Uncovering Interactions and Interactors: Joint Estimation of Head, Body Orientation and F-formations from Surveillance Videos

Elisa Ricci[1,3], Jagannadan Varadarajan[2], Ramanathan Subramanian[2],
Samuel Rota Bulò[1], Narendra Ahuja[2,4], Oswald Lanz[1]
[1] *Fondazione Bruno Kessler, Trento, Italy*
[2] *Advanced Digital Sciences Center, University of Illinois at Urbana-Champaign, Singapore*
[3] *Department of Engineering, University of Perugia, Italy*
[4] *University of Illinois at Urbana-Champaign, IL USA*
{eliricci,rotabulo,lanz}@fbk.eu, {vjagan,subramanian.r}@adsc.com.sg, {n-ahuja}@illinois.edu

Abstract

We present a novel approach for jointly estimating targets' head, body orientations and conversational groups called F-formations from a distant social scene (e.g., a cocktail party captured by surveillance cameras). Differing from related works that have (i) coupled head and body pose learning by exploiting the limited range of orientations that the two can jointly take, or (ii) determined F-formations based on the mutual head (but not body) orientations of interactors, we present a unified framework to jointly infer both (i) and (ii). Apart from exploiting spatial and orientation relationships, we also integrate cues pertaining to temporal consistency and occlusions, which are beneficial while handling low-resolution data under surveillance settings. Efficacy of the joint inference framework reflects via increased head, body pose and F-formation estimation accuracy over the state-of-the-art, as confirmed by extensive experiments on two social datasets.

1. Introduction

Following decades of research progress, head and body pose estimation (PE) is now possible in challenging settings where persons are captured at prohibitively low-resolution with blurred facial and body parts, or moving unconstrained in an environment with uneven illumination. Buoyed by the success of PE algorithms that can robustly handle facial appearance variations [7, 36] and learn with limited training data by exploiting anatomic constraints [5, 8], computer vision research has begun to focus on complex phenomena like social interactions.

The ability to detect conversational groups or *F-formations* [11] in social scenes (Fig.1 (left)) is critical for a variety of applications such as surveillance, social robotics and behavior analysis. Formally, an F-formation arises naturally in conversational settings whenever two or more individuals in close proximity orient their bodies such that each of them has an easy, direct and equal access to every other participant's transactional segment [11]. For example, when two persons interact, the typical formation arrangements are vis-a-vis, L-shape and side-by-side (Fig.1 (right)). The fact that F-formations are characterized by the shared physical locations and head, body orientations of interactors has been exploited by several works [12, 33]. In these works, an F-formation is typically computed by determining interacting members and the center of the O-space [11], *i.e.*, the center of the smallest empty convex space encompassed by the interactors (Fig.1 (center)).

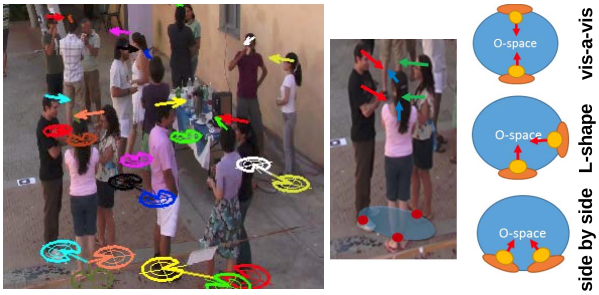

Figure 1: **Problem overview:** (Left) Social scene from the Coffeebreak dataset [12]. We *jointly* estimate conversational groups and the head, body pose of conversing targets, both of which are non-trivial due to low-resolution and extreme occlusions. Circles around targets' feet positions denote body pose, arrows denote head pose, and lines connecting the circles signify F-formations estimated using our method. (Center) Three member F-formation where feet positions, head and body orientations are shown– corresponding O-space is denoted using the blue ellipse. (Right) Exemplar F-formations with two targets.

Challenges in visual analysis of conversational groups

1550-5499/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCV.2015.529

are manifold (Fig.1 (left)). First, accurately determining head and body pose of the conversing targets is non-trivial due to the presence of heavy facial and bodily occlusions, background clutter and the difficulty in characterizing body pose due to extensive variability in clothing. Also, employing cues such as walking direction, as proposed in prior works considering pedestrian scenes [5, 8], is ineffective as F-formations are defined by relatively static arrangements of individuals. State-of-the-art F-formation detection approaches [12, 33] rely on pre-trained head pose classifiers and coarse orientation quantization. However, F-formation discovery is hard when pose classifiers are not adapted to the considered social scene.

To tackle the above problems, we present the first work to *jointly* estimate targets' head and body pose and F-formations in a social scene captured by a surveillance camera, as illustrated in Fig.1. Different from prior works which have focused solely on (i) jointly learning head and body pose of individuals based on anatomic constraints [5, 8], or (ii) F-formation detection from spatial and head orientation cues [12, 33], we present a joint framework to infer both (i) and (ii). Our approach exploits the synergetic interaction-interactor relationship, *i.e.*, F-formations are characterized by mutual locations and head, body orientations of interactors, while conversely, interactors are constrained in terms of the head and body pose they can exhibit, motivating the need for joint learning. Specifically, our novel learning framework (i) exploits both annotated and unlabeled data to learn the range of joint head-body orientations of individuals, (ii) exploits positional and pose-based constraints relating interactors to discover F-formations, and (iii) further refines pose estimates of interactors based on the gained knowledge concerning F-formations and vice-versa.

