Computer Vision and Image Understanding 143 (2016) 11–24

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Detecting conversational groups in images and sequences: A robust game-theoretic approach☆

Sebastiano Vascon[a,*], Eyasu Z. Mequanint[b], Marco Cristani[a,c], Hayley Hung[d], Marcello Pelillo[b], Vittorio Murino[a,c]

[a] Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Genova, Italy
[b] Deptartment of Environmental Sciences, Informatics and Statistics, University Ca' Foscari of Venice, Italy
[c] Deptartment of Computer Science, University of Verona, Italy
[d] Faculty of Electrical Engineering, Mathematics and Computer Science, Technical University of Delft, Netherlands

ARTICLE INFO

Article history:
Received 9 February 2015
Accepted 18 September 2015
Available online 1 October 2015

Keywords:
Group detection
F-formation detection
Conversational groups
Game-theory
Scene understanding

ABSTRACT

Detecting groups is becoming of relevant interest as an important step for scene (and especially activity) understanding. Differently from what is commonly assumed in the computer vision community, different types of groups do exist, and among these, standing conversational groups (a.k.a. F-formations) play an important role. An F-formation is a common type of people aggregation occurring when two or more persons sustain a social interaction, such as a chat at a cocktail party. Indeed, detecting and subsequently classifying such an interaction in images or videos is of considerable importance in many applicative contexts, like surveillance, social signal processing, social robotics or activity classification, to name a few. This paper presents a principled method to approach to this problem grounded upon the socio-psychological concept of an F-formation. More specifically, a game-theoretic framework is proposed, aimed at modeling the spatial structure characterizing F-formations. In other words, since F-formations are subject to geometrical configurations on how humans have to be mutually located and oriented, the proposed solution is able to account for these constraints while also statistically modeling the uncertainty associated with the position and orientation of the engaged persons. Moreover, taking advantage of video data, it is also able to integrate temporal information over multiple frames utilizing the recent notions from multi-payoff evolutionary game theory. The experiments have been performed on several benchmark datasets, consistently showing the superiority of the proposed approach over the state of the art, and its robustness under severe noise conditions.

1. Introduction

The visual analysis of groups is becoming more and more widespread in computer vision, after decades of research on the automated modeling of individuals (which still remains an open problem), the goal has moved from encoding simple actions performed by a single subject to capturing dyads or clusters of social interactions [1–10]. This is of extreme importance in many fields and applications, also addressing social and life sciences [11,12]. This seems to be a necessary step, since humans are essentially a social species, as demonstrated by the fact that in everyday life people continuously interact with each other to achieve goals or simply to exchange states of mind. In this paper, we exploit a recent taxonomy presented in [13], which indicates that many types of groups can be defined. In particular, we target standing conversational groups, also known as *F-formations* [14], that is, groups of people who spontaneously decide to be in each other's immediate presence to converse with each and every member of that group.

Standing conversational groups are of primary importance in many contexts, such as video surveillance [7], social signal processing [1,2,4,6], multimedia [3], social robotics [15], and activity recognition [16], as we will discuss extensively in Section 2.

Many studies have been carried out by social psychologists to understand how people behave in public. By exploiting the theory behind these findings, we propose novel and more socio-psychologically principled ways of designing methods for automatically analyzing human behavior. For example, Hall [17] proposed that relationships and levels of interactions could be inferred by considering different physical distances.

Goffman [18] observed that group interactions can be categorized into those that are 'focused' and those that are 'unfocused'. Focused

☆ This paper has been recommended for acceptance by Gang Hua.

* Corresponding author.
E-mail address: sebastiano.vascon@iit.it, me@xwasco.com (S. Vascon).

http://dx.doi.org/10.1016/j.cviu.2015.09.012
1077-3142/

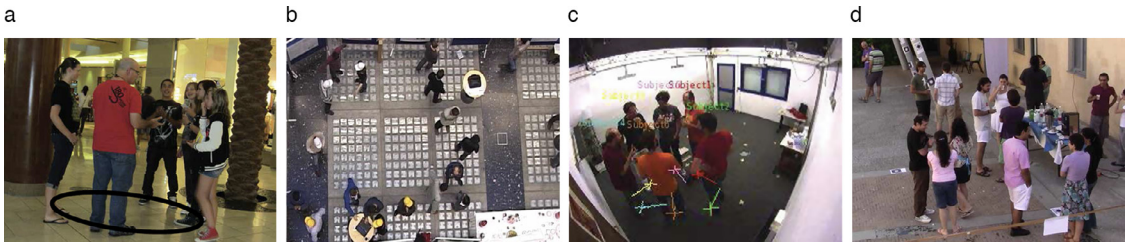

Fig. 1. Standing conversational groups: (a) in black, graphical depiction of overlapping space within an F-formation: the o-space; (b) a poster session in a conference, where different groupings are visible; (c) circular F-formation; (d) a typical surveillance setting where camera is located at 2.5–3 m from the floor, for which detecting groups is challenging.

