

# Analyzing Free-standing Conversational Groups: A Multimodal Approach

Xavier Alameda-Pineda  
University of Trento  
Trento, Italy  
xavier.alamedapineda@unitn.it

Yan Yan  
Univ. of Trento / ADSC, UIUC  
Trento, Italy / Singapore  
yan.yan@unitn.it

Elisa Ricci  
FBK / Univ. of Perugia  
Trento, Italy / Perugia, Italy  
eliricci@fbk.eu

Oswald Lanz  
Fondazione Bruno Kessler  
Trento, Italy  
lanz@fbk.eu

Nicu Sebe  
University of Trento  
Trento, Italy  
sebe@disi.unitn.it

## ABSTRACT

During natural social gatherings, humans tend to organize themselves in the so-called free-standing conversational groups. In this context, robust head and body pose estimates can facilitate the higher-level description of the ongoing interplay. Importantly, visual information typically obtained with a distributed camera network might not suffice to achieve the robustness sought. In this line of thought, recent advances in wearable sensing technology open the door to multimodal and richer information flows. In this paper we propose to cast the head and body pose estimation problem into a matrix completion task. We introduce a framework able to fuse multimodal data emanating from a combination of distributed and wearable sensors, taking into account the temporal consistency, the head/body coupling and the noise inherent to the scenario. We report results on the novel and challenging SALSA dataset, containing visual, auditory and infrared recordings of 18 people interacting in a regular indoor environment. We demonstrate the soundness of the proposed method and the usability for higher-level tasks such as the detection of F-formations and the discovery of social attention attractors.

## 1. INTRODUCTION

Several wearable sensing devices became available for the general public in the past few years. Often, such consumer platforms include several sensors, *e.g.* microphone, accelerometer or blue-tooth. Investigating methods to process the data gathered with wearable devices is worthwhile for many reasons. First of all, considering that modern consumer platforms, such as smart-phones or smart-watches, are at the reach of the general public, many are already using them for data acquisition. Second, these platforms are



**Figure 1: People interacting naturally organized in free-standing conversational groups during a poster session. The social interplay is captured by a distributed camera network and by sociometric badges worn by each participant (small white boxes).**

inherently multimodal, thus they provide a rich description of the environment. Third, since these devices are easy-to-wear, the continuous flow of information emanating from them is a precious resource to study casual situations of the real life, far from laboratory-controlled conditions.

Importantly, wearable technologies are complementary to distributed sensor networks. For example, when analyzing social interactions, wearable sensing devices can be exploited *inter alia* to localize people and to roughly estimate their activities. However, a fine-grained analysis of the social scene requires additional information gathered with alternative distributed sensing networks, *e.g.* visual data [29]. Developing methods that robustly fuse data from wearable devices and distributed networks remains largely unexplored, and many challenges arise in this context. As with any consumer device, data gathered with wearable technology are corrupted by severe noise. Therefore, unimodal approaches fail to provide a robust and accurate representation of the environment and smart multimodal fusion strategies need to be developed [27]. Moreover, these strategies should also account for the specificities of the combination of a distributed sensor network and the wearable technology.

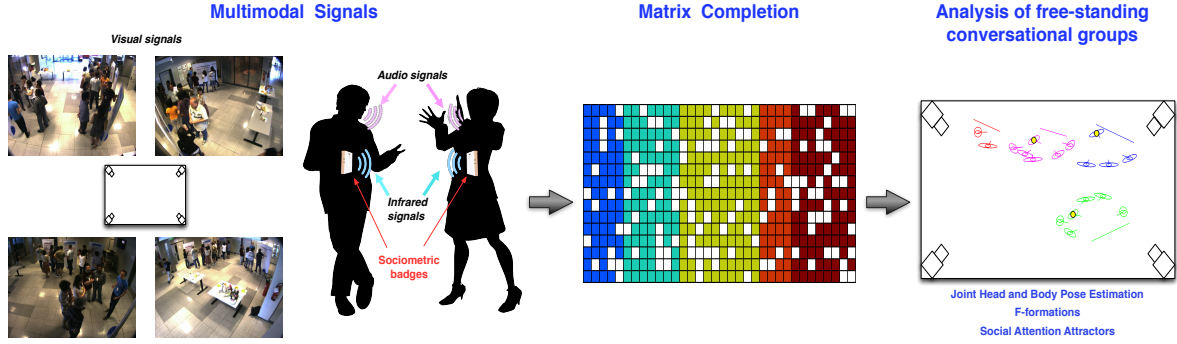
The focus of this research study is the analysis of spontaneous social interactions in natural indoor environments, see Figure 1. In such social events, human beings tend to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [Permissions@acm.org](mailto:Permissions@acm.org).