Our work has several unique aspects. Firstly, we use body orientation as the primary cue for determining F-formations. While prior works [12, 33] have acknowledged the importance of body pose for deducing F-formations, they nevertheless use head pose estimates in their analysis given the adverse impact of occlusions on body pose estimation. The use of head orientation is nevertheless spurious as it is prone to frequent changes during social interactions. In contrast, body pose is a more stable cue, and can better express the geometrical F-formation arrangement. Secondly, in order to estimate body pose precisely, our learning framework couples head and body pose learning as in [8], but also handles occlusions by adopting multiple occlusion-specific regression functions. Finally, temporal consistency is also enforced to ensure smoothness in head, body and F-formation estimates over time.

Contributions: (i) We present a novel framework for jointly estimating individuals' head, body orientations and F-formations in social scenes. Via thorough experiments on two challenging social datasets, we demonstrate the benefits of our joint learning framework against competing pose and F-formation estimation methods. (ii) In contrast to existing methods, we employ body orientation as the primary cue for estimating F-formations. Computation of precise body pose estimates is achieved via coupled head and body pose learning, knowledge gained regarding F-formations and handling varying levels of body occlusion with multiple regressors. (iii) Our model also enforces temporal consistency with respect to estimated pose and group memberships which is particularly useful as tracking and cropping errors are commonplace in low-resolution surveillance videos.

2. Related Work

We now review prior work in topics most related to this work, namely, head and body pose estimation (HPE and BPE) from surveillance video, and detection of social interactions and conversational groups from social scenes.

Head and body pose estimation. Recently, coarse head and body PE from surveillance videos has been investigated by many works [5, 7, 8, 16, 27, 36], as pose represents an important cue in surveillance and human behavior analysis. Pioneering work in [27] proposes HPE with eight directional classes. In [5], unsupervised HPE is presented by exploiting weak labels in the form of walking direction of pedestrians. A similar idea is also exploited in [7]. Chen *et al.* [8] novelly compute head pose by introducing two coupling factors, one between head and body pose and another between body pose and velocity direction. Furthermore, they introduce classifier adaptation via manifold learning. An adaptive transfer learning framework for multi-view HPE under target motion is proposed by Rajagopal *et al.* [26]. Yan *et al.* [36] address the same problem by modeling facial similarities and differences among neighboring scene regions using multi-task learning. Recent HPE approaches are able to cope with label noise [15] and integrate temporal consistency [13].

BPE from surveillance video has been studied by few works [19, 27], which only consider body orientation as a link between walking direction and head pose, but do not explicitly learn body pose classifiers. Recent works of Chen *et al.* [8] and Liem *et al.* [21] demonstrate the benefits of coupling HPE and BPE. However, most PE works focus on pedestrian scenes involving non-interacting individuals, while we expressly consider complex social scenes. Typically, prior approaches do not work well when targets remain static (for BPE), or are observed under large occlusions (as most methods are monocular). For instance, experiments in [5, 8] show poor PE performance when targets are either static or their velocity is noisy. Similarly, Yan *et al.* [36] alleviate the occlusion problem by considering multi-view images, but do not implement specific strategies for handling varying levels of body occlusion.

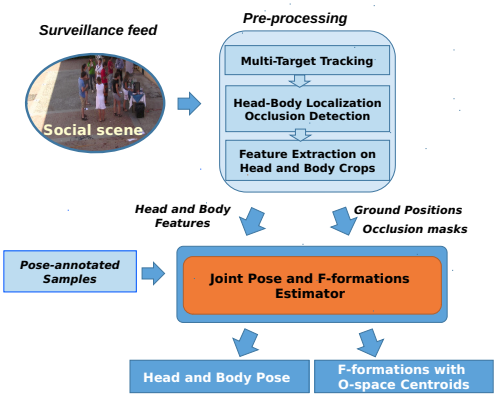

Figure 2: Overview of our social scene analysis framework.

Social interactions and conversational groups. Recently, there has been considerable interest in analyzing social interactions and social scenes. Jimenez *et al.* , [22] propose continuous HPE using Gaussian process regression, and evaluate several methods for detecting dyadic interactions in a video shot. Perez *et al.* [24] achieve spatio-temporal localization of dyadic interactions from TV videos using a structured SVM, combining information from local (pose-based) and global (position-based) descriptors. Choi *et al.* [10] recognize group activities by analyzing the spatial configuration of group members. Other works have focused on (i) detecting groups instead of individuals in static images to overcome partial occlusions [14, 31] and (ii) leveraging information about groups to improve multi-target tracking performance [20, 25].

Detecting conversational groups or F-formations in social scenes has generated interest lately due to security and commercial applications [12, 29, 30]. Cristani *et al.* [12] 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 [29]. Social interactions are detected in egocentric videos in [3]. Vascon *et al.* [33] propose a game-theoretic approach for determining F-formations using position and head pose cues which allows for systematic integration of temporal information. While orientation relationships among interactors have been exploited for detecting conversational groups, joint estimation of such groups and the head, body pose of targets has not been attempted. In this paper, we show that joint learning benefits both pose and F-formation estimation.

