Int J Comput Vis
DOI 10.1007/s11263-017-1026-6

Joint Estimation of Human Pose and Conversational Groups from Social Scenes

Jagannadan Varadarajan[1] · **Ramanathan Subramanian**[2,3] · **Samuel Rota Bulò**[4,5] · **Narendra Ahuja**[1,6] · **Oswald Lanz**[5] · **Elisa Ricci**[5,7]

Received: 15 March 2016 / Accepted: 2 June 2017

Abstract Despite many attempts in the last few years, automatic analysis of social scenes captured by wide-angle camera networks remains a very challenging task due to the low resolution of targets, background clutter and frequent and persistent occlusions. In this paper, we present a novel framework for jointly estimating (i) head, body orientations of targets and (ii) conversational groups called *F-formations* from social scenes. In contrast to prior works that have (a) exploited the limited range of head and body orientations to jointly learn both, or (b) employed the mutual head (but not body) pose of interactors for deducing F-formations, we propose a weakly-supervised learning algorithm for joint inference. Our algorithm employs body pose as the primary cue for F-formation estimation, and an alternating optimization strategy is proposed to iteratively refine F-formation and pose estimates. We demonstrate the increased efficacy of joint inference over the state-of-the-art via extensive experiments on three social datasets.

Keywords Head and body pose estimation · F-formation estimation · Semi-supervised learning · Convex optimization · Conversational groups · Video surveillance

Communicated by Bernt Schiele.

This work is supported by the research grant for the Human-Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR). We thank NVIDIA for GPU donation.

Electronic supplementary material The online version of this article (doi:10.1007/s11263-017-1026-6) contains supplementary material, which is available to authorized users.

✉ Jagannadan Varadarajan
vjagan@gmail.com

Ramanathan Subramanian
ramanathan.Subramanian@glasgow.ac.uk

Samuel Rota Bulò
rotabulo@fbk.eu

Narendra Ahuja
n-ahuja@illinois.edu

Oswald Lanz
lanz@fbk.eu

Elisa Ricci
eliricci@fbk.eu

1 Advanced Digital Sciences Center, Singapore, Singapore

2 International Institute of Information Technology, Hyderabad, India

3 University of Glasgow, Glasgow, UK

4 Mapillary Research, Graz, Austria

5 Fondazione Bruno Kessler, Trento, Italy

6 University of Illinois Urbana Champaign, Champaign, IL, USA

7 Department of Engineering, University of Perugia, Perugia, Italy

1 Introduction

Major strides in computer vision research have made head and body pose[1] estimation possible even when pedestrians are captured at prohibitively low-resolution by distant cameras in public spaces. Under such conditions, the scene is cluttered and facial and body parts appear blurred; also, per-

[1] We use the term *pose* to refer to orientation in the ground plane (pan) rather than the articulated spatial configuration of the human body. In line with several previous works (Benfold and Reid 2011; Chen and Odobez 2012), we will use the terms *pose* and *orientation* interchangeably.

sons tend to move unconstrained in the environment with (typically) uneven illumination. Recent pose estimation algorithms have been shown to be robust to facial appearance variations (Chamveha et al. 2013; Yan et al. 2016), and can also learn with little training data by exploiting anatomic constraints (Benfold and Reid 2011; Chen and Odobez 2012). Consequently, vision-based algorithms are now equipped to handle complex phenomena like *social interactions*.

Being able to determine conversational groups or *F-formations* (Ciolek and Kendon 1980) in social scenes (Fig. 1) can facilitate social computing, surveillance and robotics. F-formation (FF) is a loose geometric arrangement of interactors arising naturally in conversational settings. It is defined by the interactors who are in close proximity, and orienting their bodies such that each has equal, exclusive and unhindered access to the convex *O-space* between them (Ciolek and Kendon 1980). For instance, one can expect vis-a-vis or L-shape FF arrangements when two persons are interacting as seen in Fig. 1. Many works Cristani et al. (2011), Vascon et al. (2014) have exploited the fact that FFs are characterized by the shared physical locations and head, body orientations of interactors, and characterize an FF via group members and the O-space center (Ciolek and Kendon 1980).

Analyzing conversational groups in surveillance settings is highly challenging. Determining head and body pose of interactors is non-trivial due to (i) low facial resolution coupled with frequent and extreme (facial and bodily) occlusions, (ii) background clutter and (iii) body pose modeling owing to clothing variability. Furthermore, employing cues such as walking direction (Benfold and Reid 2011; Chen and Odobez 2012) is ineffective for social scenes as FFs denote relatively static arrangements. State-of-the-art FF detection methods (Cristani et al. 2011; Vascon et al. 2014) rely on pre-trained classifiers upon quantizing the range of possible head movements. Nevertheless, FF discovery is hard when pose classifiers are not adapted to the considered social scene, and some works therefore use annotated pose information to this end.

Addressing the above problems, we propose *joint* estimation of target head, body orientations and FFs from social scenes captured in surveillance settings. Different from prior works that have focused either on (i) joint learning of head and body pose exploiting human anatomic constraints (Benfold and Reid 2011; Chen and Odobez 2012), or (ii) FF discovery when positional and head pose information of interactors are precomputed (Cristani et al. 2011; Vascon et al. 2014), we present a unified framework to infer both (i) and (ii). Our model exploits the synergetic interaction-interactor relationship, i.e., FFs are characterized by mutual scene locations and head, body pose of interactors, who conversely are constrained in terms of the possible range of head and body orientations they can exhibit, motivating the need for joint learning. Specifically, our novel learning framework (i) exploits labeled and unlabeled data via manifold regularization to learn the range of jointly possible head-body orientations, (ii) exploits positional and pose-based constraints relating interactors to discover FFs, and (iii) iteratively refines pose estimates of interactors based on FF knowledge and vice-versa.

The salient aspects of our work are the following. (1) In contrast to prior works, we mainly use targets' body orientation for discovering F-formations. While previous works (Cristani et al. 2011; Vascon et al. 2014) have acknowledged the utility of body pose for deducing FFs, they still employ head pose as the primary cue for FF discovery given the difficulty in estimating body pose under extreme occlusions. Nevertheless, using head orientation is fundamentally spurious as it can frequently change during social interactions. In contrast, body pose is an inherently more stable cue, and can more accurately define the geometrical FF arrangement (Fig. 1). (2) In order to robustly estimate body pose under occlusions, our learning framework limits the possible range of body orientations based on the head pose to achieve joint head and body pose learning, as in Chen and Odobez (2012), and tackles varying levels of body occlusion by incorporating multiple, occlusion-adaptive regression functions. Our approach to jointly estimate head, body orientations and FF is inspired by the coupled head and body pose estimation framework proposed in Chen and Odobez (2012). However, we differ from Chen and Odobez (2012) as we couple head and body pose estimation with F-formation detection. (3) We explore a novel methodology for FF detection and propose an algorithm where each target votes for its O-space center, thereby indirectly modeling FF discovery as a clustering problem. (4) Temporal consistency is also enforced to ensure smoothness in pose and FF estimates over time.

To summarize, the main research contributions of this work are: (i) We present joint estimation of F-formations and targets' head, body orientations in distant social scenes. Via thorough experiments on three challenging datasets, we demonstrate the benefits of our joint learning framework against competing pose and FF detection approaches. (ii) Different from prior works, we employ body pose as the primary handle for discovering FFs. Robust computation of body pose estimates is achieved via coupled learning of head and body pose, and knowledge gained regarding FFs. Furthermore, body occlusions are handled via the use of multiple, occlusion-adaptive pose regressors. (iii) Enforcing temporal consistency with respect to estimated pose (classes) and FF memberships is particularly useful while analyzing surveillance scenes, where low video resolution and occlusions make target localization and facial feature extraction challenging and considerably error-prone.

This paper improves over our previous work (Ricci et al. 2015) in the following ways. An in-depth review of related

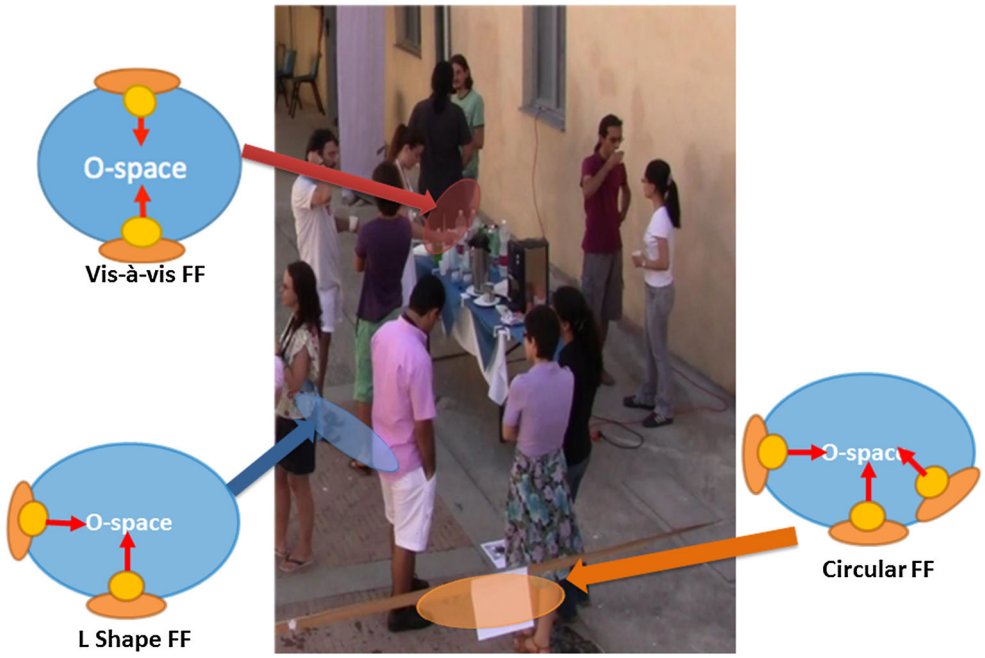

Fig. 1 Problem overview (*left*) social scene from the Coffee break dataset (Cristani et al. 2011). We *jointly* estimate conversational groups and the head, body pose of conversing targets. Exemplar F-formations are denoted by *ellipses* connecting feet positions of conversing targets, and the type of each F-formation is also specified

work is presented to better motivate the need for our framework. Furthermore, the proposed joint inference algorithm is elaborately detailed along with the associated optimization procedure, and our experimental evaluation is significantly extended with the use of a third, considerably more challenging dataset (as compared to the two used in Ricci et al. 2015). Remainder of the paper is organized as follows. The following section outlines related work on pose estimation and social scene analysis to motivate the need for our framework. Section 3 describes our algorithm and its salient aspects, and experimental results to demonstrate the benefits of our approach are presented in Sect. 4. We conclude with key remarks in Sect. 5.

2 Related Work

Research areas closely related to this work are (i) head and body pose estimation (HPE and BPE) from surveillance videos, (ii) semi or weakly-supervised learning, and (iii) detection of face-to-face interactions and conversational groups from social scenes. Below, we present a review of each of these topics.

2.1 Head and Body Pose Estimation

Head orientation serves as a useful cue to determine one's direction of attention and interest, and therefore, head pose estimation is of critical importance in studies examining non-verbal communication, social attention, surveillance and human behavior analysis (Benfold and Reid 2011; Chamveha et al. 2013; Chen and Odobez 2012; Yan et al. 2016; Robertson and Reid 2006; Heili et al. 2014). Several taxonomies have been proposed to categorize prior works in HPE, which can be found in Murphy-Chutorian and Trivedi (2009). The categorizations are mainly methodology-based, but HPE methods also differ based on their operating domain, i.e., HPE under low-resolution views acquired from far-field cameras, high resolution views from near-field cameras, involving static and moving targets, etc.

In this work, we are expressly interested in HPE from low-resolution or surveillance videos involving static and moving targets. HPE for moving targets is an interesting problem which presents both challenges and opportunities. Person tracking, head localization and HPE are inherently hard in this scenario as targets undergo appearance changes due to varying camera perspective and scale, illumination and body articulation as they move around. However, the motion direction of moving targets can be effectively used as a proxy for body or head pose in such scenes.

In Robertson and Reid (2006), efficient gaze determination is achieved by combining walking direction, obtained via target tracking and serving as a proxy for body pose, with head pose computed based on skin characteristics. In Smith et al. (2008), the authors jointly perform head tracking and HPE, and show that knowledge of head pose can

be used for improving head localization accuracy and vice-versa. This shared knowledge is exploited to obtain the visual focus of attention of pedestrians in outdoor scenes. Unsupervised HPE is proposed in Benfold and Reid (2011) by exploiting weak labels in the form of pedestrians' walking direction. A similar idea is also investigated in Chamveha et al. (2013). Chen and Odobez (2012) compute head pose by introducing two coupling factors, one between head and body pose and another between body pose and velocity direction. Differently, with explicit focus on appearance variations with scene location (or alternatively target motion), authors of Yan et al. (2013) partition the monitored scene space into a dense uniform spatial grid to learn region-specific head pose estimators along with the scene regions where facial appearance is more or less stable. Alternatively, an adaptive framework for multi-view HPE under target motion is proposed by Rajagopal et al. (2014), but this work suffers from the limitation of having to specify the number of scene partitions a priori for learning (as many) region-specific pose estimators.