MM'15, October 26–30, 2015, Brisbane, Australia.

© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...\$15.00.

DOI: <http://dx.doi.org/10.1145/2733373.2806238>.



**Figure 2: Overview of the proposed approach for detection of attentional patterns.** The scene is captured by a distributed camera network and sociometric badges worn by each participant. The estimation of the head and body poses, required to further extract F-formations and social attention attractors, is cast into a matrix completion problem. The HBPE are used to provide a higher-level description of the social interplay.

organize themselves in free-standing conversational groups (FCGs). Contrary to the staticity proper to round-table meetings, FCGs are dynamic by nature (they increase and decrease in size, move and split) and therefore their analysis is inherently difficult [11, 33]. In this context, robust human pose estimates can ease the further mining of higher-level descriptions, such as F-formations or attentional patterns. In this paper we are expressly interested in analyzing FCGs using head and body pose estimates obtained through the processing of the multimodal flow of data emanating from a distributed camera network and sociometric badges [10] worn by each participant.

Providing a complete description of the human pose (positions of the limbs) is chimeric in video surveillance settings, *i.e.* when people are monitored with distant and large field-of-view cameras. Indeed, the low-resolution images and the numerous occlusions have a relentless negative effect on current approaches. Fortunately, the head and body pose can be used as a surrogate of the full pose when analyzing free-standing conversational groups. Indeed, *all* features delineating group social interplays, are often extracted from head and body pose estimates (HBPE). More precisely, accurate and robust HBPEs are the prelude to detect face-to-face interactions, identify the potential addressee and addresser, describe the geometric structure of the group (or F-formation) and recognize personality traits. Many research studies focused on HBPE from visual data in the recent past, and demonstrated the synergistic effect of joint head and body pose estimation, when compared to independent estimation [9]. Furthermore, as previously discussed, wearable sensing devices provide a complementary source of information to the surveillance distributed camera network. Nowadays, up to the authors’ knowledge, there are no existing research studies that address the challenging task of head and body pose estimation in crowded indoor environments using multimodal data gathered with a combination of distributed visual sensors and wearable devices.

In this paper we propose to integrate multimodal signals acquired in regular indoor environments to robustly extract HBPE of several people naturally arranged in FCG. More precisely, we use auditory and proximity (infrared) information to automatically label visual features of head and body respectively. All this information, together with the visual features extracted from the surveillance camera network, are pulled together into a linear classification problem. Seeking for the non-available labels, *i.e.* estimating the head

and body pose of all people, is cast into a matrix completion problem. Within this framework, the joint head and body pose estimation adds an additional coupling term to the formulation. Similarly, the temporal structure inherent to the problem is induced by means of a Laplacian matrix that regularizes the matrix completion task. The resulting HBPE are then used to infer the F-formations and to further retrieve the social attention attractors. Figure 2 shows an overview of the proposed approach.

The main contributions of this study are: (i) address multimodal HBPE within the analysis of FCGs (ii) propose a framework versatile enough to handle not only data coming from different modalities, but also to process features extracted with both wearable sensors and a distributed camera network (iii) cast HBPE into a matrix completion problem, accounting for the noisy multimodal labels, and the inherent temporal structure of the head and body poses (iv) induce joint HBPE by coupling the two incomplete matrices (v) derive a novel optimization approach to solve for the matrix completion task and (vi) conduct an extensive evaluation of the proposed method on the novel SALSA dataset, showing the usability of the HBPE for detecting F-formations and discovering social attention attractors.