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. Thanks to several state-of-the-art tracking methods [6, 18, 34], we can now deal with complexities in social scenes due to occlusion and clutter. Target feet positions are estimated with the multi-target tracking approach in [18] and are used for head localization and cropping via a 3D head-plus-shoulder model registered through shape matching as in [36]. Each target's body region is determined as the portion between head and feet coordinates. 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 for the head and body regions (see Subsection 3.2.3). Targets' positions, head and body features along with occlusion masks are input to our joint learning algorithm that outputs for each target (i) head and body pose and (ii) F-formation membership as described in the following.

3.2. Inferring head, body pose and F-formations

3.2.1 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 $(\boldsymbol{x}^{\mathtt{B}}_{kt}, \boldsymbol{x}^{\mathtt{H}}_{kt}, \boldsymbol{p}_{kt})$, providing for each frame t the body and head descriptors denoted by $\boldsymbol{x}^{\mathtt{B}}_{kt} \in \mathcal{X}_{\mathtt{B}}$ and $\boldsymbol{x}^{\mathtt{H}}_{kt} \in \mathcal{X}_h$ respectively, and the target's feet position $\boldsymbol{p}_{kt} \in I\!R^2$. Here, $\mathcal{X}_{\mathtt{B}}$ and $\mathcal{X}_{\mathtt{H}}$ represent the feature spaces associated to body and head samples respectively. Information concerning all video targets is collected in $\mathcal{S} = \{(\boldsymbol{x}^{\mathtt{B}}_{kt}, \boldsymbol{x}^{\mathtt{H}}_{kt}, \boldsymbol{p}_{kt})\}_{kt}$, with $k \in \langle N_K \rangle$ and $t \in \langle N_T \rangle$, where $\langle N \rangle = \{1, \ldots, N\}$ for notational convenience.

The goal of the inference task is to estimate the body pose $\alpha^{\mathtt{B}}_{kt} \in [0, 2\pi)$, the head pose $\alpha^{\mathtt{H}}_{kt} \in [0, 2\pi)$ 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 [26, 36], we estimate only the head and body *pan*. F-formations are determined by all targets sharing the membership z_{kt} (*i.e.* at frame t two targets k and h belong to the same group 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 $\mathcal{T}_{\mathtt{B}} = \{(\hat{\boldsymbol{x}}^{\mathtt{B}}_i, \boldsymbol{y}^{\mathtt{B}}_i)\}_{i=1}^{N_{\mathtt{B}}} \subseteq \mathcal{X}_{\mathtt{B}} \times \mathcal{Y}$ and $\mathcal{T}_{\mathtt{H}} = \{(\hat{\boldsymbol{x}}^{\mathtt{H}}_i, \boldsymbol{y}^{\mathtt{H}}_i)\}_{i=1}^{N_{\mathtt{H}}} \subseteq \mathcal{X}_{\mathtt{H}} \times \mathcal{Y}$ to enhance the head and body pose estimation capabilities of our model. Each training sample in $\mathcal{T}_\diamond$, where $\diamond \in \{\mathtt{B}, \mathtt{H}\}$, 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[1]

[1]Most available datasets on HBPE in low resolution settings only provide quantized pose annotations.

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).

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 the $\diamond$ is used as a placeholder for H or B.

3.2.2 Jointly inferring pose and F-formations

The inference problem that we face is semi-supervised, as we have both annotated data from $\mathcal{T} = (\mathcal{T}_\mathtt{B}, \mathcal{T}_\mathtt{H})$ and *non-annotated* observations $\mathcal{S}$ from the video under analysis. The head and body pose annotations from $\mathcal{T}$ implicitly provide a prior for estimating the pose of targets in $\mathcal{S}$. No annotation of F-formations is used during learning.

In order to exploit the distribution of descriptors corresponding to annotated data and scene targets, we introduce two regression functions $f_\mathtt{B}$ and $f_\mathtt{H}$ for the body and head pose respectively, which are two unknowns in our model. Intuitively, $f_\diamond : \mathcal{X}_\diamond \to I\!R^{N_C}$ 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 $I\!R^{N_C}$. 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 $\boldsymbol{A}\, f_\diamond(\boldsymbol{x}^\diamond_{kt}) \in I\!R^2$ can finally be cast in polar coordinates to recover the pose angles $\alpha^\diamond_{kt}$ corresponding to $\boldsymbol{x}^\diamond_{kt}$ in $\mathcal{S}$.

Assignment of targets to F-formations is modeled indirectly by letting each target vote for the center of the F-formation 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 center of the F-formation for target k in frame t. We assume these centers, which will become additional unknowns of our model, to be stacked into a $2 \times N_K N_T$-dimensional matrix $\boldsymbol{C}$. We denote by $\mathcal{C} = I\!R^{2 \times N_K N_T}$, the set of all such matrices. Given $\boldsymbol{C}$, the corresponding F-formation assignments z_{kt} can be easily recovered as shown in [17]. 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.

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 $\mathcal{S}$:

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

where $\mathcal{F}_\diamond$ is the space of pose regressors $f_\diamond$ (details on pose regressor spaces are given in Subsection 3.2.3). 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 F-formations' center votes given by $\boldsymbol{C}$. Given the optimal solution to (2), we recover the head, body pose $\alpha^\diamond_{kt}$ and F-formation assignment z_{kt} of each target at every frame as discussed above. We now describe L_P and L_F in detail.