When targets in the scene are mostly static, lack of motion cues makes the HPE problem harder. This is particularly the case with social scenes (e.g., a cocktail party) (Alameda-Pineda et al. 2016), where targets are not only static but also severely occluded owing to clutter. Researchers have explored novel means to perform HPE in meeting situations. For instance, in one of the first works to examine round-table meetings, Voit and Stiefelhagen (2009) use visual information from multiple views along with a neural network for single and multi-view HPE. Their algorithm is applicable to the CHIL (Butko et al. 2011) and AMI (Carletta et al. 2006) datasets where interacting targets are seated and captured by far-field cameras. In Demirkus et al. (2014), a hierarchical graphical model is proposed, and integrated with temporal smoothing to achieve improved HPE. The visual focus of attention modeled via head pose is examined for several meeting scenes in Ba and Odobez (2006). Some HPE algorithms are designed to explicitly handle occlusions (Meyer et al. 2015), while others learn person-specific appearance variations in their model (Yan et al. 2009).

Although estimation of full/articulated body pose has received much attention (Andriluka et al. 2009; Toshev and Szegedy 2014; Tompson et al. 2014), very few works have addressed body pose estimation in surveillance settings. BPE from surveillance video has been studied by few works (Robertson and Reid 2006; Krahnstoever et al. 2011; Chen et al. 2011), but they only consider body orientation as a link between walking direction and head pose, and do not explicitly estimate body pose. Recent works of Chen and Odobez (2012) and Liem and Gavrila (2014) demonstrate the benefits of jointly learning head and body pose. In Chen and Odobez (2012), this is further extended to social interactions involving 2–3 individuals. Extending this further, we expressly consider social scenes involving groups of interactors, and such scenes are characterized by little motion and considerable occlusions. Typically, prior BPE approaches do not work well in these conditions as most methods are monocular. For instance, experiments in Benfold and Reid (2011) show poor PE performance when targets are either static or their velocity is noisy. Similarly, Yan et al. (2016) alleviate the occlusion problem by considering multi-view images, but do not implement specific strategies for handling body occlusion. Very recently, joint HPE and BPE in social scenes is considered in Alameda-Pineda et al. (2015). However, additional information from non-visual sensors (i.e. microphones and infrared beam detectors) is employed as a form of weak supervision, thus simplifying the pose estimation task.

2.2 Weakly-Supervised Learning Methods for HPE and BPE

The joint learning framework we propose shares similarities with previous semi-supervised learning, transfer learning, and multi-task learning methods for head and body pose estimation. We briefly review these ideas and highlight their similarities and differences with respect to each in the proposed method.

Semi-supervised learning (see Zhu 2005 for a survey) leverages both labeled and unlabeled data to achieve improved classification with limited training data. In this work, given the difficulty in acquiring large amounts of labeled training data from the monitored social scene, we employ labeled data derived from an *auxiliary* dataset to improve pose estimation performance. Similarities in terms of appearance features between labeled and unlabeled data acquired from the social scene are then enforced via manifold regularization to jointly learn head and body pose classifiers. A similar idea is exploited in Chen and Odobez (2012) and Alameda-Pineda et al. (2015), where semi-supervised learning frameworks for HPE and BPE are proposed. Similar to Chen and Odobez (2012), we also adopt a semi-supervised approach to estimate head and body pose. However, our work is unique on multiple aspects. Firstly, Chen and Odobez (2012) mainly focus on head and body pose estimation, while our method focuses on simultaneous estimation of head and body pose as well as F-Formations in social interactions, which is challenging to analyze as targets are mostly static and severely occluded. Secondly, F-formations are estimated by adding a further term in the loss function which simultaneously regularizes BPE. Finally, we include temporal smoothing and occlusion-specific classifiers that improve pose and FF estimation performance.

In transfer learning, a classifier is initially built upon learning from a *source* dataset, and later adapted to a *target* dataset having a different feature distribution. A few transfer learning approaches have been proposed recently for HPE.

In Rajagopal et al. (2014), transfer learning techniques are used to address variations between *source* and *target* datasets owing to the differing range of head orientations and person motion in the *target*. Similarly, in Heili et al. (2014) a domain adaptation technique is used to 'align' the underlying structure of the *source* and *target* datasets. In our framework, we attempt to enforce consistency between head and body samples extracted from the auxiliary dataset and the social scene under analysis, which have different attributes—this can be regarded as knowledge transfer. As detailed in Sect. 4, our algorithm performs poorly when only the auxiliary data samples are employed for learning. However, performance improves considerably when samples from the scene are additionally utilized, facilitating the model to *adapt* to the target scene.

In multi-task learning, several classifiers are learned for related tasks so that their commonalities as well as differences are modeled by the classifiers (e.g., learning pose classifiers for different views). This procedure requires that the input and output spaces for all tasks remain the same. For instance, Yan et al. (2016) address the problem of HPE under target motion by modeling facial similarities and differences among neighboring scene regions using multi-task learning. Our learning framework connects the body pose of interactors with FFs, which are highly functionally related, but correspond to different input spaces.

2.3 Social Interactions and Conversational Groups

There has been considerable interest in analyzing social interactions and social scenes of late. Marin-Jimenez et al. (2014) propose continuous HPE using Gaussian process regression, and evaluate several methodologies for detecting dyadic interactions in a video shot. Ba and Odobez (2008) propose the joint estimation of visual focus of attention (VFOA) by the use of interaction models and contextual cues. Here, they exploit the fact that people in a group interaction naturally tend to share VFOA targets. Patron-Perez et al. (2012) achieve spatio-temporal localization of dyadic interactions from TV videos using a structured support vector machine, combining information from pose-based and position-based descriptors. Choi et al. (2014) recognize group activities by analyzing the spatial arrangement of group members. Other works have focused on (a) detecting groups instead of individuals in static images to overcome partial occlusions (Eichner and Ferrari 2010; Tang et al. 2014) and (b) leveraging information concerning groups to improve multi-target tracking performance (Pellegrini et al. 2010; Leal-Taixé et al. 2014).

Detecting conversational groups or F-formations in social scenes has generated interest lately due to security, behavioral and commercial applications (Cristani et al. 2011; Setti et al. 2013; Gan et al. 2013). Cristani et al. (2011) analyze spatial arrangements and head orientations, and propose a voting strategy based on the Hough transform to detect F-formations. This work is extended via multi-scale analysis in Setti et al. (2013). Social interactions are detected from ego-centric videos via a correlation clustering algorithm in Alletto et al. (2014). Vascon et al. (2014) detect FFs by applying a game-theoretic clustering algorithm to an affinity matrix. This is achieved by first creating a frustum (a two dimensional histogram capturing social attention) for each individual using their head orientation and feet position. A game-theoretic clustering algorithm is then employed over the affinity matrix, which is calculated using frustum overlaps, to detect FF. A more comprehensive study of this approach is presented in Vascon et al. (2016). Although both Vascon et al. (2014), Vascon et al. (2016) propose an elegant approach to deal with noise and systematically integrate temporal information to obtain superior results, availability of head pose using either manual or semi-automatic method is assumed. Our approach differs significantly from Vascon et al. (2014), Vascon et al. (2016) as we estimate both head and body pose and detect FFs simultaneously and without assuming the availability of head pose orientations.

Recently, the effectiveness of the graph-cut approach for individuating FFs is demonstrated in Setti et al. (2015). While orientation relationships among interactors have been exploited for detecting conversational groups, joint estimation of FFs, and the head, body pose of the FF members has never been attempted. In this paper, we show that joint learning benefits both pose and FF estimation.

The following section describes the various aspects of our pose and FF estimation framework in detail.

3 Framework for Analyzing Social Scenes

3.1 Overview

In this section, we describe our approach to jointly infer conversational groups and the head and body pose of each target in a social scene. An overview of our social scene analysis pipeline is presented in Fig. 2. Given a distant video of a social gathering (e.g., cocktail party), we first apply multi-target tracking to estimate the feet positions of persons in the scene. Target feet positions are estimated with the Hybrid Joint Separable-Particle Filter (HJS-PF) multi-target tracking approach (see supplemetary material for more elaborate description of how HJS-PF is deployed for our purpose) described in Hu et al. (2015), and are used for head localization and cropping via a 3D head-plus-shoulder model registered through shape matching as in Yan et al. (2016). Each target's body region is determined as the portion between head and feet coordinates.

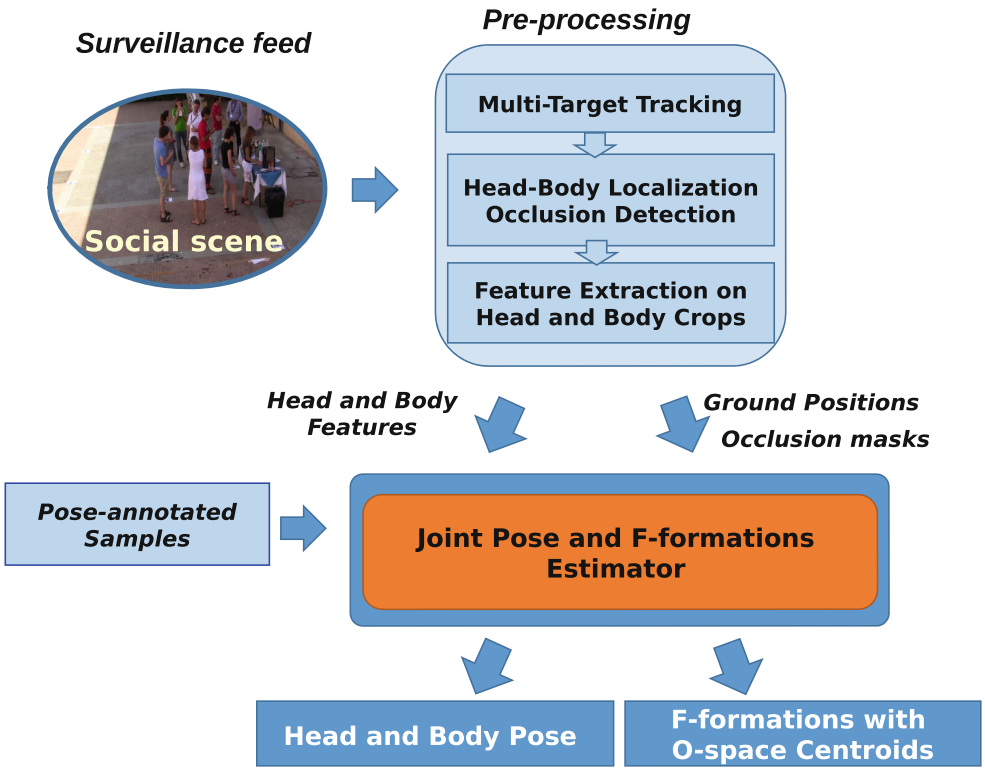

Fig. 2 Overview of our social scene analysis framework

We also estimate the extent of occlusion for each target by accounting for shape-projections of targets closer to the camera. In practice, we associate a binary occlusion mask to each of the computed body crops. Camera calibration information is used for tracking, head/body localization, as well as for occlusion detection. We then extract visual descriptors (HOG descriptors as detailed in Sect. 3.4) for the head and body regions. Targets' positions, head and body features along with occlusion masks are input to our joint learning algorithm that outputs (i) head and body pose and (ii) F-formation membership for each target as described below.

3.2 Problem Setting

We consider a N_T-frame video depicting N_K persons involved in a social gathering. Each target k is characterized by a time-dependent triplet $(\boldsymbol{x}^{\mathrm{B}}_{kt}, \boldsymbol{x}^{\mathrm{H}}_{kt}, \boldsymbol{p}_{kt})$, providing for each frame t the body and head descriptors denoted by $\boldsymbol{x}^{\mathrm{B}}_{kt} \in \mathcal{X}_{\mathrm{B}}$ and $\boldsymbol{x}^{\mathrm{H}}_{kt} \in \mathcal{X}_h$ respectively, and the target's feet position $\boldsymbol{p}_{kt} \in \mathbb{R}^2$. Here, $\mathcal{X}_{\mathrm{B}}$ and $\mathcal{X}_{\mathrm{H}}$ represent the feature spaces associated to body and head samples respectively. We use HOG descriptors to characterize the head and body patches (see Sect. 3.4 for more details). The target's feet positions are obtained using tracking and registration. Information regarding all video targets are collected in a set $\mathcal{S} = \{(\boldsymbol{x}^{\mathrm{B}}_{kt}, \boldsymbol{x}^{\mathrm{H}}_{kt}, \boldsymbol{p}_{kt}) : k \in \langle N_K \rangle, t \in \langle N_T \rangle\}_{kt}$, where $\langle N \rangle = \{1, \ldots, N\}$ for notational convenience.

The goal of the inference task is to estimate the body pose $\alpha^{\mathrm{B}}_{kt} \in [0, 2\pi)$, the head pose $\alpha^{\mathrm{H}}_{kt} \in [0, 2\pi)$ and the conversational group membership $z_{kt} \in \langle N_K \rangle$ of each target k at each frame t. As in previous works considering a low resolution setting (Yan et al. 2016; Rajagopal et al. 2014), we estimate only the head and body *pan*.[2] Analogous to clustering methods, F-formations are determined by all targets sharing the membership z_{kt} (i.e., two targets k and h belong to the same group at frame t if $z_{kt} = z_{ht}$). Singleton conversational groups represent non-interacting targets.