The rest of the paper is structured as follows. Related approaches for head and body pose estimation are discussed in the next section. The proposed matrix completion model, accounting for noise, temporal regularization and joint head and body pose estimation is presented in Section 3, together with the associated optimization algorithm. Extensive evaluations of the proposed approach, and comparisons with the state of the art are reported in Section 4, before concluding and delineating future research directions in Section 5.

## 2. RELATED WORK

The applicative context of the present study is the multimodal analysis of social interactions in informal gatherings. Special focus is given to the estimation of the head and body pose, easing the automatic study of free-standing conversational groups. Moreover, the proposed solution is based on a matrix completion framework.

### 2.1 Multimodal analysis of social interactions

Combining data from heterogeneous modalities is of utmost importance for understanding human behaviors [36]. Approaches relying on audio-visual cues are probably the most popular and successful examples [18, 30, 37, 1]. In

[18], the role of non-verbal cues for the automatic analysis of face-to-face social interactions is investigated. Social interactions in political debates are studied in [37]. A multimodal approach for detecting laughter episodes from audio-visual data and a novel dataset for studying this problem are introduced in [30]. In the last few years, mobile and wearable devices opened novel opportunities for multimodal analysis of social interactions [6, 16, 15, 28]. In [15] a probabilistic approach is proposed to automatically discover interaction types from large-scale dyadic data (*e.g.* proximity blue-tooth data, phone call network or email network). In [28] social interactions on a small spatio-temporal scale combining proximity data and accelerometers are analyzed.

On the one hand, traditional distributed camera and microphone networks are able to capture subtle features of human behavior, but their performance drops significantly when dealing with a crowded scene. On the other hand, wearable devices permit the ubiquitous localization of people for a long time span, as they typically embed proximity sensors, but they fail to provide accurate and detailed descriptions of the ongoing interplay. Therefore, it is obvious that the nuances and the complexity of social interactions require leveraging information from both wearable and traditional sensors. This paper develops from this intuition and, to our knowledge, it is the first work jointly employing sociometric badges and external cameras for HBPE.

## 2.2 Head and body pose estimation

As already outlined, the analysis of the human behavior will definitely benefit from automatic human pose estimation. Multimodal approaches based on RGB-D data [35, 42], eventually combined with audio [17], and with visual-inertial sensors [13], have recently proved very successful in tracking human limbs. However, the former methods are appropriate when only one or few people move in proximity of the camera, while the latter approaches are often not practical solutions, as they require to instrument the person's body with multiple sensors. In addition, the robust estimation of the full human pose is chimeric in crowded scenes, which is typically the case of informal social gatherings. Consequently, when considering social interactions among several people, the head and body pose can be used as a surrogate of the full human pose.

Previous research has demonstrated that head and body orientations can be successfully detected from visual data [3, 9, 41, 31] and used as primary cues to infer high level information, such as visual focus of attention [38] and conversational groups [11]. In order to alleviate the human annotation effort, previous research has focused on using related cues to estimate head and body pose. For example, head pose labels generated using walking direction are considered by Benfold and Reid [3]. Similarly, coupling of head and body pose due to anatomical constraints is enforced by Chen and Odobez [9]. The body orientation is considered as a link between walking direction and head pose in [32, 25]. A transfer learning framework for head pose estimation under target motion is proposed in [31]. In [41] a multi-task learning approach is introduced for accurate estimates of the head pose from large field-of-view surveillance cameras. More recently, approaches to specifically cope with label noise [19] and integrate temporal consistency [14] are introduced. Audio-visual head pose estimation is attempted in [8]. However, up to the authors' knowledge, there are no

research studies combining visual features from a distributed camera network with multimodal cues extracted from wearable sensors to jointly estimate the head and body poses.