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

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

The first two loss terms penalize pose regressor 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 $\mathfrak{G}_\mathtt{H}$ and $\mathfrak{G}_\mathtt{B}$ for the available head and body samples. For each $\diamond \in \{\mathtt{H}, \mathtt{B}\}$, 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 i^{th} one denoted by $\boldsymbol{v}^\diamond_i \in \mathcal{X}_\diamond$. For all annotated samples in $\mathcal{V}_\diamond$, *i.e.*, $\forall i \in \langle N_\diamond \rangle$, we indicate the corresponding pose label by $\boldsymbol{y}^\diamond_i$. The set $\mathcal{E}_\diamond \subseteq \langle |\mathcal{V}_\diamond| \rangle^2$ indexes pairs of neighboring vertices, while $\omega^\diamond_{ij} \geq 0$ is a non-negative weight indicating the strength of the (i, j)-edge connection. More details will be given in Subsection 3.2.3.

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

$$\begin{aligned} L_\diamond(f; \mathcal{T}_\diamond, \mathcal{S}) = \sum_{i=1}^{N_\diamond} \|f(\boldsymbol{v}^\diamond_i) - \boldsymbol{y}^\diamond_i\|^2_M + \lambda_R \|f\|^2_{\mathcal{F}_\diamond} \\ + \lambda_U \sum_{(i,j) \in \mathcal{E}_\diamond} \omega^\diamond_{ij} \|f(\boldsymbol{v}^\diamond_i) - f(\boldsymbol{v}^\diamond_j)\|^2_M\,, \end{aligned} \tag{4}$$

where $\|\cdot\|_{\mathcal{F}_\diamond}$ is a semi-norm for the function space $\mathcal{F}_\diamond$, and $\|\boldsymbol{a}\|_M = \sqrt{\boldsymbol{a}^\top \boldsymbol{M} \boldsymbol{a}}$ is a semi-norm on $I\!R^{N_C}$ induced by the symmetric, positive semi-definite matrix $\boldsymbol{M} \in I\!R^{N_C \times N_C}$, which accounts for the semantic mapping from the pose label vectors $\in I\!R^{N_C}$ to angles in $\boldsymbol{\alpha}$ (see Subsection 3.2.3).

The first term in $L_\diamond$ measures the prediction error of $f \in \mathcal{F}_\diamond$ with respect to the annotated training set; the second 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 nonnegative 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 [38]. 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):

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

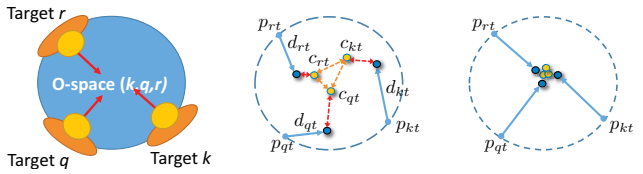

Figure 3: (left) O-space of the F-formation involving three targets k, r and q and their body pose. (center) Direction vectors $\boldsymbol{d}_{(.)}$ obtained via body pose regressor are shown using blue arrows, while $\boldsymbol{c}_{(.)}$ (yellow points) denote voted center locations. By minimizing (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 $\lambda_T = 0$.

where λ_C is a free, nonnegative parameter.

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 F-formations from body pose estimates of interacting targets.

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

$$L_F(f_{\mathtt{B}}, \boldsymbol{C}; \mathcal{S}) = \lambda_F \sum_{k=1}^{N_K} \sum_{t=1}^{N_T} \|\boldsymbol{c}_{kt} - (\boldsymbol{p}_{kt} + D\boldsymbol{A}\, f_{\mathtt{B}}(\boldsymbol{x}^{\mathtt{B}}_{kt}))\|_2^2$$
$$+ \gamma_c \sum_{k,h=1}^{N_K} \sum_{t=1}^{N_T} \|\boldsymbol{c}_{kt} - \boldsymbol{c}_{ht}\|_1 + \lambda_T \sum_{k=1}^{N_K} \sum_{t=2}^{N_T} \|\boldsymbol{c}_{kt} - \boldsymbol{c}_{k(t-1)}\|_1 \tag{6}$$

where $\|\cdot\|_p$ is the p-norm, and λ_F, D, γ_c and λ_T are nonnegative, 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 [12, 33]). The body orientation for the kth target at frame t in $\mathbb{R}^2$ is obtained as $\boldsymbol{A}\, f_{\mathtt{B}}(\boldsymbol{x}^{\mathtt{B}}_{kt})$, since the output of f is the prediction of the pose label. Hence, his/her ideal F-formation center position $\boldsymbol{c}_{kt}$ is given by $\boldsymbol{p}_{kt} + \boldsymbol{d}_{kt}$, where $\boldsymbol{d}_{kt} = D\boldsymbol{A}\, f_{\mathtt{B}}(\boldsymbol{x}^{\mathtt{B}}_{kt})$. 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 F-formations in the scene. Note that the 1-norm induces the centroids of targets belonging to the same F-formation to merge. Finally, the third term enforces temporal consistency of the targets' center votes, given the fact that conversational groups do not change rapidly over time.

