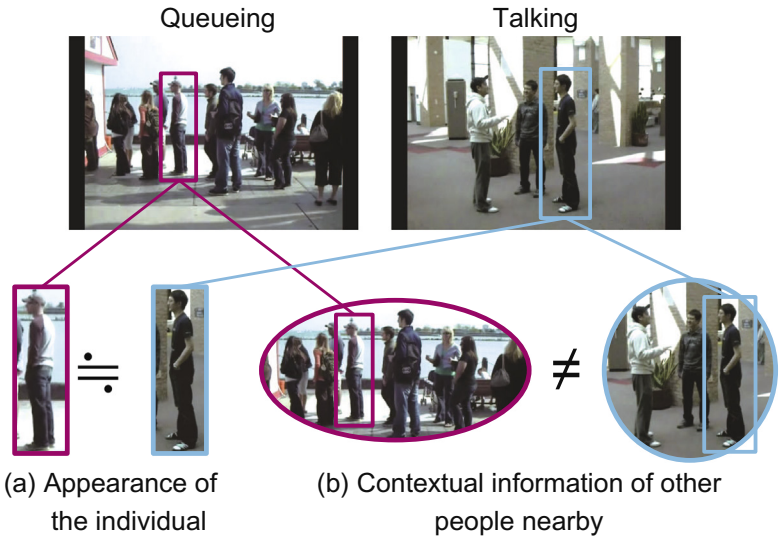

Fig. 1. Useful contexts for collective activity recognition. It is often difficult to differentiate between collective activities by the appearance of only an individual person (a). When we have identified the activities of people nearby, it immediately becomes clear that the left person is queuing and the right person is talking (b).

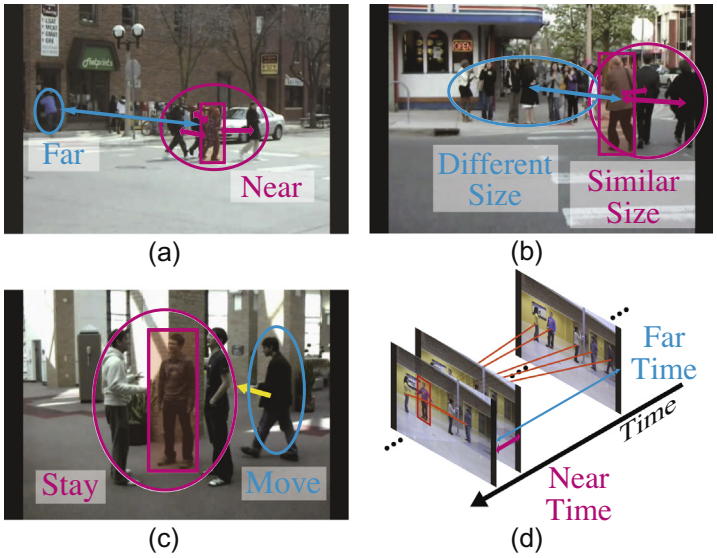

Fig. 2. Which people are in the same group? For dividing people into groups, various criteria, such as (a) position, (b) size, (c) motion, and (d) time sequence, can be used.

Lan et al. [21], the person–person interactions were latent and learned automatically; however, their model was restricted to modeling contextual information in a single frame, and was not designed such that temporal consistency was ensured. Choi et al. [7] and Khamis et al. [15,16] considered temporal consistency for a person or group; however, in their model, the person–person interactions that were considered were restricted only in consecutive frames to compute reasonably. The results of these models are likely to be affected by temporary misclassification. Choi and Savarese [5] exploited a hierarchical model to classify collective activities jointly; however, they assumed there exists only one collective activity in a certain time frame. Therefore, the method cannot model multiple collective activities in a scene, such as that shown in Fig. 2(b), where some persons are waiting at a street intersection, while others are crossing. Considering real-world applications, such as surveillance monitoring, this assumption is not natural.

In contrast, our proposed model describes the "multi-scale" interactions in various features, i.e., position, size, motion, and time sequence. This means that our model is able not only to describe the long-range relationships among people in both time and space, but also to consider multiple collective activities in a certain time frame. In particular, we use fully connected conditional random fields (CRFs), which consider all the interactions among the people in a video clip, and alter the interaction strength according to the degree of their similarity. This model is able to represent the various features over a "multi-scale" in a single unified model. In general, the calculation cost of a fully connected model is intractable when strict estimation is conducted; however, the cost is reduced to linear in the number of detected persons using a highly efficient approximation method in which the pairwise potentials are modeled using Gaussian kernels [18].

We summarize the main contributions of this paper. (1) We propose a novel method for consistent collective activity recognition in video images using a fully connected model. In the model, we do not restrict the person–person interactions that are considered heuristically, but instead consider all the interactions among the people in a video clip. (2) We describe the person–person