## 2.3 Matrix Completion

We propose to cast multimodal head and body pose estimation into a matrix completion problem, which has been shown to be equivalent to learning a classifier in a transductive setting [21]. This formulation is particularly advantageous when data and labels are noisy or in the case of missing data. In the computer vision and multimedia communities, this fact has been exploited in several applications, such as multi-label image classification [5], image retrieval [40] and facial expression analysis [39]. A recent study [24] extends the matrix completion problem to take into account an underlying graph structure inducing a weighted relationship between the columns and between the rows of the matrix. We also utilize this structure to model the temporal smoothness of the head and body pose estimates. However, from the technical point of view our approach is novel because: (i) we address two matrix completion problems (for the head and body poses) that are coupled and cannot be reduced to a joint matrix completion problem, (ii) the proposed framework is able to deal with data coming from multiple modalities, handling signals generated from different types of sensing devices (distributed and wearable).

## 3. MATRIX COMPLETION FOR MULTIMODAL POSE ESTIMATION

The technical challenge of the present study is the robust estimation of the body and head pose in crowded scenarios for further analysis of FCGs. We assume a regular indoor environment populated with  $K$  people and captured with a distributed camera network and by sociometric badges [10] worn by the participants. On the one hand, we extract visual features of the head and body of all the participants, localized in the scene thanks to a multi-target visual tracking algorithm. On the other hand, since manual annotation is tedious and resource-demanding, we automatically label part of the sequence using the information gathered with the sociometric badges. The full preprocessing pipeline is described in Section 4.2, but we outline here the main insights and notations used in our research.

The visual descriptors of the head and body of person  $k$  at time  $t$  are denoted by  $\mathbf{v}_{kt}^h \in \mathbb{R}^{d_h}$  and by  $\mathbf{v}_{kt}^b \in \mathbb{R}^{d_b}$  respectively, where  $d_h$  and  $d_b$  are the features' dimension. Similarly, the head and body orientations of person  $k$  at time  $t$  are denoted by  $\theta_{kt}^h$  and  $\theta_{kt}^b$ . Inspired by previous research on the topic [9], we discretize the space of angular orientations into  $C$  sectors. Therefore the body and head orientations are  $C$ -dimensional vectors.  $(\theta_{kt}^b)_c$  should be understood as the probability that the associated body visual feature lies in the  $c^{\text{th}}$  sector (analogously for the head). For instance,  $\theta_{kt}^b = [1, 0, \dots, 0]$ , means that  $\theta_{kt}^b$  belongs to the first sector.

As outlined before, we assume the existence of a labeled training set  $\mathcal{L}$  containing the raw features and noisy labels of the head and body pose of all the  $K$  persons involved in the interaction for  $T_0 < T$  frames, where  $T$  is the total number of frames. That is:  $\mathcal{L} = \{\mathbf{v}_{kt}^b, \theta_{kt}^b, \mathbf{v}_{kt}^h, \theta_{kt}^h\}_{k=1, t=T_0+1}^{K, T}$ . Complementary, the set of unlabeled features is defined as:  $\mathcal{U} = \{\mathbf{v}_{kt}^b, \mathbf{v}_{kt}^h\}_{k=1, t=T_0+1}^{K, T}$ . It is important to remark that part of the training set is manually annotated and the rest

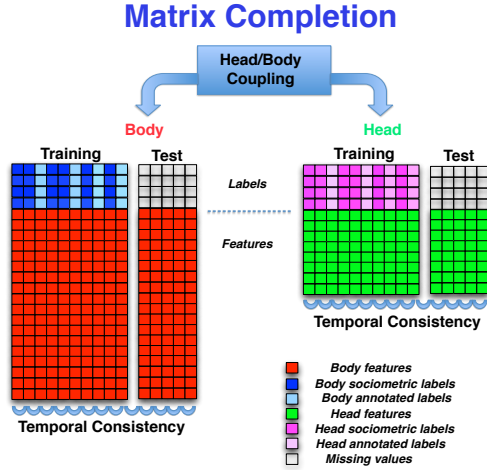


Figure 3: Illustration of the proposed matrix completion framework for multimodal head and body pose estimation (MC-HBPE). Two matrix completion problems, for the head and body poses, are regularized to account for the temporal consistency and for head/body coupling.

is automatically labeled. More precisely, we use the infrared detections as a proxy for the body pose and auditory correlations to automatically label the head pose. Indeed, it is reasonable to assume that the audio signals have a strong influence on the gaze direction of the targets, while the infrared detections naturally correlate with the body orientation. Finally, it is worth noticing that while visual data are continuously available, audio and infrared signals are sparse observations. In the following we describe the proposed approach for fusing these multiple modalities, so to robustly estimate the head and body poses of all participants.

