Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

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

Humans in groups: The importance of contextual information for understanding collective activities

Nicoletta Noceti *, Francesca Odone

DIBRIS - Università di Genova, via Dodecaneso, 35 16146-IT, Italy

ARTICLE INFO

Article history:
Received 1 October 2013
Received in revised form
29 April 2014
Accepted 12 May 2014

Keywords:
Collective activity recognition
People spatial orientation classification
Context-aware people description
Graph kernel
Semi-supervised learning

ABSTRACT

In this work we consider the problem of modeling and recognizing collective activities performed by groups of people sharing a common purpose. For this aim we take into account the social contextual information of each person, in terms of the relative orientation and spatial distribution of people groups. We propose a method able to process a video stream and, at each time instant, associate a collective activity with each individual in the scene, by representing the individual – or target – as a part of a group of nearby people – the target group. To generalize with respect to the viewpoint we associate each target with a reference frame based on his spatial orientation, which we estimate automatically by semi-supervised learning. Then, we model the social context of a target by organizing a set of instantaneous descriptors, capturing the essence of mutual positions and orientations within the target group, in a graph structure. Classification of collective activities is achieved with a multi-class SVM endowed with a novel kernel function for graphs. We report an extensive experimental analysis on benchmark datasets that validates the proposed solution and shows significant improvements with respect to state-of-art results.

1. Introduction

Understanding human activities automatically is a challenging task which finds application in a variety of domains including video retrieval, video-surveillance, and human–machine (or human-robot) interaction. Also, in the last few years there has been an increasing interest in analyzing actions performed in dynamic environments and understanding the activities of people groups. In all these cases interactions among people and contextual information gain a strong semantic relevance.

The importance of context has first emerged in scene understanding [6,48,25], but can also be appreciated in social signal processing [50]. Indeed, social cues are knowingly more likely to be misinterpreted if they are taken out of their context.

Inspired by these observations, in this work we consider the problem of modeling and recognizing complex human activities involving more than one individual. Such activities, commonly referred to as *collective activities*, can be better understood and disambiguated by analysing the social context of the person, formed by nearby people, instead than observing one person at a time. Typical collective activities include, for instance, people standing in a queue or being involved in a conversation, but the application is not limited to daily life situations. In Fig. 1 we report some examples depicting actions from a soccer game. If considered singularly, the information we can infer about each player might be limited to the specific action per-se (e.g. running in Fig. 1 (a)). What we cannot infer is their final *purpose* – in this case participating to semantically different soccer actions – which is clearer when looking at their context. Also, single actions might be misinterpreted due to the lack of information: In Fig. 1(b) the players may appear as standing, in their context it is apparent they are participating to a jumping action.

The method we propose to classify collective activities analyses a video stream and associates each person in the scene at a given time instant t with a potential collective activity, learnt by observing the person (or target) and people in his/her vicinity (the target group). To generalize with respect to the observation view-point, we associate each target group with a 2D reference frame relative to the target apparent orientation. Each target is represented by two descriptions which account for the orientations frequencies and the spatial layout of people in the target group. Further, we characterize each target group with a dynamic descriptor which evaluates the affinity of the group motion with respect to the collective activities of interest. Then the information related to the whole group is organized in a graph structure which

* Corresponding author. Tel.: +39 0103536626; fax: +39 0103536699.
E-mail addresses: nicoletta.noceti@unige.it (N. Noceti), Francesca.Odone@unige.it (F. Odone).

http://dx.doi.org/10.1016/j.patcog.2014.05.008
0031-3203/

Fig. 1. Examples of soccer players, whose actions can be better interpreted if considering their context (see text).

contains the essence of mutual positions and orientations within the group. Finally with a multi-class SVM classification architecture, we associate each target with the most appropriate collective activity. Since each target is represented by a graph, we devise a kernel function based on the spectral method for graph matching proposed in [33]. The outputs are finally regularized over time to get to the final activity labels.

Since our method strongly relies on a robust estimate of people spatial orientations, we also propose a novel approach to address this task. We describe the bounding box of each person using HOG descriptors [15] and apply a multi-class structured sparsification technique [20] to select only the most relevant groups of features. Then we classify each element with respect to a quantized orientation, by means of semi-supervised learning [4] to deal with the possible ambiguity of the data annotations, which may be due to the smooth transitions between classes of adjacent orientations. Finally, we model the multi-class classification problem with a binary all-vs-all strategy and make use of a decoding matrix to associate each sample with a unique label.

Unlike previous approaches to the classification of people appearance with respect to the camera viewpoint [2,3,10] our work takes into account explicitly the peculiarities and the ambiguity of the data by means of semi-supervised learning which, to the best of our knowledge, has never been adopted in this context. Similar to [10] we apply a feature selection step, but our method is able to consider all the classes simultaneously, and so reduce the amount of features to be computed.

As for the problem of classifying collective activities, our work shares similarities with [12,13]. However, unlike these approaches where the basic descriptors have an inherent temporal component, our approach is frame-based. The organization of such instantaneous descriptors in a more global and structured graph representation allows us to improve the classification stage.

In summary, the main contributions of our work are (i) a robust method for the classification of spatial orientation of people and (ii) an architecture for modeling and recognizing collective activities which relies on a novel kernel function for graphs tailored for our representations.

We validate our contributions on two benchmark datasets, the *TUD Multiview Pedestrian* dataset [2] and the *Collective Activity* dataset [12], which are a reference for the classification of orientations and of collective activities respectively. The experimental analysis shows that our approaches perform better or comparably than state-of-the-art methods on the two problems.

Notice that, although we may account for temporal information, the structure of our method allows us to provide an instantaneous feedback, if needed, by skipping the final temporal regularization. The only temporal descriptor we consider – the instantaneous velocity – can be approximated by two adjacent frames only, unlike other methods in the literature [12,13] which require a longer temporal window. As we will see in the experimental session, our frame-based results are very close to state of the art methods accounting for temporal sequences.

The remainder of the paper is organized as follows. Section 2 discusses the related works, while Section 3 briefly reviews the schema of our approach. Section 4 and 5 are devoted to present in details the methods we propose for classifying spatial orientation of people and collective activities, respectively. We report the experimental analysis in Section 6, while Section 7 is left to conclusions.

2. Related works

The problem of characterizing people with respect to *body pose and orientation* has been addressed from different perspectives. Accurate pose estimation has been based mainly on the estimate of position and orientation of body parts, as head, face, eyes or limbs [8,9,18,19,26]. This specific class of approaches usually address 2D or 3D body modeling, resulting in a 2D body skeleton or a more accurate 3D model.Other approaches aim at estimating the overall body orientation relative to the camera view point, a line of research not largely developed yet, with the exception of a few recent works. Andriluka et al. [2] represent people appearance with HOG and then classify their relative orientation by means of SVM. The obtained labels are exploited to reduce the ambiguity of a 3D pose estimate. Alternatively, body location and pose may be estimated jointly, as in [10], where people are described with a multi-level HOG sparsified by minimizing a regularized functional with ℓ_1-penalty. Recently, Baltieri et al. [3] described the people appearance with multi-level HOG and then classify their orientations using an array of Extremely Randomized Trees classifiers. Their output is integrated in a global probability density function using a Mixture of Approximated Wrapped Gaussian distributions.

Please cite this article as: N. Noceti, F. Odone, Humans in groups: The importance of contextual information for understanding collective activities, Pattern Recognition (2014), http://dx.doi.org/10.1016/j.patcog.2014.05.008