In contrast to prior works which use head orientations to infer F-formations, 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 F-formations, and this reflects via improved F-formation and body pose estimation accuracy as discussed in Section 4.

3.2.3 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 $\boldsymbol{\Theta}_{\mathtt{B}} \in \mathbb{R}^{N_C \times M_{\mathtt{B}}}$ and $\boldsymbol{\Theta}_{\mathtt{H}} \in \mathbb{R}^{N_C \times M_{\mathtt{H}}}$, 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. In our experiments we consider a linear kernel.

To facilitate comparisons with previous works [8,26], 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 [8, 26] and consistently with annotations of most datasets in 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 i^{th} sample $\boldsymbol{v}^\diamond_i \in \mathcal{V}_\diamond$ is among the k-nearest neighbors of the j^{th} sample $\boldsymbol{v}^\diamond_j \in \mathcal{V}_\diamond$ under the standard Euclidean metric. Moreover, temporal smoothing is enforced by imposing that $(i, j) \in \mathcal{E}_\diamond$ if samples $\boldsymbol{v}^\diamond_i$ and $\boldsymbol{v}^\diamond_j$ correspond to samples $\boldsymbol{x}^\diamond_{kt}$ and $\boldsymbol{x}^\diamond_{kt'}$ 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.*, $\omega^\diamond_{ij} = 1$. 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 Section 4.

3.2.4 Optimization

By taking (8) as the regressors' space and by rewriting the minimization in (2) in terms of $\boldsymbol{\Theta}_\diamond$ as mentioned in Subsection 3.2.3, we obtain a *convex* optimization problem with variables $(\boldsymbol{\Theta}_\mathtt{B}, \boldsymbol{\Theta}_\mathtt{H}, \boldsymbol{C})$, 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 $\boldsymbol{\Theta}_\mathtt{B}$, $\boldsymbol{\Theta}_\mathtt{H}$ and $\boldsymbol{C}$. Before delving into details, we introduce the following matrices: $\boldsymbol{X}_\diamond = (\boldsymbol{x}^\diamond_{11}, \dots, \boldsymbol{x}^\diamond_{N_K N_T})$, $\hat{\boldsymbol{X}}_\diamond = (\hat{\boldsymbol{x}}^\diamond_1, \dots, \hat{\boldsymbol{x}}^\diamond_{N_\diamond})$, $\boldsymbol{Y}_\diamond = (\boldsymbol{y}^\diamond_1, \dots, \boldsymbol{y}^\diamond_{N_\diamond})$ 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 Subsection 3.2.2, 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 $(\diamond, \star) \in \{(\mathtt{H}, \mathtt{B}), (\mathtt{B}, \mathtt{H})\}$, $\otimes$ is the Kronecker product and $\boldsymbol{I}$ is a properly-sized identity matrix. In the following we briefly describe our iterative optimization framework. Further details are provided in the supplementary material.

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

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

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

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

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

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

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 [9].

3.2.5 Handling Occlusions

We now show how the proposed framework can be extended to integrate information about body occlusions. To factor in the level of occlusion while estimating body pose, we first calculate an occlusion map using the camera geometry and target locations (Subsection 3.1). 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 (6). A similar strategy has been used by previous pedestrian detection works [23, 35] producing significant improvement over single classifiers. To our knowledge, no prior work has adopted such an approach for estimating body pose. In this work, we consider $O = 4$ different pose regressors $f^o_\mathtt{B}$, $o = 1, \dots, O$. In previous approaches [23, 35], a convex combination of the occlusion-specific classifier scores is considered at test time. Differently, to keep the computational cost limited, we partition the body samples extracted from the social scene into four groups, according to the detected level of occlusion (a region is considered occluded if at least 50% of the pixels are not visible). 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 Subsection 3.2.4) reduces to solving a set of O independent optimization problems while learning $f^o_\mathtt{B}$ and $f_\mathtt{H}$. Conversely, while learning $\boldsymbol{C}$, the appropriate occlusion-specific regressor $f^o_\mathtt{B}$ is invoked for each sample $\mathbf{x}^\mathtt{B}_{k,t}$, according to its occlusion level. While our approach can be also used to model head occlusions, we consider only body occlusions as they more severely impact PE performance and to keep the computational cost limited.

Table 1: Mean HBPE error (degrees).

	CP		CB	
Method	Head	Body	Head	Body
AUX ($\lambda_U = \lambda_F = \lambda_C = \lambda_T = 0$)	58.2	65.3	64.3	68.6
AUX + SS ($\lambda_F = \lambda_C = \lambda_T = 0$)	51.3	54.7	56.8	58.6
AUX + SS + H/B ($\lambda_F = \lambda_T = 0$)	49.4	53.6	52.8	55.6
AUX + SS + H/B + FF ($\lambda_T = 0$)	46.5	50.3	46.6	49.4
AUX + SS + H/B + FF + T	45.8	48.2	45.3	47.4
AUX + SS + H/B + FF + T + O	**44.5**	**46.6**	**44.2**	**46.9**
Chen *et al.* [8]	48.3	51.7	56.1	57.3

4. Experimental Results

4.1. Datasets and Experimental Setup

Datasets. We found only two datasets with time-continuous F-formation annotations for evaluating our algorithm, and present experimental results on the same.