### 3.1 The Model

Recent theoretical results and practical advances in matrix completion theory [5, 21], motivated us to consider the head and body problem from this perspective. We thoroughly cast the estimation into a matrix completion problem, thus introducing Matrix Completion for Head and Body Pose Estimation (MC-HBPE), a graphical illustration of which is shown in Figure 3. In the following, in order to ease the exposition, all the claims, definitions and properties are stated only for the body pose, but they remain true for the head pose, unless it is explicitly written otherwise. Before describing the model, we set a few useful notations: the matrices of labeled and unlabeled body features of person  $k$ ,  $\mathbf{V}_{\mathcal{L},k}^b = [\mathbf{v}_{kt}^b]_{t=1}^{T_0}$  and  $\mathbf{V}_{\mathcal{U},k}^b = [\mathbf{v}_{kt}^b]_{t=T_0+1}^T$ ; their concatenation,  $\mathbf{V}_k = [\mathbf{V}_{\mathcal{L},k}^b \mathbf{V}_{\mathcal{U},k}^b] = [\mathbf{v}_{kt}^b]_{t=1}^T$ ; and the concatenation over all persons, *i.e.*, the concatenation of all body visual features,  $\mathbf{V}^b = [\mathbf{V}_k^b]_{k=1}^K$ . The matrices  $\Theta_{\mathcal{L},k}^b$ ,  $\Theta_{\mathcal{U},k}^b$ ,  $\Theta_k^b$  and  $\Theta^b$  are analogously built from  $\theta_{kt}^b$ .

The objective is to estimate the matrix  $\Theta_{\mathcal{U}} = [\Theta_{\mathcal{U},k}^b]_{k=1}^K$ , under the assumption of a linear classifier:

$$\Theta^b = \mathbf{W}^b \begin{bmatrix} \mathbf{V}^b \\ \mathbf{1}^\top \end{bmatrix}, \quad (1)$$

with  $\mathbf{W}^b \in \mathbb{R}^{C \times (d_b+1)}$ . Remarkably, the joint feature-label matrix  $\mathbf{J}_b = [\Theta^b \mathbf{V}^b \mathbf{1}^\top]^\top \in \mathbb{R}^{(C+d_b+1) \times KT}$  is low rank, since the linear classifier in (1) imposes a linear dependency between the rows of  $\mathbf{J}_b$ . Therefore, we set the following

optimization problem for  $\Theta_{\mathcal{U}}^b$ :

$$\Theta_{\mathcal{U}}^{b*} = \underset{\Theta_{\mathcal{U}}^b}{\operatorname{argmin}} \operatorname{rank}(\mathbf{J}_b). \quad (2)$$

Minimizing the rank of a matrix is a NP-hard problem, but it can be exactly relaxed by means of the nuclear norm [7]. Indeed, since the nuclear norm  $\|\cdot\|_*$  is the tightest convex envelope of the rank, the previous optimization problem is equivalent to:

$$\Theta_{\mathcal{U}}^{b*} = \underset{\Theta_{\mathcal{U}}^b}{\operatorname{argmin}} \|\mathbf{J}_b\|_*. \quad (3)$$