In addition to the social scene information provided by $\mathcal{S}$, we exploit annotated training sets $\mathcal{T}_{\mathrm{B}} = \{(\hat{\boldsymbol{x}}^{\mathrm{B}}_i, \boldsymbol{y}^{\mathrm{B}}_i)\}_{i=1}^{N_{\mathrm{B}}} \subseteq \mathcal{X}_{\mathrm{B}} \times \mathcal{Y}$ and $\mathcal{T}_{\mathrm{H}} = \{(\hat{\boldsymbol{x}}^{\mathrm{H}}_i, \boldsymbol{y}^{\mathrm{H}}_i)\}_{i=1}^{N_{\mathrm{H}}} \subseteq \mathcal{X}_{\mathrm{H}} \times \mathcal{Y}$ to enhance the head and body pose estimation capabilities of our model. Each training sample in $\mathcal{T}_\diamond$, where $\diamond \in \{\mathrm{B}, \mathrm{H}\}$, is a descriptor $\hat{\boldsymbol{x}}^{\diamond}_i$ for head/body with an associated pose label $\boldsymbol{y}^{\diamond}_i$. The pose labels are N_C-dimensional binary vectors[3] with a single non-zero entry indexing an angle in $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{N_C}]^\top \in [0, 2\pi)^{N_C}$ (i.e., $\mathcal{Y} \in \{0, 1\}^{N_C}$, where N_C denotes the number of quantized angles).

[2] The head and body angles are orientations in the ground plane.

[3] Most available datasets on head and body pose estimation in low resolution settings only provide quantized pose annotations.

For convenience, we also define a re-parametrization of $\boldsymbol{\alpha}$ in terms of a matrix of 2-dimensional vectors:

$$\boldsymbol{A} = \begin{bmatrix} \cos\alpha_1 & \cdots & \cos\alpha_{N_C} \\ \sin\alpha_1 & \cdots & \sin\alpha_{N_C} \end{bmatrix}. \tag{1}$$

In the following, symbol $\diamond$ is used as a placeholder for H or B.

3.3 Jointly Inferring Pose and F-Formations

The inference problem that we face is semi-supervised, as we have both annotated data from $\mathcal{T} = (\mathcal{T}_{\mathrm{B}}, \mathcal{T}_{\mathrm{H}})$ and *non-annotated* observations $\mathcal{S}$ from the social scene under analysis. The head and body pose annotations from $\mathcal{T}$ implicitly provide a prior for estimating the pose of targets in $\mathcal{S}$. F-formation annotations are not used during learning, and are therefore discovered in an unsupervised manner.

In order to exploit the distribution of descriptors corresponding to annotated data and scene targets, we introduce two regression functions f_{B} and f_{H} for the body and head pose respectively, which are two unknowns in our model. Intuitively, $f_\diamond : \mathcal{X}_\diamond \to \mathbb{R}^{N_C}$ provides for each sample in $\mathcal{X}_\diamond$, a prediction for the pose label in $\mathcal{Y}$ that is relaxed to a real vector in $\mathbb{R}^{N_C}$. The output of $f_\diamond$ can be used to linearly combine the columns of $\boldsymbol{A}$ in (1), which are a vectorial representation of the discretized angles in $\boldsymbol{\alpha}$. The resulting 2-dimensional vector $\boldsymbol{A}\, f_\diamond(\boldsymbol{x}_{kt}^\diamond) \in \mathbb{R}^2$ can finally be cast in polar coordinates to recover the pose angles $\alpha_{kt}^\diamond$ corresponding to $\boldsymbol{x}_{kt}^\diamond$ in $\mathcal{S}$.

Assignment of targets to FFs is modeled indirectly by letting each target vote for the center of the FF he/she belongs to. In practice, we introduce a *latent* 2-dimensional vector $\boldsymbol{c}_{kt}$ for each target $k \in \langle N_K \rangle$ and frame $t \in \langle N_T \rangle$, which intuitively represents the voted FF center for target k in frame t. We assume these centers, which will be additional unknowns of our model, to be stacked into a $2 \times N_K N_T$-dimensional matrix $\boldsymbol{C}$. We denote by $\mathcal{C} = \mathbb{R}^{2\times N_K N_T}$, the set of all such matrices. Given $\boldsymbol{C}$, the corresponding F-formation assignments z_{kt} can be easily recovered as shown in Hocking et al. (2011). Intuitively, two targets k and h are considered members of the same group, i.e., $z_{kt} = z_{ht}$, if their voted centers $\boldsymbol{c}_{kt}$ and $\boldsymbol{c}_{ht}$ for the O-space center are close enough. We use a very small value (1e-6) to merge two clusters and we verified empirically that values from 1e-3 to 1e-6 are suitable to our purpose.

Our goal is to jointly infer the head and body poses and F-formations, i.e., to find pose regressors and center votes that minimize the following loss, given $\mathcal{T}$ and $\mathcal{S}$:

$$\begin{aligned} \min \quad & L_P(f_{\mathrm{B}}, f_{\mathrm{H}}; \mathcal{T}, \mathcal{S}) + L_F(f_{\mathrm{B}}, \boldsymbol{C}; \mathcal{S}) \\ \text{s.t.} \quad & f_{\mathrm{B}} \in \mathcal{F}_{\mathrm{B}},\ f_{\mathrm{H}} \in \mathcal{F}_{\mathrm{H}},\ \boldsymbol{C} \in \mathcal{C} \end{aligned} \tag{2}$$

where $\mathcal{F}_\diamond$ is the space of pose regressors $f_\diamond$ (details on pose regressors spaces are given in Sect. 3.4). The loss in (2) has two terms. The first term, L_P, enforces pose regressors to reflect the distribution of annotated samples in $\mathcal{T}$ under a regularization that also accounts for the manifold of unlabeled samples in $\mathcal{S}$. The second term, L_F, enforces the body pose estimates of the targets in $\mathcal{S}$ to be consistent with the FF center votes given by $\boldsymbol{C}$. Given the optimal solution to (2), we recover the head/body pose $\alpha_{kt}^\diamond$ and FF assignment z_{kt} of each target at every frame as discussed above.

We now describe L_P and L_F in detail.

3.3.1 Pose-Related Loss Term

The pose-related loss term L_P decomposes into three terms:

$$L_P(f_{\mathrm{B}}, f_{\mathrm{H}}; \mathcal{T}, \mathcal{S}) = \sum_{\diamond\in\{\mathrm{H},\mathrm{B}\}} L_\diamond(f_\diamond; \mathcal{T}_\diamond, \mathcal{S}) + L_C(f_{\mathrm{B}}, f_{\mathrm{H}}; \mathcal{S}). \tag{3}$$

The first two loss terms L_{H} and L_{B} penalize pose regressors errors with respect to the annotated training sets under harmonic regularization, also accounting for the data manifold of $\mathcal{S}$. To this end, we introduce two graph-based manifolds $\mathfrak{G}_{\mathrm{H}}$ and $\mathfrak{G}_{\mathrm{B}}$ for the available head and body samples. For each $\diamond \in \{\mathrm{H}, \mathrm{B}\}$, the graph is defined as $\mathfrak{G}_\diamond = (\mathcal{V}_\diamond, \mathcal{E}_\diamond, \omega^\diamond)$, where $\mathcal{V}_\diamond$ comprises all body/head samples (depending on $\diamond$) from $\mathcal{T}_\diamond$ and $\mathcal{S}$, the first $N_\diamond$ being samples from $\mathcal{T}_\diamond$ and the rest from $\mathcal{S}$. In total, $\mathcal{V}_\diamond$ contains $N_\diamond + N_K N_T$ elements, the ith one denoted by $\boldsymbol{v}_i^\diamond \in \mathcal{X}_\diamond$. For all annotated samples in $\mathcal{V}_\diamond$, i.e., $\forall i \in \langle N_\diamond \rangle$, we indicate the corresponding pose label by $\boldsymbol{y}_i^\diamond$. The set $\mathcal{E}_\diamond \subseteq \langle |\mathcal{V}_\diamond| \rangle^2$ indexes pairs of neighboring vertices, while $\omega_{ij}^\diamond \geq 0$ is a non-negative weight indicating the strength of the (i, j)-edge connection. More details will be given in Sect. 3.4.

Given $\mathfrak{G}_{\mathrm{H}}$ and $\mathfrak{G}_{\mathrm{B}}$, we define the loss term $L_\diamond$ as

$$\begin{aligned} L_\diamond(f; \mathcal{T}_\diamond, \mathcal{S}) = {} & \frac{1}{N_\diamond} \sum_{i=1}^{N_\diamond} \| f(\boldsymbol{v}_i^\diamond) - \boldsymbol{y}_i^\diamond \|_{\boldsymbol{M}}^2 + \lambda_R \| f \|_{\mathcal{F}_\diamond}^2 \\ & + \lambda_U \frac{1}{|\omega^\diamond|_0} \sum_{(i,j)\in\mathcal{E}_\diamond} \omega_{ij}^\diamond \| f(\boldsymbol{v}_i^\diamond) - f(\boldsymbol{v}_j^\diamond) \|_{\boldsymbol{M}}^2, \end{aligned} \tag{4}$$

where $\| \cdot \|_{\mathcal{F}_\diamond}$ is a semi-norm for the function space $\mathcal{F}_\diamond$, $|\omega|_0$ is the number of edges with non-zero weights and $\|\boldsymbol{a}\|_{\boldsymbol{M}} = \sqrt{\boldsymbol{a}^\top \boldsymbol{M} \boldsymbol{a}}$ is a semi-norm on $\mathbb{R}^{N_C}$ induced by the symmetric, positive semi-definite matrix $\boldsymbol{M} \in \mathbb{R}^{N_C \times N_C}$, which accounts for the semantic mapping from the pose label vectors $\in \mathbb{R}^{N_C}$ to angles in $\boldsymbol{\alpha}$ (see, Sect. 3.4).

The first term in $L_\diamond$ measures the prediction error of $f \in \mathcal{F}_\diamond$ with respect to the annotated training set; the sec-

ond term regularizes f in the respective function space; the last term performs harmonic regularization of f with respect to the manifold of data samples in $\mathcal{T}_\diamond$ and $\mathcal{S}$. Finally, we have two free non-negative parameters λ_R and λ_U to balance the contribution of the regularization terms. Note that losses akin to (4) are typically encountered in the context of semi-supervised learning (Zhu and Goldberg 2009).

The last term in (3) enforces consistency between head and body poses predicted on $\mathcal{S}$ by penalizing configurations violating human anatomic constraints (e.g., head and body oriented in opposite directions):

$$L_C(f_{\mathrm{B}}, f_{\mathrm{H}}; \mathcal{S}) = \lambda_C \frac{1}{N_K N_T} \sum_{k=1}^{N_K} \sum_{t=1}^{N_T} \| f_{\mathrm{B}}(\boldsymbol{x}_{kt}^{\mathrm{B}}) - f_{\mathrm{H}}(\boldsymbol{x}_{kt}^{\mathrm{H}}) \|_{\boldsymbol{M}}^2, \tag{5}$$

where λ_C is a free, non-negative parameter. The pose related loss terms in (3) is similar to the objective function used in Chen and Odobez (2012). However, in Chen and Odobez (2012) the F-formation-related loss term described in the following subsection is not considered.

3.3.2 The F-Formation-Related Loss Term

The second term of the objective function in (2) is specifically defined to exploit the relationship between targets' body orientation and F-formations. Our purpose is to exploit the targets' group membership for refining body pose estimates as group members tend to orient towards the O-space center and, conversely, to accurately detect FFs from body pose estimates of interacting targets.

The following loss term depends on a body regressor $f_{\mathrm{B}} \in \mathcal{F}_{\mathrm{B}}$, and on a matrix of votes $\boldsymbol{C} \in \mathcal{C}$ concerning F-formation center for each target and at each frame:

$$\begin{aligned} &L_F(f_{\mathrm{B}}, \boldsymbol{C}; \mathcal{S}) \\ &= \frac{1}{N_K N_T} \Bigg[\lambda_F \sum_{k=1}^{N_K} \sum_{t=1}^{N_T} \| \boldsymbol{c}_{kt} - (\boldsymbol{p}_{kt} + D\boldsymbol{A}\, f_{\mathrm{B}}(\boldsymbol{x}_{kt}^{\mathrm{B}})) \|_2^2 \\ &\quad + \gamma_c \sum_{k,h=1}^{N_K} \sum_{t=1}^{N_T} \| \boldsymbol{c}_{kt} - \boldsymbol{c}_{ht} \|_1 + \lambda_T \sum_{k=1}^{N_K} \sum_{t=2}^{N_T} \| \boldsymbol{c}_{kt} - \boldsymbol{c}_{k(t-1)} \|_1 \Bigg], \end{aligned} \tag{6}$$

where $\| \cdot \|_p$ is the p-norm, and λ_F, D, γ_c and λ_T are non-negative, free parameters.

Since interactors typically orient their bodies towards the O-space center, we expect the center vote of each target at each frame to be located D units from the target in the direction predicted by the body pose regressor, where D denotes the expected target distance from a hypothetical O-space center (akin to previous works Cristani et al. 2011; Vascon et al. 2014). The body orientation for the kth target at frame t in $\mathbb{R}^2$ is obtained as $\boldsymbol{A}\, f_{\mathrm{B}}(\boldsymbol{x}_{kt}^{\mathrm{B}})$, since the output of f is the predicted pose label. Hence, the ideal FF center position $\boldsymbol{c}_{kt}$ for each target is given by $\boldsymbol{p}_{kt} + \boldsymbol{d}_{kt}$, where $\boldsymbol{d}_{kt} = D\boldsymbol{A}\, f_{\mathrm{B}}(\boldsymbol{x}_{kt}^{\mathrm{B}})$. This is accounted by the first term in (6). The second term induces a spatial clustering of the center votes of all targets at each frame, which is regulated by the parameter γ_c: large values of γ_c tend to favor the concentration of the votes into few cluster points, while low values reduce the mutual influence of the targets' votes. Computed cluster centroids represent putative O-space centers of FFs in the scene. Note that the 1-norm induces the centroids of targets belonging to the same FF to merge. This effect may not be achieved with other norms such as the L_2-norm. Finally, the third term enforces temporal consistency of the targets' center votes, given the fact that FFs do not change rapidly over time. It is worth nothing that in (6), no information regarding group membership (z_{kt}) is employed to force clustering votes of only associated targets. The general formulation employed may therefore result in a biased cluster centroid estimate pointing to the scene center. However, this biasing rarely happens in practice when the value of γ_c is properly chosen, and does not hinder accurate recognition of conversational groups.