The **CocktailParty** dataset [37] (CP) contains a 30-minute video recording of a cocktail party in a $30\mathrm{m}^2$ room involving six subjects. The social event is recorded using four synchronized wall-mounted cameras (512×384 pixels, jpeg). Consistently with previous works [12,33], we use data from camera 1. This sequence is challenging for video analysis due to low-resolution of the targets' faces, background clutter as well as frequent and persistent occlusions. Target positions are logged via a 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

Table 2: Performance on F-formations detection (F1-score).

Method	CP	CB
AUX ($\lambda_U = \lambda_C = \lambda_T = 0$)	0.79	0.78
AUX + SS ($\lambda_C = \lambda_T = 0$)	0.80	0.82
AUX + SS + H/B ($\lambda_T = 0$)	0.82	0.84
AUX + SS + H/B + T	0.85	0.85
AUX + SS + H/B + T + O	**0.85**	**0.86**

those frames where F-formation annotations are available. F-formation annotations are available every five seconds.

The **CoffeeBreak** dataset [12] (CB) again depicts a social event and comprises a maximum of 14 targets, organized in groups of 2-3 persons. Target positions are annotated using a tracker, while head and body pose are annotated by an expert to incorporate eight classes (original dataset has only head annotations with four classes). F-formations are annotated for two sequences of lengths 45 and 75 frames respectively.

We additionally use samples extracted from the DPOSE dataset [26] as auxiliary labeled data for training. DPOSE contains head pose measurements acquired using inertial sensors, while body pose in each frame is determined using walking direction as in [5]. Note that in our approach only labels from DPOSE are used during learning while the annotations in CB and CP are only used for evaluation.

Experimental Setup. Algorithm parameters are fixed using a small validation set. Specifically, $\lambda_U = 0.5$, $\lambda_F = 0.2$, $\lambda_C = 0.2$, $\lambda_R = 0.1$ and these values are identical for the two datasets. Parameter D, which indicates the associated O-space radius, is set equal to 0.5 meters on the ground plane. This is consistent with previous approaches [12, 33], and with sociological studies [11] which fix an upper bound of about 1.2 m for the typical distance between interacting targets in casual/personal relations. As different temporal smoothness constraints need to be enforced for the CP and CB datasets due to social dynamics and frequency of annotated frames. temporal parameter λ_T is set to 0.1 and 0.01 respectively. Finally, the parameter γ_c is particularly important and its role is discussed in the following subsection.

To evaluate HPE and BPE accuracy, we use the mean angular error (in degrees). Specifically, given a sample $\boldsymbol{x}^{\diamond}_{kt}$ from the social scene, the associated head/body pose $\alpha^{\diamond}_{kt}$ is recovered by computing $\alpha^{\diamond}_{kt} = \text{atan2}(a^{\diamond}_{sin}, a^{\diamond}_{cos})$, where $\boldsymbol{a}^{\diamond} = [a^{\diamond}_{cos}, a^{\diamond}_{sin}]^T = \boldsymbol{A}\, f_{\diamond}(\boldsymbol{x}^{\diamond}_{kt})$. F-formation estimation accuracy is evaluated using F1-score [12,33]. In each frame we consider a group as correctly estimated if at least $T \cdot |G|$ of the members are correctly found and if no more than $1 - (T \cdot |G|)$ non-members are wrongly identified, where $|G|$ is the cardinality of the group G and $T = 2/3$.
Our method runs on a desktop with a quad-core Intel processor (3.3GHz) and 8GB RAM. The tracking and head/body localization modules, implemented in C++, run in real-time. The HBPE and F-formation detection are coded in MATLAB and take about 1 sec each 10 frames.

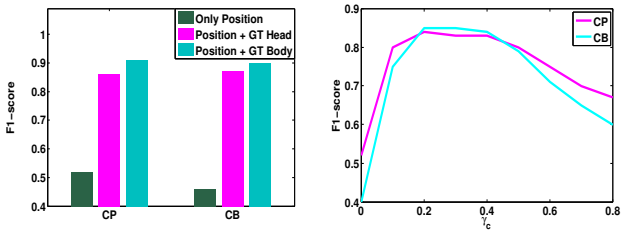

Figure 4: (left) F1-score computed with annotated data. (right) F1-score at varying γ_c (best viewed in color).

4.2. Results and Discussion

Head, body pose estimation (HBPE). We firstly evaluate the effectiveness of our joint estimation framework on head and body pose estimation. Table 1 shows the average HPE, BPE errors on the CP and CB datasets. Maximum error of about $68°$ is obtained for both datasets 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 HBPE error for both datasets. Thereafter, integrating the F-formation (FF) term in (6) further improves pose estimates by about $3°$ and $6°$ for CP and CB respectively. This improvement confirms the benefit of jointly estimating body pose of interactors and F-formations. Including additional information concerning occlusions (O) and temporal consistency (T) further reduces HBPE error, implying that all cues considered in this work are beneficial.

We also compare our approach with the state-of-the-art for joint HBPE [8]. It is worth noting that other recent methods [5,7,36] operating on a low resolution setting only consider head pose and do not estimate body pose. Evidently, the algorithm from [8] performs similar to the AUX + SS + H/B setting in our approach, as both these methods focus on coupled learning of head and body pose of individuals. However, our algorithm performs significantly better than [8] 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.