In real applications both the observations and the training labels are noisy. Therefore, it is of crucial importance to account for the observational and label noise. In practice, we assume  $\tilde{\mathbf{J}}_b = \mathbf{J}_b + \mathbf{E}_b$ , where  $\tilde{\mathbf{J}}_b$  represents the noisy features and labels, and  $\mathbf{E}_b$  represents the noise. The objective is to estimate the low-rank matrix  $\mathbf{J}_b$  that best approximates the observed matrix  $\tilde{\mathbf{J}}_b$ . This is done by constraining (3), and relaxing it into [21]:

$$\min_{\mathbf{J}_b} \nu_b \|\mathbf{J}_b\|_* + \frac{\lambda_b}{2} \left\| P_{\mathcal{O}}^b (\tilde{\mathbf{J}}_b - \mathbf{J}_b) \right\|_{\mathcal{F}}^2, \quad (4)$$

where  $\nu_b$  and  $\lambda_b$  are regularization parameters,  $\|\mathbf{X}\|_{\mathcal{F}}^2 = \operatorname{Tr}(\mathbf{X}^\top \mathbf{X})$  denotes the Frobenius norm of  $\mathbf{X}$  and  $P_{\mathcal{O}}^b$  denotes the projection onto the “set of observations”, defined as  $P_{\mathcal{O}}^b \left( [\Theta^b, \mathbf{V}^b, \mathbf{1}^\top]^\top \right) = [\Theta_{\mathcal{L},k}^b, \mathbf{0}]_{k=1}^K \mathbf{V}^b, \mathbf{1}^\top]^\top$ , and thus setting to zero all elements associated to  $\Theta_{\mathcal{U}}^b$ . We remark that this *unitary* regularization term, encompasses the error on the visual features as well as on the training labels.

One of the prominent factors alleviating the matrix completion problem with noisy observations is the inherent temporal structure of the problem. Intuitively, we would expect the labels to be *smooth* in time, that is  $\theta_{kt}^b \approx \theta_{kt+1}^b$ ,  $\forall k, t$ . We reinforce this intuition through a set of *binary* regularization terms, grouped in the following loss function:

$$\mathcal{L}_b(\mathbf{J}_b) = \frac{1}{2} \operatorname{Tr} \left( P_{\Theta}^b(\mathbf{J}_b) \mathbf{T}_b P_{\Theta}^b(\mathbf{J}_b)^\top \right), \quad (5)$$

where  $P_{\Theta}^b$  is the projection onto the labels,  $\Theta_b$ , defined as  $P_{\Theta}^b \left( [\Theta^b, \mathbf{V}^b, \mathbf{1}^\top]^\top \right) = [\Theta^b, \mathbf{0}]^\top$ , and  $\mathbf{T}_b \in \mathbb{R}^{KT \times KT}$  is the Laplacian matrix associated to the graph encoding the relations between the variables. In our case,  $\mathbf{T}_b = \mathbf{I}_K \otimes \mathbf{L}$ , where  $\mathbf{I}_K$  is the  $K$ -dimensional identity matrix,  $\otimes$  is the Kronecker product and  $\mathbf{L} \in \mathbb{R}^{T \times T}$  is a tri-diagonal matrix defined as follows: the elements of the two subdiagonals are set to  $-1$ , and the elements of the main diagonal are set to  $2$ , except  $\mathbf{L}_{11} = \mathbf{L}_{TT} = 1$ . In practice this means that only contiguous pose labels of the same person are embolden to be equal.

Until this point, we discussed the problem of body and head pose estimation separately. However, previous research studies in head and body pose estimation demonstrated the synergistic effect of joint estimation [9]. Subsequently, the estimation of head and body pose benefits from a coupling structure, in addition to the temporal structure already discussed. We write,  $\theta_{kt}^b \approx \theta_{kt}^h$  and formalize the coupling structure by means of the following regularization term:

$$\mathcal{C}(\mathbf{J}_b, \mathbf{J}_h) = \frac{1}{2} \left\| P_{\Theta}^b(\mathbf{J}_b) - P_{\Theta}^h(\mathbf{J}_h) \right\|_{\mathcal{F}}^2, \quad (6)$$

where  $P_{\Theta}^b$  is defined analogously<sup>1</sup> as  $P_{\Theta}^h$ .