In contrast to other prior works which use head orientations to infer FFs, we propose a coupled inference framework. The loss term L_F allows for coupled estimation of body pose and O-space centroids via the center votes of targets (Fig. 3). Indeed, we exploit the fact that body pose is a more stable cue than head pose for inferring FFs, and this reflects via improved FF and body pose estimation accuracy as discussed in Sect. 4.

3.4 Implementation Details

We model each regressor $f_\diamond$ as a generalized, linear function parametrized by a matrix $\boldsymbol{\Theta} \in \mathbb{R}^{N_C \times M_\diamond}$, i.e.

$$f_\diamond(\boldsymbol{x}; \boldsymbol{\Theta}) = \boldsymbol{\Theta} \Phi_\diamond(\boldsymbol{x}), \tag{7}$$

where $\Phi_\diamond : \mathcal{X}_\diamond \to \mathbb{R}^{M_\diamond}$ is a feature mapping. The set of all regressors $f^\diamond$ is thus given by

$$\mathcal{F}_\diamond = \{ f_\diamond(\boldsymbol{x}; \boldsymbol{\Theta}) \,:\, \boldsymbol{\Theta} \in \mathbb{R}^{N_C \times M_\diamond} \}, \tag{8}$$

In light of the surjection between parameters $\boldsymbol{\Theta} \in \mathbb{R}^{N_C \times M_\diamond}$ and regressors $f_\diamond \in \mathcal{F}_\diamond$, we can re-write the minimization in (2) with variables $\boldsymbol{\Theta}_{\mathrm{B}} \in \mathbb{R}^{N_C \times M_{\mathrm{B}}}$ and $\boldsymbol{\Theta}_{\mathrm{H}} \in \mathbb{R}^{N_C \times M_{\mathrm{H}}}$, by substituting $f_\diamond$ with its definition in (7) and by taking the following seminorm on the space $\mathcal{F}_\diamond$: $\| f(\cdot; \boldsymbol{\Theta}) \|_{\mathcal{F}_\diamond} = \| \boldsymbol{\Theta} \|_F$, where $\| \cdot \|_F$ denotes the Frobenious norm. Note that the feature mapping $\Phi_\diamond$ can be specified by implicitly defining a kernel function, as in kernel methods. We consider a linear kernel in our experiments.

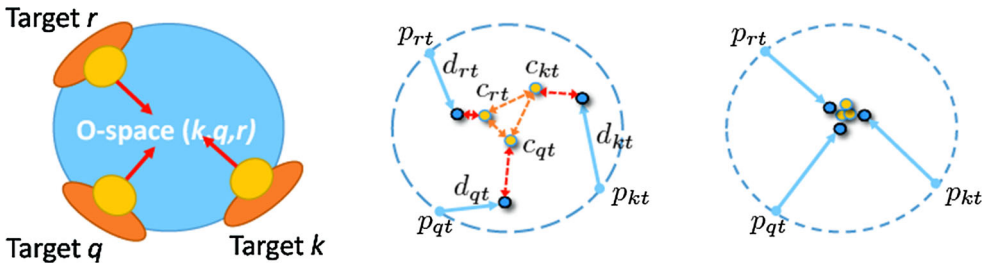

Fig. 3 (*Left*) O-space of the F-formation involving three targets k, r and q and their body pose. (*Center*) direction vectors $\boldsymbol{d}_{(.)}$ obtained via body pose regressor are shown using *blue arrows*, while $\boldsymbol{c}_{(.)}$ (*yellow points*) denote voted center locations. By minimizing (2), we refine body pose and F-formation estimates to arrive at the least loss configuration (*right*), where the voted centers for each target cluster at the O-space centroid. For sake of simplicity, we illustrate the minimization of (2) for a single frame t and for $\lambda_T = 0$ (Color figure online)

To facilitate comparisons with previous works (Chen and Odobez 2012; Rajagopal et al. 2014), we consider HOG features to describe the head and body regions. Head crops are first normalized to 20×20 pixels and HOG features are computed over 4×4 cells. Similarly, body images are resized to 80×60 pixels, and HOG features are extracted over 4×4 cells. Similar to previous works (Chen and Odobez 2012; Rajagopal et al. 2014) and consistent with annotations for most datasets in a low-resolution setting, we set $N_C = 8$.

The graph-based data manifolds $\mathfrak{G}_\diamond = (\mathcal{V}_\diamond, \mathcal{E}_\diamond, \omega^\diamond)$, used in (4) for harmonic regularization of pose regressors, are defined such that head/body samples similar in appearance should correspond to similar pose. Specifically, $(i, j) \in \mathcal{E}_\diamond$ if the ith sample $\boldsymbol{v}_i^\diamond \in \mathcal{V}_\diamond$ is among the k-nearest neighbors of the jth sample $\boldsymbol{v}_j^\diamond \in \mathcal{V}_\diamond$ under the standard Euclidean metric. Moreover, temporal smoothing is enforced by imposing that $(i, j) \in \mathcal{E}_\diamond$ if samples $\boldsymbol{v}_i^\diamond$ and $\boldsymbol{v}_j^\diamond$ correspond to samples $\boldsymbol{x}_{kt}^\diamond$ and $\boldsymbol{x}_{kt'}^\diamond$ in $\mathcal{S}$, where $|t - t'| = 1$, i.e., they correspond to the same target in contiguous frames. Also, we do not impose any preference over edges and set a constant strength equal to one, i.e., $\omega_{ij}^\diamond = 1$. The metric matrix $\boldsymbol{M}$ adopted in (4) and in (5) is defined as $\boldsymbol{M} = \boldsymbol{A}^\top \boldsymbol{A}$, and the parameters λ_R, λ_U, λ_C, λ_F, λ_T and γ_c are fixed using a validation set. Details are provided in Sect. 4.

3.5 Optimization

By taking (8) as the regressors' space and by rewriting the minimization in (2) in terms of $\boldsymbol{\Theta}_\diamond$ as mentioned in Sect. 3.4, we obtain a *convex* optimization problem with variables $(\boldsymbol{\Theta}_\mathrm{B}, \boldsymbol{\Theta}_\mathrm{H}, \boldsymbol{C})$, which can be reformulated as a Quadratic Program (QP). The convexity is implied by the fact that we have a sum of positively-rescaled terms being the composition of a norm (or semi-norm) with an affine function of the variables to be optimized. Accordingly, any local solver can be used to find a global solution, irrespective of the initial starting point.

The optimization strategy we propose involves alternating updates of $\boldsymbol{\Theta}_\mathrm{B}$, $\boldsymbol{\Theta}_\mathrm{H}$ and $\boldsymbol{C}$. Before delving into details, we introduce the following matrices: $\boldsymbol{X}_\diamond = (\boldsymbol{x}_{11}^\diamond, \ldots, \boldsymbol{x}_{N_K N_T}^\diamond)$, $\hat{\boldsymbol{X}}_\diamond = (\hat{\boldsymbol{x}}_1^\diamond, \ldots, \hat{\boldsymbol{x}}_{N_\diamond}^\diamond)$, $\boldsymbol{Y}_\diamond = (\boldsymbol{y}_1^\diamond, \ldots, \boldsymbol{y}_{N_\diamond}^\diamond)$ and $\boldsymbol{V}_\diamond = (\hat{\boldsymbol{X}}_\diamond, \boldsymbol{X}_\diamond)$. Moreover, let $\boldsymbol{L}_\diamond$ denote the Laplacian matrix of the graph $\mathfrak{G}_\diamond$ defined in Sect. 3.3, and let:

$$\boldsymbol{E}_\diamond = \lambda_R \boldsymbol{I} + (\hat{\boldsymbol{X}}_\diamond \hat{\boldsymbol{X}}_\diamond^\top + \lambda_U \boldsymbol{V}_\diamond \boldsymbol{L}_\diamond \boldsymbol{V}_\diamond^\top + \lambda_C \boldsymbol{X}_\diamond \boldsymbol{X}_\diamond^\top) \otimes \boldsymbol{M},$$
$$\boldsymbol{F}_\diamond = \boldsymbol{M} \left[\boldsymbol{Y}_\diamond \hat{\boldsymbol{X}}_\diamond^\top + \lambda_C \boldsymbol{\Theta}_\star \boldsymbol{X}_\star \boldsymbol{X}_\diamond^\top \right],$$

where $(\diamond, \star) \in \{(\mathrm{H}, \mathrm{B}), (\mathrm{B}, \mathrm{H})\}$, $\otimes$ is the Kronecker product and $\boldsymbol{I}$ is a properly-sized identity matrix. In the following, we briefly describe our iterative optimization framework. Detailed derivations can be found in the "Appendix".

Update of $\boldsymbol{\Theta}_\mathrm{H}$ The optimization problem in (2) is quadratic and unconstrained in $\boldsymbol{\Theta}_\mathrm{H}$. Accordingly, the update rule that we find by setting the first-order derivatives to zero has the following closed-form:

$$\mathrm{vec}(\boldsymbol{\Theta}_\mathrm{H}) \leftarrow \boldsymbol{E}_\mathrm{H}^{-1} \mathrm{vec}(\boldsymbol{F}_\mathrm{H}),$$

where $\mathrm{vec}(\cdot)$ denotes vectorization of a matrix.

Update of $\boldsymbol{\Theta}_\mathrm{B}$ Similarly, the update for $\boldsymbol{\Theta}_\mathrm{B}$ is given by

$$\mathrm{vec}(\boldsymbol{\Theta}_\mathrm{B}) \leftarrow \left[\boldsymbol{E}_\mathrm{B} + \lambda_F D^2 \boldsymbol{X}_\mathrm{B} \boldsymbol{X}_\mathrm{B}^\top \otimes \boldsymbol{A}^\top \boldsymbol{A} \right]^{-1} \mathrm{vec}(\boldsymbol{G}),$$

where $\boldsymbol{G} = \boldsymbol{F}_\mathrm{B} + \lambda_F D \boldsymbol{A}^\top (\boldsymbol{C} - \boldsymbol{P}) \boldsymbol{X}_\mathrm{B}^\top$.

Update of $\boldsymbol{C}$ Computing a minimizer of (2) with respect to $\boldsymbol{C}$ (with $\boldsymbol{\Theta}_\diamond$ fixed) is equivalent to finding a minimizer of L_F with respect to $\boldsymbol{C}$, as L_P does not depend on $\boldsymbol{C}$. The resulting optimization problem can be solved efficiently with the alternating direction method of multipliers (Chi and Lange 2015).

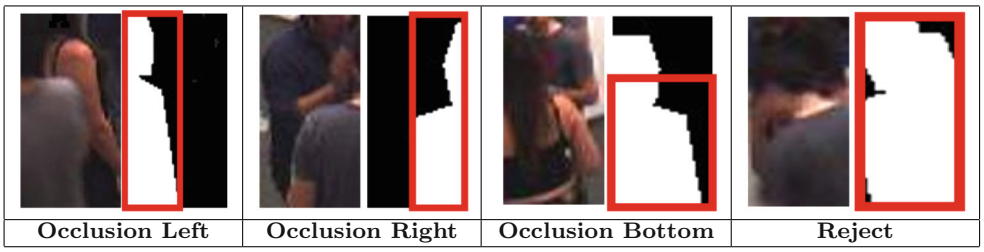

Fig. 4 The different occlusion categories illustrated with samples from the Cocktail Party dataset. *White pixels* show the occlusion mask generated by a tracked foreground target. *Red rectangles* show the regions associated to the assigned category (for left/right category we consider the left-/right-half of the crop, while for the bottom category we use half of the torso area which covers the bottom-most 70% of the crop area) (Color figure online)

3.6 Handling Occlusions

We now show how the proposed framework can be extended to integrate information about body occlusions. In previous approaches (Mathias et al. 2013; Wojek et al. 2011), a convex combination of the occlusion-specific classifier scores is considered at test time. Following a similar idea, we partition the body samples extracted from the social scene into four groups according to the detected level of occlusion, namely (i) fully visible and occlusion (ii) to the left, (iii) to the right, (iv) at the bottom. To determine the occlusion category of a sample we first generate an occlusion mask using a coarse shape model (see supplementary material) and considering the target position obtained by visual tracking and camera calibration information. A region is considered occluded if at least 50% of its pixels are masked. The sample is assigned to the category group with the highest occlusion level. If the occlusion level is above 80% of the entire mask the sample is rejected. Figure 4 shows the regions used to determine the level of occlusion and illustrates the different occlusion categories with real samples. Based on the detected level of occlusion, we propose to learn multiple occlusion-specific regression functions for body pose estimation and invoke the appropriate model in Eq. (6). In this work, we consider $O = 4$ different pose regressors f^o_{B}, $o = 1, \ldots, O$, one for each group.