F-formation estimation. The benefit of our joint learning framework on F-formation estimation can be noted from Table 2. Similar to HBPE experiments, using unlabeled samples from the social scene in addition to auxiliary data is beneficial for F-formation estimation. This is consistent with our expectation that accurate estimation of the body pose of interacting targets can aid detection of conversational groups. Incorporating additional information such as H/B coupling, temporal consistency and occlusion-specific classifiers in our framework further raises the F1-score.

In order to conceive the best F-formation detection performance using our method, we computed detection accuracy using target positions and the ground-truth head

Table 3: F-Formation estimation evaluation via precison (pre), recall (rec) and F1-scores (F1).

	CP			CB		
Method	pre	rec	F1	pre	rec	F1
IRPM [4]	0.67	0.65	0.66	0.68	0.50	0.57
IGD [32]	0.81	0.61	0.70	0.69	0.65	0.67
HVFF lin [12]	0.59	0.74	0.65	0.73	0.86	0.79
HVFF ent [28]	0.78	0.83	0.80	0.81	0.78	0.79
HVFF ms [29]	0.81	0.81	0.81	0.76	0.86	0.81
Game-Th. [33]	0.86	0.82	0.84	0.83	0.89	**0.86**
Our method	0.87	0.83	**0.85**	0.84	0.88	**0.86**

and body pose labels (Fig.4 (left)). A significant increase in F-formation detection performance is observed when pose cues are used along with positional information, and maximum performance is achieved with position and body pose cues, which confirms our intuition that body pose is more important than head pose for detecting F-formations. A comparison with state-of-the-art F-formation estimation approaches for the two datasets is presented in Table 3. These include frustrum-based (IRPM [4], IGD [32]), Hough transforms-based (HVFF lin [12], HVFF ent [28], HVFF ms [29]) and Game-theoretic [33] methods. We obtain F1-scores of 0.85 and 0.86 on CP and CB, thereby achieving best performance on CP and state-of-the-art results on CB. 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 among different runs.

Finally, we examine the effect of the clustering parameter γ_c on F-formation detection performance (Fig.4 (right)). Low values of γ_c preclude clustering of target positions and result in only singleton groups being discovered, thereby implying low F1-scores. Conversely, large γ_c values result in multiple F-formations to merge as all O-space centroids are constrained to be close to each other in such cases (see (Eqn.6)), which again adversely impacts detection performance. Interestingly, for both datasets, γ_c values in the range $[0.2, 0.4]$ correspond to the best performance.

Qualitative Results. Fig.5 depicts some qualitative results associated with our method on the CP dataset. Specifically we compare the inferred body poses and F-formations with ground truth annotations. Fig.5(top) shows one case where body pose is mostly accurately estimated and one conversational groups is correctly detected. Fig.5(center-bottom) depicts two challenging situations where our method fails. In Fig.5(center) one subject is close to a conversational group with other three targets and his pose is wrongly estimated, leading to a incorrect F-formation detection. In Fig.5(bottom), despite the body pose of all the targets is correctly estimated, our algorithm is not able to detect two conversational groups. This leaves room for further improving our method, *e.g.*by adopting a multiscale approach or considering a time-varying parameter γ_c.

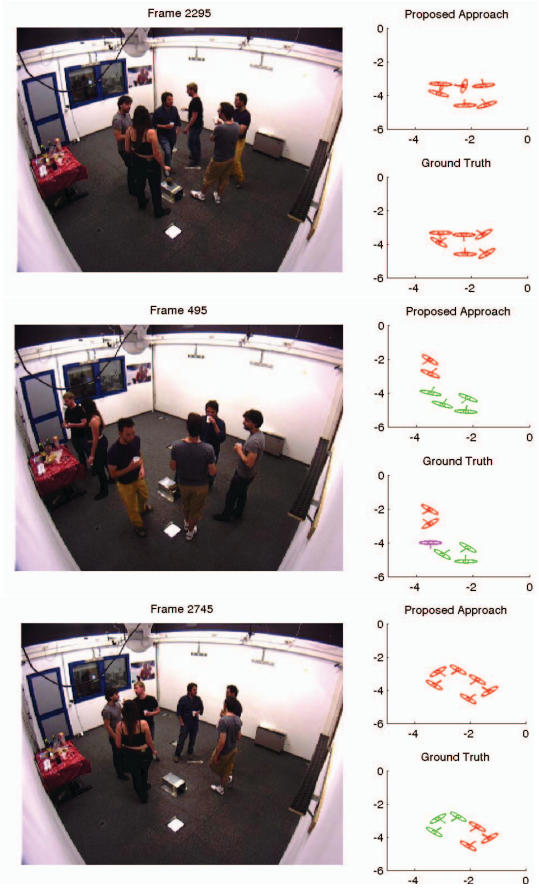

Figure 5: CP dataset: qualitative results. (left) Original frames. (right) Estimates body pose and F-formations compared with ground truth annotations. Ellipses indicate targets, arrows emerging from the ellipses indicate body pose. Members of the same group are shown in the same color.