Summarizing, the head and body pose estimation is cast into a matrix completion problem encompassing the temporal and coupling structural constraints, by considering the following optimization problem:

$$\begin{aligned} \min_{\mathbf{J}_b, \mathbf{J}_h} \quad & \nu_b \|\mathbf{J}_b\|_* + \frac{\lambda_b}{2} \left\| P_{\Theta}^b (\tilde{\mathbf{J}}_b - \mathbf{J}_b) \right\|_{\mathcal{F}}^2 \\ & + \nu_h \|\mathbf{J}_h\|_* + \frac{\lambda_h}{2} \left\| P_{\Theta}^h (\tilde{\mathbf{J}}_h - \mathbf{J}_h) \right\|_{\mathcal{F}}^2 \\ & + \frac{\tau_b}{2} \text{Tr} \left( P_{\Theta}^b (\mathbf{J}_b)^{\top} \mathbf{T}_b P_{\Theta}^b (\mathbf{J}_b) \right) + \frac{\tau_h}{2} \text{Tr} \left( P_{\Theta}^h (\mathbf{J}_h)^{\top} \mathbf{T}_h P_{\Theta}^h (\mathbf{J}_h) \right) \\ & + \frac{\lambda_c}{2} \left\| P_{\Theta}^b (\mathbf{J}_b) - P_{\Theta}^h (\mathbf{J}_h) \right\|_{\mathcal{F}}^2. \end{aligned} \quad (7)$$

This is a non-linear convex optimization problem whose critical points cannot be expressed in closed-form solution. In the following we describe the proposed optimization algorithm to find the optimal solution for (7).

### 3.2 Optimization method

The alternating-direction method of multipliers (ADMM) is a well-known optimization procedure, and a review of theoretical results, practical implementation considerations and applications can be found in [4]. Recently, an ADMM-based optimization procedure for the matrix completion problem has been presented [24]. We also considered ADMM to solve the matrix completion problem. However, the proposed method is intrinsically different from previous approaches because of two reasons:

- First, we consider a coupled matrix completion problem, which is different from solving for the concatenated head/body matrix. Absolutely, completing the concatenation matrix implicitly assumes, not only that the body and head visual features have the same dimension, but also that the linear classifier for the body and head poses are equal  $\mathbf{W}^b = \mathbf{W}^h$ . Therefore, the method developed in [24] cannot be used in our case. Instead, we propose and develop a *Coupled Alternating Direction Method of Multipliers (C-ADMM)*.
- Second, the temporal structure of the problem is modeled with a column-wise regularizer on the matrices, which is a particular case of [24]. Importantly, as will be shown in the following, column-wise regularization can be solved in closed-form, while the formulation of [24] cannot. Consequently, all the steps of the proposed C-ADMM have a closed-form solution, and the convergence is guaranteed.

Classically, the ADMM arises from the construction of the *augmented Lagrangian* [4], which in our case writes:

$$\begin{aligned} \mathfrak{L} = \quad & \nu_b \|\mathbf{J}_b\|_* + \frac{\lambda_b}{2} \left\| P_{\Theta}^b (\tilde{\mathbf{J}}_b - \mathbf{K}_b) \right\|_{\mathcal{F}}^2 \\ & + \nu_h \|\mathbf{J}_h\|_* + \frac{\lambda_h}{2} \left\| P_{\Theta}^h (\tilde{\mathbf{J}}_h - \mathbf{K}_h) \right\|_{\mathcal{F}}^2 \\ & + \frac{\tau_b}{2} \text{Tr} \left( P_{\Theta}^b (\mathbf{K}_b)^{\top} \mathbf{T}_b P_{\Theta}^b (\mathbf{K}_b) \right) + \frac{\tau_h}{2} \text{Tr} \left( P_{\Theta}^h (\mathbf{K}_h)^{\top} \mathbf{T}_h P_{\Theta}^h (\mathbf{K}_h) \right) \\ & + \frac{\lambda_c}{2} \left\| P_{\Theta}^b (\mathbf{K}_b) - P_{\Theta}^h (\mathbf{K}_h) \right\|_{\mathcal{F}}^2 + \frac{\phi_b}{2} \|\mathbf{K}_b - \mathbf{J}_b\|_{\mathcal{F}}^2 + \frac{\phi_h}{2} \|\mathbf{K}_h - \mathbf{J}_h\|_{\mathcal{F}}^2 \\ & + \langle \mathbf{M}_b, \mathbf{J}_b - \mathbf{K}_b \rangle + \langle \mathbf{M}_h, \mathbf{J}_h - \mathbf{K}_h \rangle. \end{aligned} \quad (8)$$