Similarly, we generate four sets of virtual samples from the auxiliary training dataset, creating artificial occlusions. In this way, solving (2) with the proposed iterative approach (see, Sect. 3.5) reduces to solving a set of O independent optimization problems while learning f^o_{B} and f_{H}. Conversely, while learning $\boldsymbol{C}$, the appropriate occlusion-specific regressor f^o_{B} is invoked for each sample $\mathbf{x}^{\mathrm{B}}_{k,t}$, according to its occlusion level. While our approach can be also used to model head occlusions, we consider only body occlusions to (i) limit computational costs, and (ii) address the fact that body pose estimation is more severely impeded by ambient occlusions as compared to head pose.

4 Experimental Results

In this section, we demonstrate the effectiveness of the proposed approach and present the results of our experimental evaluation, conducted on three publicly available social datasets.

4.1 Datasets

We considered three benchmark datasets, namely the Cocktail Party (CP) (Zen et al. 2010), Coffee Break (CB) (Cristani et al. 2011) and the recently proposed SALSA (Alameda-Pineda et al. 2016) for our evaluation. Figure 5 shows one sample frame extracted from each of these datasets and Table 1 gives details about the datasets used in this paper. Brief descriptions of these three datasets follow.

The **Cocktail Party** dataset (Zen et al. 2010) contains a 30-min video recording of a cocktail party in a 30 m^2 room involving six subjects. The social event is recorded using four synchronized wall-mounted cameras at 15 Hz (512 × 384 pixels, jpeg). Target positions are logged via a multi-camera tracker, while head and body orientations are manually assigned to one of $N_C = 8$ class labels denoting a quantized 45° head/body pan, for those frames where FF annotations are available. However, consistent with prior works (Cristani et al. 2011; Vascon et al. 2014), we only used video data from a single camera for pose and FF estimation. FF annotations are available every five seconds resulting in a total of 320 frames for evaluation.

The **Coffee Break** dataset (Cristani et al. 2011) depicts a social event and comprises a maximum of 14 targets, organized in groups of 2-3 persons. Target positions are given and have been obtained by processing the sequences with a visual tracking algorithm. Moreover, the dataset provides annotation only for head poses quantized into four orientations with respect to the *image* plane. We enriched the ground-truth by annotating both the head and body samples using a fine-grained quantization of orientations with respect to the *ground* plane. The dataset consists of two sequences

Fig. 5 Sample frames extracted from the considered datasets: (*left*) Cocktail Party, (*center*) Coffee Break and (*right*) SALSA

Table 1 Datasets used in this paper

Dataset	Length (min)	Resolution	#Annotated frames	Head, body pose	FF	#Targets
DPOSE	–	1024 × 768	50,000	Yes	No	16
Cocktail party (CP)	30	512 × 384	320	Yes	Yes	6
Coffee break (CB)	–	1440 × 1080	120	Yes	Yes	14
SALSA	60	512 × 384	1200	Yes	Yes	18

with annotations for 45 and 75 frames in the two sequences. The actual duration of the sequences is not known. The FF annotations are possibly interleaved.

SALSA (Alameda-Pineda et al. 2016) is a recent and challenging dataset comprising of two recordings of a social event (30 min each) involving 18 participants. The first of the two videos covers targets interacting during a poster presentation session, while the second depicts a cocktail party, where targets freely move around a table with food and beverages. Both sequences are extremely challenging for visual analysis due to low-resolution of the targets' faces, background clutter and persistent occlusions (SALSA also contains sensor data logged from sociometric badges worn by targets, which is not utilized in this work). The social scene in SALSA is visually captured by four synchronized cameras at (1024 × 768 resolution) operating at 15Hz. Targets are tracked using a multi-camera tracker. Also, manual annotations are available every three seconds, indicating ground position, head and body orientation of each target as well as FFs, thus leading to 1200 frames for evaluation.

FF annotations are stated as a partition of the target identities set. Target Ids within the same partition belong to an F-formation. For instance, annotation given as $\{\{1, 2\}, \{3, 4\}, \{5\}\}$ indicates that target ids 1 and 2 belong to an FF, while 3 and 4 belong to another FF and 5 is a singleton. Pre-processing steps are applied to each dataset as described in Sect. 3.1 (camera calibration, required by our method, is provided with each dataset). For sake of completeness, we also report the histograms of ground truth group cardinalities for different datasets in Fig. 6. The plots demonstrate that SALSA dataset is richer with a wide range of group cardinalities.

Coffee Break dataset comes with its own tracking annotations and we use them for fair comparison with previously-published results. For Cocktail Party and SALSA datasets, we exploit the multi-view recordings available for tracking with our implementation of HJS-PF multi-camera person tracking.[4] Also, ground positions of subjects are determined with the tracker. Finally, data from only one view is used for HPE, BPE and F-formation estimation.

We additionally used samples extracted from the DPOSE dataset (Rajagopal et al. 2014) as auxiliary labeled data for training. DPOSE contains multi-view video sequences depicting a target freely moving in a room monitored by four synchronized cameras with overlapping field of view. In the dataset there are 16 different subjects. Head pose measurements (pan, tilt and roll) are available as acquired using inertial sensors, while body pose is not annotated. Therefore, in our experiments the body pose in each frame was determined using the walking direction as in Benfold and Reid (2011).

4.2 Experimental Setup

Algorithm Parameters The parameters of the algorithm were tuned using a small validation set consisting of random samples amounting to 5% of each dataset. The best values we obtained, namely $\lambda_U = 0.5$, $\lambda_F = 0.2$, $\lambda_C = 0.2$ and $\lambda_R = 0.1$, were fixed for all our experiments. The parameter D, which indicates the O-space radius on the ground plane,

[4] Details on tracking can be found in the supplementary material. Tracking data for Cocktail Party and SALSA datasets are made available at tev.fbk.eu/datasets/cp and tev.fbk.eu/datasets/salsa respectively.

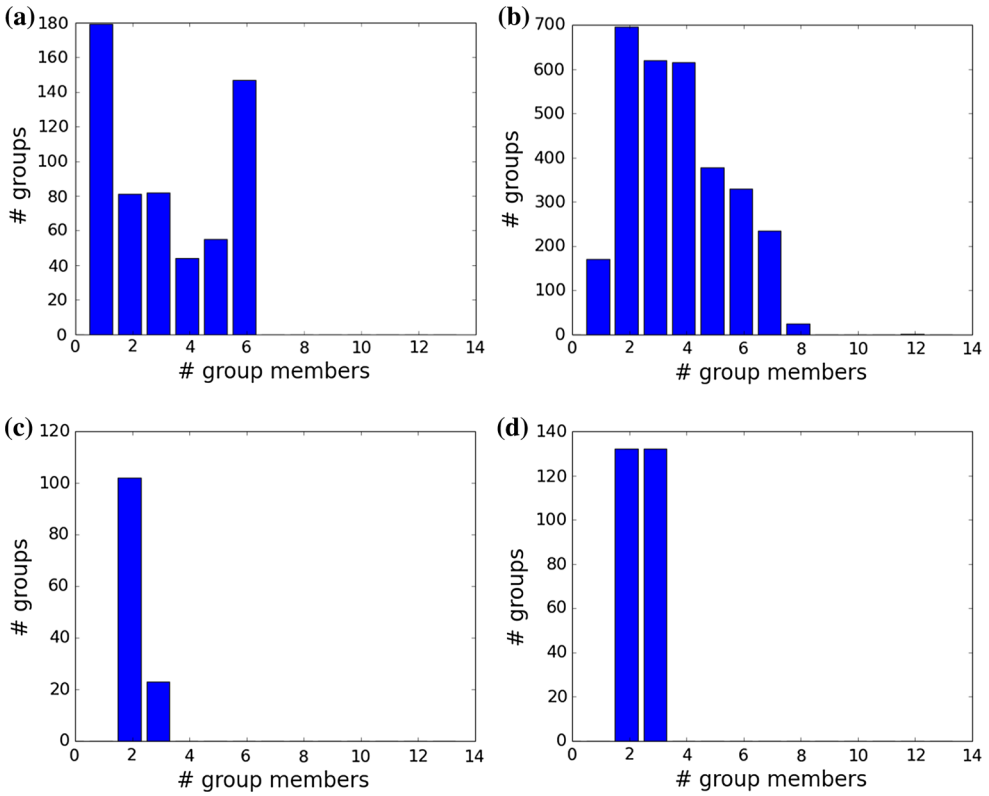

Fig. 6 Histogram of number of groups found in different datasets used in this paper. **a** Cocktailparty, **b** SALSA, **c** Coffeebreak - Seq1, **d** Coffeebreak - Seq 2

was set equal to 0.5 meters in the experiments on Cocktail Party and Coffee Break, while it was set to 0.6 meters for SALSA. This is consistent with previous approaches (Cristani et al. 2011; Vascon et al. 2014), and with sociological studies (Ciolek and Kendon 1980) which fix an upper bound of about 1.2 meters for the typical distance between interacting targets in casual/personal relations. Different temporal smoothness constraints were enforced for the three datasets due to social dynamics and frequency of annotated frames. Specifically, the CB and SALSA datasets are sparsely annotated, with high temporal distance between annotated frames. Therefore, the temporal parameter λ_T was set to low value of 0.01 for those datasets, while we used a higher one, namely 0.1 for CP. The clustering parameter was set to $\gamma_c = 0.2$ for the experiments on Coffee Break and Cocktail Party datasets and $\gamma_c = 0.5$ for SALSA. As shown in Fig. 10, values in the range [0.2, 0.5] provide the best performance. A sensitivity analysis study is provided in Sect. 4.3.

Performance Evaluation To evaluate head and body pose accuracy, we used the mean angular error (in degrees) as commonly used in previous works in Chen and Odobez (2012). To transform the N_C-dimensional output of our algorithm ($\hat{y}_i, i = 1, \ldots, N_C$) to a real-valued angle, we compute the weighted average. Specifically, given a sample $\boldsymbol{x}_{kt}^{\diamond}$ from the social scene, the associated head/body pose $\alpha_{kt}^{\diamond}$ is recovered by computing $\alpha_{kt}^{\diamond} = \text{atan2}(a_{sin}^{\diamond}, a_{cos}^{\diamond})$, where $\boldsymbol{a}^{\diamond} = [a_{cos}^{\diamond}, a_{sin}^{\diamond}]^T = \boldsymbol{A}\, f_{\diamond}(\boldsymbol{x}_{kt}^{\diamond})$. FF estimation accuracy was evaluated using the F1-score as per the standard protocol formulated by previous works (Cristani et al. 2011; Vascon et al. 2014). The F1-score is computed as $F1 = \frac{2PrRe}{Pr+Re}$ and precision and recall are defined as follows: $Pr = \frac{TP}{TP+FP}$ and $Re = \frac{TP}{TP+FN}$. To compute the true positives TP in each frame, we consider a group as correctly estimated if at least $T \cdot |G|$ members are accurately determined, where $|G|$ is the cardinality of the group G and $T = 2/3$. The number of false positives FP and false negatives FN rates are derived by subtracting TP from the cardinality of the detected groups and of the ground truth groups respectively.

Computational Cost Analysis Our algorithm runs on a desktop with a quad-core Intel processor (3.3 GHz) and 8GB RAM. The complexity of the tracking algorithm has a quadratic upper bound on the number of targets. With 50 particles for each target, a C++ implementation runs in real-time for CP and at 7 fps for SALSA dataset. The computational complexity of the head/body localization modules is linear in the number of targets. The joint head/body pose and FF estimation approach is coded in MATLAB (not optimized) and takes on average about 0.1 sec per frame.

Table 2 Average head and body pose estimation error (degrees)

Method	Cocktail party		Coffee break		SALSA	
	Head	Body	Head	Body	Head	Body
AUX ($\lambda_U = \lambda_F = \lambda_C = \lambda_T = 0$)	58.2	65.3	64.3	68.6	58.3	62.4
AUX + SS ($\lambda_F = \lambda_C = \lambda_T = 0$)	51.3	54.7	56.8	58.6	55.3	59.6
AUX + SS + H/B ($\lambda_F = \lambda_T = 0$)	49.4	53.6	52.8	55.6	51.2	54.3
AUX + SS + H/B + FF ($\lambda_T = 0$)	46.5	50.3	46.6	49.4	50.9	51.2
AUX + SS + H/B + FF + T	45.8	48.2	45.3	47.4	50.7	50.2
Chen and Odobez (2012)	48.3	51.7	56.1	57.3	50.2	54.3

Table 3 Performance on F-formation detection (F1-score)

	Cocktail party	Coffee break	SALSA
AUX ($\lambda_U = \lambda_C = \lambda_T = 0$)	0.79	0.78	0.58
AUX + SS ($\lambda_C = \lambda_T = 0$)	0.80	0.82	0.63
AUX + SS + H/B ($\lambda_T = 0$)	0.82	0.84	0.66
AUX + SS + H/B + T	0.85	0.85	0.67

4.3 Results and Discussion

4.3.1 Evaluating Head Pose, Body Pose and F-Formation Estimation Performance

We firstly evaluate the effectiveness of our joint estimation framework on head and body pose estimation (HPE and BPE). Table 2 shows the average HPE, BPE errors on the three considered datasets. The maximum PE error for all datasets is obtained when the objective function only involves the loss term corresponding to auxiliary labeled data (AUX). However, incorporating data from the analyzed social scene (AUX + SS) and coupling head and body pose learning (AUX + SS + H/B) as in (5) considerably reduces pose estimation error. Thereafter, integrating the FF term in (6) further improves pose estimates. This improvement confirms the benefit of jointly estimating body pose of interactors and FFs. Including additional information concerning temporal consistency (T) further reduces PE error, implying that all cues considered in this work are beneficial.