5. Conclusions

We present a novel approach to jointly learn head, body pose of targets and F-formations from social scenes. Our algorithm uniquely exploits the interaction-interactor relationship in terms of positional and pose cues to infer the above from low-resolution and crowded scenes involving extreme occlusions. Joint learning improves both pose and F-formation estimation accuracy, and we outperform the state-of-the-art on two social datasets upon incorporating information concerning occlusions and temporal consistency. Future work involves extending the current methodology to multi-view settings, and incorporating multimodal cues obtained from wearable sensors (*e.g.*, infra-red) [1,2].

Acknowledgments: 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.

References

[1] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *arXiv preprint arXiv:1506.06882*, 2015.

[2] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe. Analyzing free-standing conversational groups: a multimodal approach. *ACM MM*, 2015.

[3] S. Alletto, G. Serra, S. Calderara, F. Solera, and R. Cucchiara. From ego to nos-vision: Detecting social relationships in first-person views. In *CVPRW*, 2014.

[4] L. Bazzani, D. Tosato, M. Cristani, M. Farenzena, G. Pagetti, G. Menegaz, and V. Murino. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 2013.

[5] B. Benfold and I. Reid. Unsupervised learning of a scene-specific coarse gaze estimator. In *ICCV*, 2011.

[6] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE TPAMI*, 33(9):1806–1819, 2011.

[7] I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, and A. Sugimoto. Head direction estimation from low resolution images with scene adaptation. *CVIU*, 117(10):1502–1511, 2013.

[8] C. Chen and J.-M. Odobez. We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video. In *CVPR*, 2012.

[9] E. C. Chi and K. Lange. Splitting methods for convex clustering. *arXiv preprint arXiv:1304.0499*, 2013.

[10] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese. Discovering groups of people in images. In *ECCV*, 2014.

[11] T. Ciolek and A. Kendon. Environment and the spatial arrangement of conversational encounters. sociological inquiry. *Sociological Inquiry*, 1980.

[12] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of F-formations. In *BMVC*, 2011.

[13] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel. Probabilistic temporal head pose estimation using a hierarchical graphical model. In *ECCV*. 2014.

[14] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *ECCV*, 2010.

[15] X. Geng and Y. Xia. Head pose estimation based on multivariate label distribution. In *CVPR*, 2014.

[16] A. Heili, J. Varadarajan, B. Ghanem, N. Ahuja, and J.-M. Odobez. Improving head and body pose estimation through semi-supervised manifold alignment. In *ICIP*, 2014.

[17] T. D. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath an algorithm for clustering using convex fusion penalties. In *ICML*, 2011.

[18] T. Hu, S. Messelodi, and O. Lanz. Dynamic task decomposition for decentralized object tracking in complex scenes. *CVIU*, 134:89–104, 2015.

[19] N. Krahnstoever, M.-C. Chang, and W. Ge. Gaze and body pose estimation from a distance. In *AVSS*, 2011.

[20] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese. Learning an image-based motion context for multiple people tracking. In *CVPR*, 2014.

[21] M. C. Liem and D. M. Gavrila. Coupled person orientation estimation and appearance modeling using spherical harmonics. *IVC*, 32(10):728–738, 2014.

[22] M. Marin-Jimenez, A. Zisserman, M. Eichner, and V. Ferrari. Detecting people looking at each other in videos. *IJCV*, 106(3):282–296, 2014.

[23] M. Mathias, R. Benenson, R. Timofte, and L. V. Gool. Handling occlusions with franken-classifiers. In *ICCV*, 2013.

[24] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. Structured learning of human interactions in tv shows. *IEEE TPAMI*, 34(12):2441–2453, 2012.

[25] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *ECCV*, 2010.

[26] A. K. Rajagopal, R. Subramanian, E. Ricci, R. L. Vieriu, O. Lanz, and N. Sebe. Exploring transfer learning approaches for head pose classification from multi-view surveillance images. *IJCV*, 109(1-2):146–167, 2014.

[27] N. Robertson and I. Reid. Estimating gaze direction from low-resolution faces in video. In *ECCV*, 2006.

[28] F. Setti, H. Hung, and M. Cristani. Group detection in still images by F-formation modeling: A comparative study. In *WIAMIS*, 2013.

[29] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani. Multi-scale F-formation discovery for group detection. In *ICIP*, 2013.

[30] D. Z. T. Gan, Y. Wong and M. Kankanhalli. Temporal encoded F-formation system for social interaction detection. In *WACM MM*, 2013.

[31] S. Tang, M. Andriluka, and B. Schiele. Detection and tracking of occluded people. *IJCV*, pages 1–12, 2012.

[32] K. N. Tran, A. Bedagkar-Gala, I. A. Kakadiaris, and S. K. Shah. Social cues in group formation and local interactions for collective activity analysis. In *VISAPP*, 2013.

[33] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino. A game theoretic probabilistic approach for detecting conversational groups. In *ACCV*, 2014.

[34] B. Wang, G. Wang, K. L. Chan, and L. Wang. Tracklet association with online target-specific metric learning. In *CVPR*, 2014.

[35] C. Wojek, S. Walk, S. Roth, and B. Schiele. Monocular 3d scene understanding with explicit occlusion reasoning. In *CVPR*, 2011.

[36] Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, and N. Sebe. A multi-task learning framework for head pose estimation under target motion. *IEEE TPAMI*, 2015.

[37] G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space speaks: towards socially and personality aware visual surveillance. In *MPVA*. ACM, 2010.

[38] X. Zhu and A. B. Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.