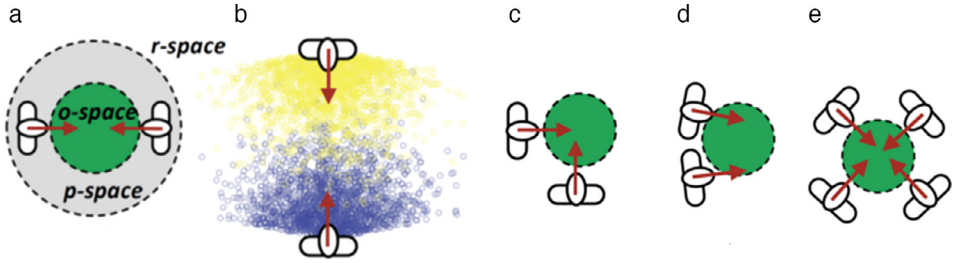

Fig. 2. F-formations; (a) components of an F-formation: o-space, p-space, r-space; in this case, a face-to-face F-formation is sketched; (b) modeling the frustum of attention by particles: in the intersection stays the o-space; (c) L-shape F-formation; (d) side-by-side F-formation; (e) circular F-formation.

interactions concern the gathering of people to participate in an activity where there is a common focus, such as playing and watching a football match, conversing, or marching in a band. Unfocused encounters involves light interactions such as avoiding people on a busy street, briefly greeting a colleague while passing them in the corridor, or indicating to let someone pass when boarding a train. This taxonomy has been exploited recently in [13] for addressing F-formations.

Within the class of focused encounters, the F-formation is a specific type of group interaction which requires more attention from our senses. Specifically, an F-formation arises "whenever two or more individuals in close proximity orient their bodies in such a way that each of them has an easy, direct and equal access to every other participant's transactional segment, and when they maintain such an arrangement" [19, p. 243]. Some examples of F-formations in real-world situations are illustrated in Fig. 1a. There can be different F-formations as shown in Fig. 2a–e. In the case of two participants, typical F-formation arrangements are vis-a-vis, L-shape, and side-by-side.

Three social spaces emerge from an F-formation: the o-space, the p-space and the r-space. The most important part is the o-space (see Fig. 2), a convex empty space surrounded by the people involved in a social interaction, in which every participant looks inward, and no external people are allowed. The p-space is a narrow strip that surrounds the o-space, and that contains the bodies of the conversing people, while the r-space is the area outward the p-space.

Our goal in this paper is to develop a robust approach to automatically detect F-formations from images and videos employing a single monocular camera. As input, the approach requires the position of the persons in the scene on the ground plane as well as their body orientation, although in most cases, head orientation is more readily captured, even under heavy occlusions. These cues are easily obtainable nowadays, even if they are not estimated very accurately, and many approaches are aimed at extracting such information from raw images/videosequences [4,20,21]. Among the few approaches of F-formation detection, a recent experimental work of Setti et al. [22] shows that substantial improvement in the performance of F-formation detection algorithms can be achieved by combining a probabilistic approach (as [7]) and graph-based clustering methods [6]. Motivated by these studies, we develop a new sociologically-based approach which combines in a natural way the modeling of the uncertainty in the position and orientation of the subjects and a game-theoretic clustering approach , allowing one to extract coherent groups in edge-weighted graphs, digraphs and hypergraphs [23,24]. The game-theoretic setting provides a conceptual framework which allows us to integrate temporal information in a principled way, in an attempt to reliably extract groups in video sequences under severe noisy conditions. This is done by using a recent approach to integrate multiple payoff functions in an evolutionary game-theoretic setting [25].

This work represents a substantial contribution to group detection in real scenarios. To date in computer vision, grouping behaviors have been analyzed mainly in dynamic situation via tracking, exploiting the oriented velocity as a primary cue, for example by associating individuals' tracklets [26–34]. In our case, F-formation are manifested primarily when people are still, so that a finer yet robust analysis is required. Our approach considers in fact the detection of groups in both still images and videos.

To test the effectiveness of the proposed approach, we performed extensive experiments over five different datasets, each one representing a particular scenario. In particular, we used a synthetic dataset [7], the Coffee Break dataset [7], the GDet dataset [7], the Idiap Poster data dataset [6], the Cocktail Party [5] dataset and two new dataset, one proposed by Choi et al. [35] and FriendsMeet2 that we propose in this work. We also carried out systematic noise resilience experiments to fully investigate the stability and robustness of our method. The results consistently show the superior or comparable performances of the proposed approach over the state of the art.

The rest of the paper is organized as follows. A detailed review of the literature on group detection approaches is presented in Section 2. Our approach is detailed in Section 3. In Section 4 we describe the game-theoretic clustering approach we use to extract F-formations and its extension to multiple affinity matrices. Finally, Section 5 presents the experimental results and Section 6 concludes the paper.

2. Literature review

2.1. Groups

During multi-party activities, we expect that there is a different underlying structure that governs the behavior of groups compared to individuals acting independently. For example, there has been