The positive impact of joint learning on F-formation estimation can be noted from Table 3.

This is consistent with our expectation that accurate BPE of interacting targets can aid detection of FFs. Similar to what is observed in Table 2, using unlabeled samples from the social scene in addition to auxiliary data is beneficial. Incorporating additional information such as H/B coupling and temporal consistency in our framework further improves F1-score. It is worth noting that, as expected, both pose and FF estimation are worse for the challenging SALSA dataset as compared to the others, due to the large number of targets and variety of FF configurations observed.

In order to investigate the highest FF detection performance using our formulation, we designed an experiment to detect conversational groups considering only the F-formation-related loss term and using (i) only the target position (human annotation), (ii) the head pose ground truth and (iii) the body pose ground truth. The results of this experiment are given in Fig. 7.

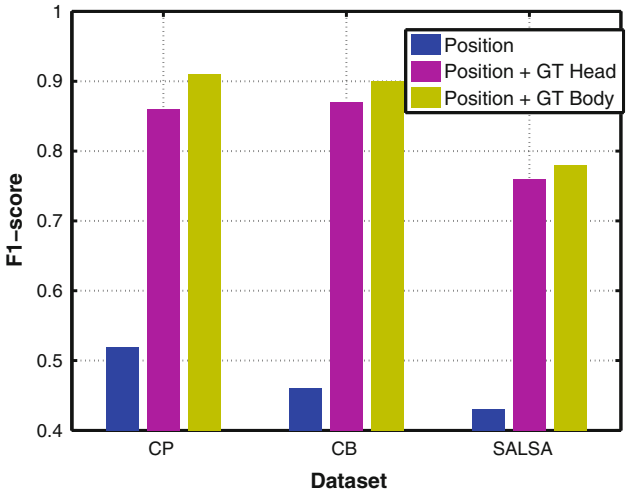

Fig. 7 F1-scores computed with manually annotated data on the CP, CB and SALSA datasets. GT denotes ground truth

Setting (i) is equivalent to putting $D = 0$ and it clearly yields poor results, indicating that feet positions are not sufficient to detect FF accurately (e.g. two targets standing back to back belong to different FFs while being close by). Settings (ii) and (iii) were implemented by using the true target positions for $\boldsymbol{p}_{kt}$ and the head and body pose ground truth instead of $\boldsymbol{A}\, f_{\mathrm{B}}(\boldsymbol{x}_{kt}^{\mathrm{B}})$ in (6). We see a significant increase in FF detection performance when pose cues are utilized along with positional information, with maximum performance achieved using position and body pose cues. This

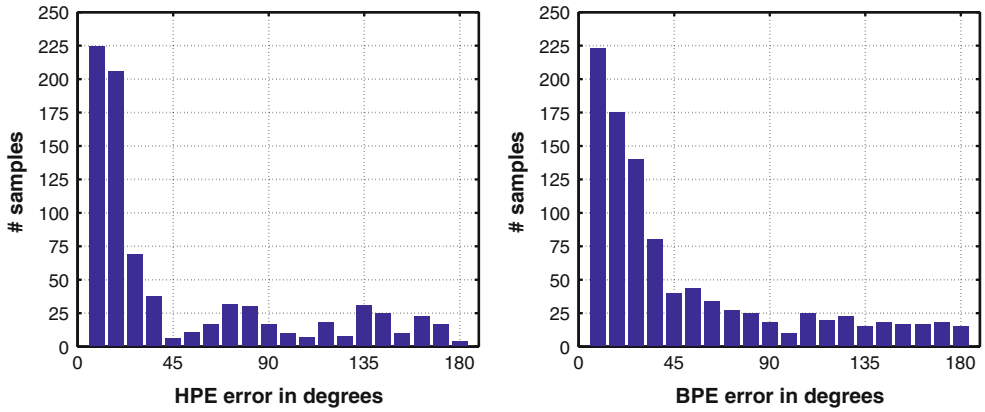

Fig. 8 SALSA dataset. Distribution of (*left*) head and (*right*) body pose estimation errors

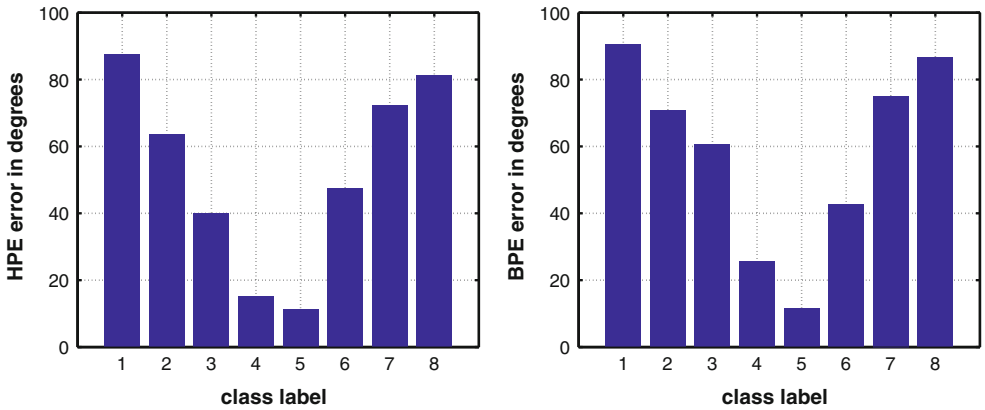

Fig. 9 SALSA dataset. Average (*left*) head and (*right*) body pose estimation errors for different class labels. The eight classes from 1 to 8 refer to angular bins [90, 135), [135, 180), [180, 225), [225, 270), [270, 315), [315, 360), [0, 45), [45, 90) in degrees respectively

confirms our intuition that body pose is a more stable cue as compared to head pose for detecting FFs.

It is also interesting to analyze the head and body pose estimation errors in detail. For our experiments on the SALSA dataset, the bar diagram in Fig. 8 shows the distribution of errors for 1000 random unlabeled samples. It is clear that for most of the samples the estimation error is below 45° (corresponding to predicting the neighboring class). This analysis confirms our claims that our approach is appropriate for the analysis of social scenes: a pose estimation error lower than 45° is enough to detect conversational groups in moderately crowded social scenes, such as poster sessions, cocktail parties or museum visits. This makes our approach suitable to be employed in tasks such as the study of group dynamics or detection of social attractors (Alameda-Pineda et al. 2015). Figure 9 also shows that, if we consider the average errors for samples of a specific class, nearly frontal samples (i.e. classes 4 and 5) are typically better classified. While this is not surprising for head pose estimation, due to better distinctive facial features, we believe that a similar behavior observed for body samples is mostly due to the coupling term.

Finally, Table 4 analyzes the performance of our approach on the CP and CB datasets upon incorporating multiple, occlusion-specific classifiers (we do not report on SALSA as the occlusion rate—per-target average is 22%, see Table 2 in Alameda-Pineda et al. 2016—is too severe for this analysis to provide a basis for conclusion). Our results indicate that the use of multiple occlusion-adaptive regressors reduces head and body pose estimation error, while the improvement in terms of FF detection are rather moderate.

4.3.2 Sensitivity Analysis

We conduct a sensitivity study to analyze the impact of the different hyper-parameters on estimation performance. First, we examine the effect of parameters γ_c and D on F-formation detection accuracy. Figure 10 (left) shows the F1-score at varying γ_c. Low γ_c values preclude clustering of target positions and result in only singleton groups being discovered, thereby resulting in low F1-scores. Conversely, large γ_c values cause multiple FFs to merge as all O-space centroids are constrained to be proximal in such cases (see Eq. (6)), which again adversely impacts detection performance. Inter-

Table 4 Performance improvement with occlusion handling strategy

Method	Cocktail party			Coffee break		
	HP Error	BP Error	FF F1	HP Error	BP Error	FF F1
Our approach	45.8°	48.2°	0.85	45.3°	47.4°	0.85
Our approach (with occlusions)	44.5°	46.6°	0.85	44.2°	46.9°	0.86

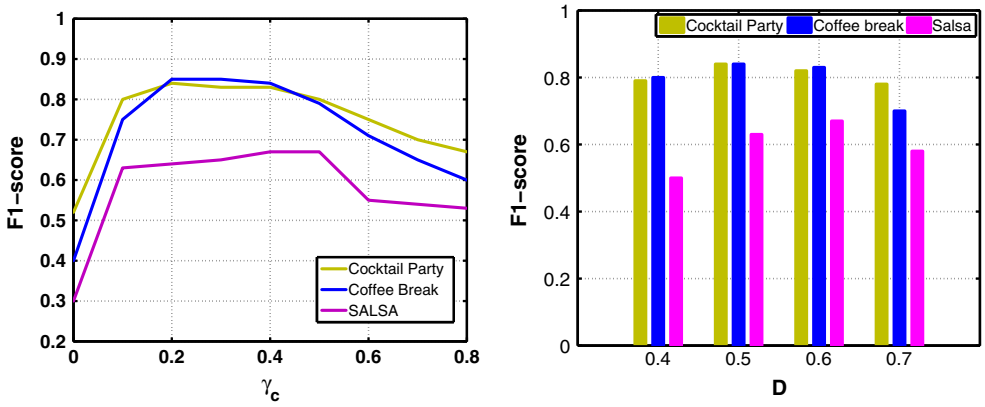

Fig. 10 F1-score obtained from all the three datasets for different γ_c and D values (best viewed in *color*) (Color figure online)

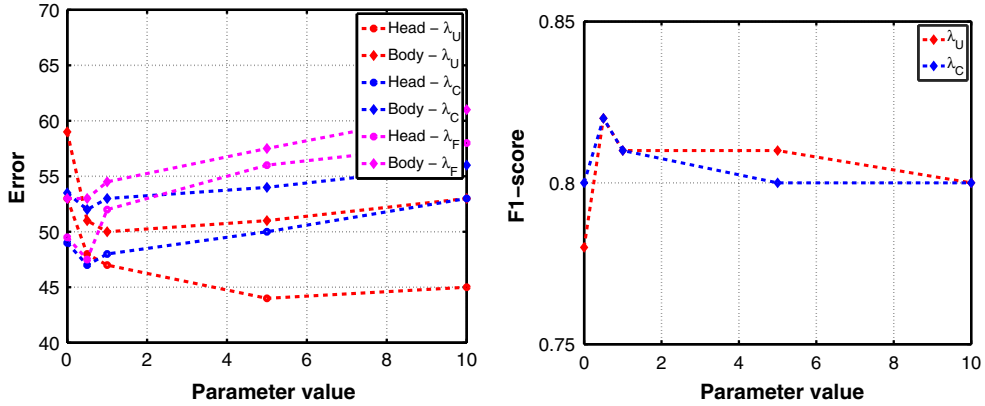

Fig. 11 Cocktail party dataset (*left*) head and body pose estimation error and (*right*) F1-score at varying parameters values (best viewed in *color*) (Color figure online)

estingly, it can be noted from Fig. 10 that for all three datasets, γ_c values in the range $[0.2, 0.5]$ produce the best performance. In our experiments we set $\gamma_c = 0.2$ for Coffee Break and Cocktail Party datasets and $\gamma_c = 0.5$ for SALSA. We also analyze the variation of performance when the parameter D changes. As shown in previous works (Setti et al. 2013), D indicates the typical O-space radius on the ground plane and it is generally set to a value of 0.5 meters. Our experiments (Fig. 10, right) confirm previous findings. Values of D in the range $[0.5, 0.6]$ produce the best performance, while the F1-score degrades for smaller and larger values.

We also evaluate the impact of the parameters on the head and body pose estimation errors. We present our analysis of these parameters on the Cocktail Party dataset. Figure 11 (left) show the results at varying values of λ_U, λ_C and λ_F. Specifically, considering samples from the social scene significantly decreases the head and body pose estimation errors ($\lambda_U \geq 0$). The parameter λ_C also plays an important role. Small values of λ_C improve the performance with respect to the case of not imposing any head and body coupling ($\lambda_C \geq 0$). However, as expected, when the importance of the coupling term increases, accuracy deteriorates (i.e., tight bonds between body and head orientations are detrimental). Similarly, enforcing constraints on body pose according to the conversational groups configuration ($\lambda_F \geq 0$) is beneficial but the performance also degrades when these constraints are too limiting. Similar observations can be made analyzing the performance on F-formation detection (see, Fig. 11 (right)).

Finally, we analyze the contribution of the temporal consistency term in (6). We report the performance at varying

Table 5 Performance on varying parameter λ_T

λ_T	Cocktail party			Coffee break			SALSA		
	HP Error	BP Error	FF F1 score	HP Error	BP Error	FF F1 score	HP Error	BP Error	FF F1 score
0.01	46.1°	50.4°	0.82	45.3°	47.4°	0.85	50.7°	50.2°	0.67
0.1	45.8°	48.2°	0.85	46.3°	49.1°	0.83	51.3°	50.3°	0.67
1	48.2°	50.3°	0.82	49.2°	53.2°	0.82	53.6°	54.4°	0.65

values of λ_T. Table 5 shows the results of our experiments. As discussed above, the relative importance of temporal consistency with respect to other terms in (6) is linked to the frame rate and to the dynamics of the social scene. Therefore, small values of λ_T guarantee the best performance for CB and SALSA, while $\lambda_T = 0.1$ is the optimal choice for CP.

4.3.3 Comparison with State-of-the-Art Methods

We compare our approach with the state-of-the-art for joint head and body pose estimation (Chen and Odobez 2012) in Table 2. To enable this comparison, we implemented the method in Chen and Odobez (2012) in MATLAB. It is worth noting that other recent approaches (Chamveha et al. 2013; Yan et al. 2016; Benfold and Reid 2011) operating on a low resolution setting only consider head pose and do not estimate body pose. The algorithm from Chen and Odobez (2012) is similar to the AUX + SS + H/B setting in our approach, as both these methods focus on coupled learning of head and body pose. However, the small variation in the performance could be ascribed to the use of different optimization techniques to learn the parameters (we adopt an iterative convex optimization method, whereas Chen and Odobez 2012 uses a closed-form solution based on Sylvester equation). However, our algorithm performs significantly better than (Chen and Odobez 2012) when the social context is taken into account, as alternative cues (e.g., velocity direction) are ineffective when targets are mostly static and heavily occluded.

A comparison with state-of-the-art F-formation estimation approaches for the two datasets is presented in Table 6. These include frustrum-based [IRPM (Bazzani et al. 2013), IGD (Tran et al. 2013)], Hough transforms-based [HVFF lin Cristani et al. (2011), HVFF ent (Setti et al. 2013), HVFF ms (Setti et al. 2013)], Graph Cut (GC) (Setti et al. 2015) and Game-theoretic (Vascon et al. 2014) methods. The performance of different methods on the CP and CB datasets are taken from the original papers, while for the SALSA dataset we provide results obtained from code made publicly available by the authors. As shown in the table, we obtain state-of-the-art results on all datasets, clearly outperforming competing methods on the SALSA dataset. It is worth noting that some previous works use orientation annotations available with datasets, and do not automatically estimate the pose. Moreover, most previous approaches are based on sampling techniques, therefore the performance may vary significantly for different runs.

4.4 Qualitative Results

Figures 12 and 13 depict qualitative results generated using our approach for the CP and CB datasets. Specifically, Fig. 12 (top) shows one sequence where our method correctly estimates the head and body pose of all targets, and all FFs are also correctly detected. Figure 12 (bottom) depicts a challenging situation where our method fails: the head and body pose of the yellow target is wrongly estimated in the middle frame owing to extreme occlusion. Similarly, Fig. 13 shows

Table 6 F-Formation estimation performance comparison

Method	Cocktail party			Coffee break			SALSA		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
IRPM (Bazzani et al. 2013)	0.67	0.65	0.66	0.68	0.50	0.57	–	–	–
IGD (Tran et al. 2013)	0.81	0.61	0.70	0.69	0.65	0.67	–	–	–
HVFF lin (Cristani et al. 2011)	0.59	0.74	0.65	0.73	0.86	0.79	0.59	0.62	0.61
HVFF ent (Setti et al. 2013)	0.78	0.83	0.80	0.81	0.78	0.79	0.60	0.61	0.61
HVFF ms (Setti et al. 2013)	0.81	0.81	0.81	0.76	0.86	0.81	0.62	0.64	0.63
GC (Setti et al. 2015)	0.84	0.86	0.85	0.85	0.91	0.88	0.64	0.67	0.65
Game-Th. (Vascon et al. 2014)	0.86	0.82	0.84	0.83	0.89	0.86	0.59	0.63	0.62
Our method	0.87	0.83	0.85	0.84	0.88	0.86	0.66	0.68	0.67

Fig. 12 Qualitative results on Cocktail Party dataset: estimated head pose (*arrows*), body pose (pie on the ground) and F-formations (connections on the ground)

Fig. 13 Qualitative results on Coffee Break dataset: estimated head pose (*arrows*), body pose (pie on the ground) and F-formations (connections on the ground)

some results from the more challenging CB dataset. Evidently, the head and body orientations of most targets are correctly estimated, facilitating reliable FF detection therefrom.

5 Discussion and Conclusions

This paper proposes a novel framework to jointly estimate the head and body pose of targets, and geometric formations involving interacting targets known as F-formations from social scenes. Our approach exploits the synergetic interactor-interaction relationship in social scenes, i.e., the body orientations of interactors facilitate discovery of FFs, and conversely, knowledge of FFs enables robust inference of the interactors' body pose. In contrast to many works that have examined either joint head and body pose estimation from pedestrian scenes, or FF discovery from social scenes, our joint learning accomplishes these twin goals via a single objective function.

The objective function, formulated as a convex function that guarantees a unique global solution, incorporates (i) a loss penalizing prediction errors with respect to the annotated auxiliary dataset, while also enforcing consistency in appearance with unlabeled examples acquired from the social scene under manifold regularization, (ii) a loss penalizing head and body pose predictions violating human anatomic constraints, thereby enabling coupled head and body pose learning, and (iii) the FF-related loss term which facilitates discovery of FFs based on the body pose of interactors, and refines targets' body pose estimates based on the detected FFs. Minimizing each of the above losses is found to improve head and body pose estimates as well as FF detection performance on three benchmark social datasets, demonstrating the efficacy of our joint learning approach. Furthermore, the use of mul-

tiple occlusion-adaptive classifiers and temporal consistency in our model allows to handle bodily occlusions, and noise associated with low-resolution videos.

The proposed approach for handling occlusion was inspired by previous works (Mathias et al. 2013; Wojek et al. 2011) and it has been devised as a first attempt to tackle this challenging problem in the context of F-formations detection. Table 4 shows that an occlusion handling strategy can indeed provide some benefits. However the limitations of occlusion handling become evident in the most challenging contexts, such as that of SALSA dataset that has an average per-target occlusion rate of 22% (cf. Table 2 in Alameda-Pineda et al. 2016). We believe that future works should be devoted to devise more effective strategies to address this problem in complex natural scenarios like that of SALSA dataset.

Future research will also focus on extending the proposed methodology from monocular videos to multi-view settings, and on devising strategies to incorporate additional cues obtained from wearable sensors (e.g., infra-red, blue-tooth and accelerometer), such as to alleviate problems associated with vision-based multi-target tracking and social scene analysis, similar to Alameda-Pineda et al. (2016), Alameda-Pineda et al. (2015).

Inspired by the recent successes of deep learning models applied to computer vision tasks, another future research direction will be replacing handcrafted features, such as HOG, with more robust head and body representations derived from Convolutional Neural Networks. For instance, a first attempt could be to input head and body patches to a pre-trained CNN model (e.g., Krizhevsky et al. 2012) and use the activations from the last fully connected layers directly as a replacement to HOG descriptors used in this paper. However, since the head and body image crops from far field views are typically of very low-resolution, pre-computed CNN features may not necessarily improve the classification performance. Therefore, presuming the availability of a large scale pose annotated dataset, a better way would be to train a CNN model from scratch for head and body pose estimation and use it to predict pose labels on the target dataset with or without fine tuning. Other modules of the proposed framework, e.g.the multi-camera multi-target tracking algorithm and the F-formation detection approach, can be also redesigned considering deep learning models.

6 Appendix: Derivation of Update Rules for $\boldsymbol{\Theta}_{\mathtt{H}}$ and $\boldsymbol{\Theta}_{\mathtt{B}}$

Consider the body and head regressors defined in Sect. 3.4. The update rules for $\boldsymbol{\Theta}_{\mathtt{H}}$ and $\boldsymbol{\Theta}_{\mathtt{B}}$ that we provide in Sect. 3.5 are obtained by setting to zero the partial derivative of the objective function in (2) with respect to $\boldsymbol{\Theta}_{\diamond}$ with $\diamond \in \{\mathtt{B}, \mathtt{H}\}$, and by solving the resulting equations, which are given by

$$l\frac{\partial}{\partial \boldsymbol{\Theta}_{\mathtt{H}}}L_{\mathtt{H}}(f_{\mathtt{H}}(\cdot;\boldsymbol{\Theta}_{\mathtt{H}});\mathcal{T},\mathcal{S}) + \frac{\partial}{\partial \boldsymbol{\Theta}_{\mathtt{H}}}L_C(f_{\mathtt{B}}(\cdot;\boldsymbol{\Theta}_{\mathtt{B}}),\, f_{\mathtt{H}}(\cdot;\boldsymbol{\Theta}_{\mathtt{H}});\mathcal{S}) = 0 \tag{A}$$

$$l\frac{\partial}{\partial \boldsymbol{\Theta}_{\mathtt{B}}}L_{\mathtt{B}}(f_{\mathtt{B}}(\cdot;\boldsymbol{\Theta}_{\mathtt{B}});\mathcal{T},\mathcal{S}) + \frac{\partial}{\partial \boldsymbol{\Theta}_{\mathtt{B}}}L_C(f_{\mathtt{B}}(\cdot;\boldsymbol{\Theta}_{\mathtt{B}}),\, f_{\mathtt{H}}(\cdot;\boldsymbol{\Theta}_{\mathtt{H}});\mathcal{S}) + \frac{\partial}{\partial \boldsymbol{\Theta}_{\mathtt{B}}}L_F(f_{\mathtt{B}}(\cdot;\boldsymbol{\Theta}_{\mathtt{B}}),\boldsymbol{C};\mathcal{S}) = 0 \tag{B}$$

where we have replaced L_P in (2) with its definition in (3). The $L_\diamond$ term is given by

$$L_\diamond(f_\diamond(\cdot;\boldsymbol{\Theta}_\diamond);\mathcal{T}_\diamond,\mathcal{S}) = \sum_{i=1}^{N_\diamond}\|\boldsymbol{\Theta}_\diamond \boldsymbol{v}_i^\diamond - \boldsymbol{y}_i^\diamond\|_{\boldsymbol{M}}^2 + \lambda_R\|\boldsymbol{\Theta}_\diamond\|_F^2 + \lambda_U \sum_{(i,j)\in\mathcal{E}_\diamond}\omega_{ij}^\diamond\|\boldsymbol{\Theta}_\diamond(\boldsymbol{v}_i^\diamond - \boldsymbol{v}_j^\diamond)\|_{\boldsymbol{M}}^2,$$

and its derivative with respect to $\boldsymbol{\Theta}_\diamond$ is

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{\Theta}_\diamond}L_\diamond(f_\diamond(\cdot;\boldsymbol{\Theta}_\diamond);\mathcal{T},\mathcal{S})\\ &= 2\boldsymbol{M}(\boldsymbol{\Theta}_\diamond\hat{\boldsymbol{X}}_\diamond - \boldsymbol{Y}_\diamond)\hat{\boldsymbol{X}}_\diamond^\top + 2\lambda_R\boldsymbol{\Theta}_\diamond + 2\lambda_U\boldsymbol{M}\boldsymbol{\Theta}_\diamond\boldsymbol{V}_\diamond\boldsymbol{L}_\diamond\boldsymbol{V}_\diamond^\top\\ &= 2\boldsymbol{M}\boldsymbol{\Theta}_\diamond(\hat{\boldsymbol{X}}_\diamond\hat{\boldsymbol{X}}_\diamond^\top + \lambda_U\boldsymbol{V}_\diamond\boldsymbol{L}_\diamond\boldsymbol{V}_\diamond^\top) + 2\lambda_R\boldsymbol{\Theta}_\diamond - 2\boldsymbol{M}\boldsymbol{Y}_\diamond\hat{\boldsymbol{X}}_\diamond^\top.\end{aligned}$$

Term L_C is given by

$$L_C(f_{\mathtt{B}}(\cdot;\boldsymbol{\Theta}_{\mathtt{B}}),\, f_{\mathtt{H}}(\cdot;\boldsymbol{\Theta}_{\mathtt{H}});\mathcal{S}) = \lambda_C\sum_{k=1}^{N_K}\sum_{t=1}^{N_T}\|\boldsymbol{\Theta}_{\mathtt{B}}\boldsymbol{x}_{kt}^{\mathtt{B}} - \boldsymbol{\Theta}_{\mathtt{H}}\boldsymbol{x}_{kt}^{\mathtt{H}}\|_{\boldsymbol{M}}^2,$$

and its derivative with respect to $\boldsymbol{\Theta}_\diamond$ is

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{\Theta}_\diamond}L_C(f_{\mathtt{B}}(\cdot;\boldsymbol{\Theta}_{\mathtt{B}}),\, f_{\mathtt{H}}(\cdot;\boldsymbol{\Theta}_{\mathtt{H}});\mathcal{S})\\ &= 2\lambda_C\boldsymbol{M}(\boldsymbol{\Theta}_\diamond\boldsymbol{X}_\diamond - \boldsymbol{\Theta}_\star\boldsymbol{X}_\star)\boldsymbol{X}_\diamond^\top\\ &= 2\lambda_C\boldsymbol{M}\boldsymbol{\Theta}_\diamond\boldsymbol{X}_\diamond\boldsymbol{X}_\diamond^\top - 2\lambda_C\boldsymbol{M}\boldsymbol{\Theta}_\star\boldsymbol{X}_\star\boldsymbol{X}_\diamond^\top,\end{aligned}$$

where $(\diamond,\star) \in \{(\mathtt{H},\mathtt{B}),(\mathtt{B},\mathtt{H})\}$.

Term L_F is given by

$$L_F(f_{\mathtt{B}}(\cdot;\boldsymbol{\Theta}_{\mathtt{B}}),\boldsymbol{C};\mathcal{S}) = \lambda_F\sum_{k=1}^{N_K}\sum_{t=1}^{N_T}\|\boldsymbol{c}_{kt} - (\boldsymbol{p}_{kt} + D\boldsymbol{A}\boldsymbol{\Theta}_{\mathtt{B}}\boldsymbol{x}_{kt}^{\mathtt{B}})\|_2^2 + \text{const},$$

where "const" indicates terms not depending on $\boldsymbol{\Theta}_{\mathtt{B}}$, and its derivative with respect to $\boldsymbol{\Theta}_{\mathtt{B}}$ is

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{\Theta}_{\mathtt{B}}}L_F(f_{\mathtt{B}}(\cdot;\boldsymbol{\Theta}_{\mathtt{B}}),\boldsymbol{C};\mathcal{S})\\ &= 2\lambda_F D\boldsymbol{A}^\top(D\boldsymbol{A}\boldsymbol{\Theta}_{\mathtt{B}}\boldsymbol{X}_{\mathtt{B}} + \boldsymbol{P} - \boldsymbol{C})\boldsymbol{X}_B^\top\\ &= 2\lambda_F D^2\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{\Theta}_{\mathtt{B}}\boldsymbol{X}_{\mathtt{B}}\boldsymbol{X}_{\mathtt{B}}^\top + 2\lambda_F D\boldsymbol{A}^\top(\boldsymbol{P} - \boldsymbol{C})\boldsymbol{X}_B^\top.\end{aligned}$$

By replacing the computed gradient terms in (A), and after few algebraic manipulations, we obtain

$$\boldsymbol{M}\boldsymbol{\Theta}_{\mathrm{H}}(\hat{\boldsymbol{X}}_{\mathrm{H}}\hat{\boldsymbol{X}}_{\mathrm{H}}^{\top}+\lambda_U\boldsymbol{V}_{\mathrm{H}}\boldsymbol{L}_{\mathrm{H}}\boldsymbol{V}_{\mathrm{H}}^{\top}+\lambda_C\boldsymbol{X}_{\mathrm{H}}\boldsymbol{X}_{\mathrm{H}}^{\top})+\lambda_R\boldsymbol{\Theta}_{\mathrm{H}}-\boldsymbol{F}_{\mathrm{H}}=0,$$

and by vectorizing both sides we get

$$\boldsymbol{E}_{\mathrm{H}}\mathrm{vec}(\boldsymbol{\Theta}_{\mathrm{H}})=\mathrm{vec}(\boldsymbol{F}_{\mathrm{H}}) \quad \Longrightarrow \quad \mathrm{vec}(\boldsymbol{\Theta}_{\mathrm{H}})=\boldsymbol{E}_{\mathrm{H}}^{-1}\mathrm{vec}(\boldsymbol{F}_{\mathrm{H}}).$$

By replacing the computed gradient terms in (B), and after few algebraic manipulations, we obtain

$$\begin{aligned}&\boldsymbol{M}\boldsymbol{\Theta}_{\mathrm{B}}(\hat{\boldsymbol{X}}_{\mathrm{B}}\hat{\boldsymbol{X}}_{\mathrm{B}}^{\top}+\lambda_U\boldsymbol{V}_{\mathrm{B}}\boldsymbol{L}_{\mathrm{B}}\boldsymbol{V}_{\mathrm{B}}^{\top}+\lambda_C\boldsymbol{X}_{\mathrm{B}}\boldsymbol{X}_{\mathrm{B}}^{\top})+\lambda_R\boldsymbol{\Theta}_{\mathrm{B}}\\&+\lambda_F D^2\boldsymbol{A}^{\top}\boldsymbol{A}\boldsymbol{\Theta}_{\mathrm{B}}\boldsymbol{X}_{\mathrm{B}}\boldsymbol{X}_{\mathrm{B}}^{\top}-\boldsymbol{G}=0,\end{aligned}$$

and by vectorizing both sides we get

$$\begin{aligned}&(\boldsymbol{E}_{\mathrm{B}}+\lambda_F D^2\boldsymbol{X}_{\mathrm{B}}\boldsymbol{X}_{\mathrm{B}}^{\top}\otimes\boldsymbol{A}^{\top}\boldsymbol{A})\mathrm{vec}(\boldsymbol{\Theta}_{\mathrm{B}})=\mathrm{vec}(\boldsymbol{G}) \implies\\&\mathrm{vec}(\boldsymbol{\Theta}_{\mathrm{B}})=(\boldsymbol{E}_{\mathrm{B}}+\lambda_F D^2\boldsymbol{X}_{\mathrm{B}}\boldsymbol{X}_{\mathrm{B}}^{\top}\otimes\boldsymbol{A}^{\top}\boldsymbol{A})^{-1}\mathrm{vec}(\boldsymbol{G}).\end{aligned}$$

References

Alameda-Pineda, X., Staiano, J., Subramanian, R., Batrinca, L., Ricci, E., Lepri, B., et al. (2016). Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(8), 1707–1720.

Alameda-Pineda, X., Yan, Y., Ricci, E., Lanz, O., & Sebe, N. (2015). Analyzing free-standing conversational groups: A multimodal approach. In *ACM multimedia*.

Alletto, S., Serra, G., Calderara, S., Solera, F., & Cucchiara, R. (2014). From ego to nos-vision: Detecting social relationships in first-person views. In *Workshop on egocentric vision*.

Andriluka, M., Roth, S., & Schiele, B. (2009). Pictorial structures revisited: People detection and articulated pose estimation. In *Computer vision and pattern recognition*, pp. 1014–1021.

Ba, S., & Odobez, J. M. (2008). Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)*.

Ba, S. O., & Odobez, J. M. (2006). A study on visual focus of attention recognition from head pose in a meeting room. In *Machine learning for multimodal interaction*. Springer, Berlin, Heidelberg, pp. 75–87.

Bazzani, L., Tosato, D., Cristani, M., Farenzena, M., Pagetti, G., Menegaz, G., et al. (2013). Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, *30*, 115–127.

Benfold, B., & Reid, I. (2011). Unsupervised learning of a scene-specific coarse gaze estimator. In *International conference on computer vision*.

Butko, T., Canton-Ferrer, C., Segura, C., Giró, X., Nadeu, C., Hernando, J., et al. (2011). Acoustic event detection based on feature-level fusion of audio and video modalities. *Eurasip Journal on Advances in Signal Processing*, *2011*, 485738. doi:10.1155/2011/485738.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., et al. (2006) The ami meeting corpus: A pre-announcement. In *International conference on machine learning for multimodal interaction*, pp. 28–39.

Chamveha, I., Sugano, Y., Sugimura, D., Siriteerakul, T., Okabe, T., Sato, Y., et al. (2013). Head direction estimation from low resolution images with scene adaptation. *Computer Vision and Image Understanding*, *117*(10), 1502–1511.

Chen, C., Heili, A., & Odobez, J. M. (2011). A joint estimation of head and body orientation cues in surveillance video. In *IEEE ICCV-SISM, international workshop on socially intelligent surveillance and monitoring*.

Chen, C., & Odobez, J. M. (2012). We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video. In *Computer vision and pattern recognition*.

Chi, E. C., & Lange, K. (2015). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, *24*(4), 994–1013.

Choi, W., Chao, Y. W., Pantofaru, C., & Savarese, S. (2014). Discovering groups of people in images. In *European conference on computer vision*.

Ciolek, T., & Kendon, A. (1980). Environment and the spatial arrangement of conversational encounters. *Sociological Inquiry*, *50*, 237–271.

Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., et al. (2011) Social interaction discovery by statistical analysis of F-formations. In *British machine vision conference*.

Demirkus, M., Precup, D., Clark, J. J., & Arbel, T. (2014). Probabilistic temporal head pose estimation using a hierarchical graphical model. In *European conference on computer vision*.

Eichner, M., & Ferrari, V. (2010). We are family: Joint pose estimation of multiple persons. In *European conference on computer vision*.

Gan, T., Wong, Y., Zhang, D., & Kankanhalli, M. (2013). Temporal encoded F-formation system for social interaction detection. In *ACM Multimedia*.

Heili, A., Varadarajan, J., Ghanem, B., Ahuja, N., & Odobez, J. M. (2014). Improving head and body pose estimation through semi-supervised manifold alignment. In *International conference on image processing*.

Hocking, T. D., Joulin, A., Bach, F., & Vert, J. P. (2011). Clusterpath an algorithm for clustering using convex fusion penalties. In *International conference on machine learning*.

Hu, T., Messelodi, S., & Lanz, O. (2015). Dynamic task decomposition for decentralized object tracking in complex scenes. *Computer Vision and Image Understanding*, *134*, 89–104.

Krahnstoever, N., Chang, M. C., & Ge, W. (2011). Gaze and body pose estimation from a distance. In *IEEE advanced video and signal-based surveillance (AVSS)*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Leal-Taixé, L., Fenzi, M., Kuznetsova, A., Rosenhahn, B., & Savarese, S. (2014). Learning an image-based motion context for multiple people tracking. In *Computer vision and pattern recognition*.

Liem, M. C., & Gavrila, D. M. (2014). Coupled person orientation estimation and appearance modeling using spherical harmonics. *Image and Vision Computing*, *32*(10), 728–738.

Marin-Jimenez, M., Zisserman, A., Eichner, M., & Ferrari, V. (2014). Detecting people looking at each other in videos. *International Journal of Computer Vision*, *106*(3), 282–296.

Mathias, M., Benenson, R., Timofte, R., & Gool, L. V. (2013). Handling occlusions with franken-classifiers. In *International conference on computer vision*.

Meyer, G. P., Gupta, S., Frosio, I., Reddy, D., & Kautz, J. (2015). Robust model-based 3d head pose estimation. In *International conference on computer vision*.

Murphy-Chutorian, E., & Trivedi, M. M. (2009). Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 607–626.

Patron-Perez, A., Marszalek, M., Reid, I., & Zisserman, A. (2012). Structured learning of human interactions in tv shows. *IEEE Transactions Pattern Analysis and Machine Intelligence*, *34*(12), 2441–2453.

Pellegrini, S., Ess, A., & Van Gool, L. (2010). Improving data association by joint modeling of pedestrian trajectories and groupings. In *European conference on computer vision*.

Rajagopal, A. K., Subramanian, R., Ricci, E., Vieriu, R. L., Lanz, O., & Sebe, N. (2014). Exploring transfer learning approaches for head pose classification from multi-view surveillance images. *International Journal of Computer Vision*, *109*(1–2), 146–167.

Ricci, E., Varadarajan, J., Subramanian, R., Rota Bulo, S., Ahuja, N., & Lanz, O. (2015). Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. In *International conference on computer vision (ICCV)*.

Robertson, N., & Reid, I. (2006). Estimating gaze direction from low-resolution faces in video. In *European conference on computer vision*.

Setti, F., Hung, H., & Cristani, M. (2013). Group detection in still images by F-formation modeling: A comparative study. In *International workshop on image analysis for multimedia interactive services (WIAMIS)*.

Setti, F., Lanz, O., Ferrario, R., Murino, V., & Cristani, M. (2013). Multi-scale F-formation discovery for group detection. In *International conference on image processing*.

Setti, F., Russell, C., Bassetti, C., & Cristani, M. (2015). F-formation detection: Individuating free-standing conversational groups in images. *PLoS ONE*, *10*(5), e0123,783.

Smith, K., Ba, S. O., Odobez, J. M., & Gatica-Perez, D. (2008). Tracking the visual focus of attention for a varying number of wandering people. *IEEE Transaction of Pattern Analysis and Machine Intelligence*, *30*(7), 1212–1229.

Tang, S., Andriluka, M., & Schiele, B. (2014). Detection and tracking of occluded people. *International Journal of Computer Vision*, *110*, 58–69.

Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27, pp. 1799–1807). Red Hook: Curran Associates.

Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Computer vision and pattern recognition*.

Tran, K. N., Bedagkar-Gala, A., Kakadiaris, I. A., & Shah, S. K. (2013). Social cues in group formation and local interactions for collective activity analysis. In *International joint conference on computer vision, imaging and computer graphics theory and applications (VISAPP)*.

Vascon, S., Mequanint, E. Z., Cristani, M., Hung, H., Pelillo, M., & Murino, V. (2014). A game theoretic probabilistic approach for detecting conversational groups. In *Asian conference on computer vision*.

Vascon, S., Mequanint, E. Z., Cristani, M., Hung, H., Pelillo, M., & Murino, V. (2016). Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding*, *143*, 11–24.

Voit, M., & Stiefelhagen, R. (2009). A system for probabilistic joint 3d head tracking and pose estimation in low-resolution, multi-view environments. In *International conference on computer vision systems*, pp. 415–424

Wojek, C., Walk, S., Roth, S., & Schiele, B. (2011). Monocular 3d scene understanding with explicit occlusion reasoning. In *Computer vision and pattern recognition*.

Yan, S., Wang, H., Fu, Y., Yan, J., Tang, X., & Huang, T. (2009). Synchronized submanifold embedding for person-independent pose estimation and beyond. *IEEE Transaction of the Image Processing*, *18*(1), 202–210.

Yan, Y., Ricci, E., Subramanian, R., Lanz, O., & Sebe, N. (2013). No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *International conference on computer vision*.

Yan, Y., Ricci, E., Subramanian, R., Liu, G., Lanz, O., & Sebe, N. (2016). A multi-task learning framework for head pose estimation under target motion. *IEEE Transaction of the Pattern Analysis and Machine Intelligence*, *38*(6), 1070–1083.

Zen, G., Lepri, B., Ricci, E., & Lanz, O. (2010). Space speaks: Towards socially and personality aware visual surveillance. In *ACM multimedia workshop on multimodal pervasive video analysis*.

Zhu, X. (2005). *Semi-supervised learning literature survey*. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *3*(1), 1–130.