<sup>1</sup>Even if  $P_{\Theta}^b$  and  $P_{\Theta}^h$  are conceptually equivalent, they are formally different since the features' dimension is different in general,  $d_b \neq d_h$ .

The Lagrangian  $\mathfrak{L}$  has to be minimized with respect to  $\mathbf{J}_b$ ,  $\mathbf{J}_h$ ,  $\mathbf{K}_b$ ,  $\mathbf{K}_h$ ,  $\mathbf{M}_b$  and  $\mathbf{M}_h$ , where the last two are matrices of Lagrangian multipliers,  $\mathbf{K}_b$  and  $\mathbf{K}_h$  are auxiliary variables, and  $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{ij} \mathbf{A}_{ij} \mathbf{B}_{ij}$  is the scalar product in the matrix space.

At iteration  $r + 1$ , the C-ADMM updates are:

$$(\mathbf{J}_b^{r+1}, \mathbf{J}_h^{r+1}) = \arg \min_{\mathbf{J}_b, \mathbf{J}_h} \mathfrak{L}(\mathbf{J}_b, \mathbf{J}_h, \mathbf{K}_b^r, \mathbf{K}_h^r, \mathbf{M}_b^r, \mathbf{M}_h^r), \quad (9)$$

$$(\mathbf{K}_b^{r+1}, \mathbf{K}_h^{r+1}) = \arg \min_{\mathbf{K}_b, \mathbf{K}_h} \mathfrak{L}(\mathbf{J}_b^{r+1}, \mathbf{J}_h^{r+1}, \mathbf{K}_b, \mathbf{K}_h, \mathbf{M}_b^r, \mathbf{M}_h^r), \quad (10)$$

$$\mathbf{M}_b^{r+1} = \mathbf{M}_b^r + \phi_b (\mathbf{J}_b^{r+1} - \mathbf{K}_b^{r+1}), \quad (11)$$

$$\mathbf{M}_h^{r+1} = \mathbf{M}_h^r + \phi_h (\mathbf{J}_h^{r+1} - \mathbf{K}_h^{r+1}). \quad (12)$$

From (8) it is straightforward that the coupling term does not have any effect on (9), but only on (10). Therefore, solving for (9) the two matrices  $\mathbf{J}_b^{r+1}$   $\mathbf{J}_h^{r+1}$  can be computed independently. Furthermore, the optimal solution of (9) is:

$$\mathbf{J}_b^{r+1} = \mathbf{U}_b S_{\frac{\nu_b}{\phi_b}}(\mathbf{D}_b) \mathbf{V}_b \quad (13)$$

where  $\mathbf{U}_b \mathbf{D}_b \mathbf{V}_b$  is the singular-value decomposition of the matrix  $\mathbf{K}_b^r - \frac{1}{\phi_b} \mathbf{M}_b^r$  and  $S_{\lambda}(x) = \max(x - \lambda, 0)$  is the shrinkage operator with constant  $\lambda$  applied element-wise to the diagonal matrix of singular values,  $\mathbf{D}_b$ .  $\mathbf{J}_h^{r+1}$  is found analogously. The details of the derivation can be found in the supplementary material.

The second C-ADMM update, *i.e.* (10), is also solved in closed-form. The critical point is computed by canceling the derivative of the objective function in (10) with respect to  $\mathbf{K}_b$  and  $\mathbf{K}_h$ . As shown in the supplementary material, these derivatives lead to block-diagonal linear systems that can be efficiently solved. In order to sketch the proof, we define  $\mathbf{k}_b = \text{vec}(\mathbf{K}_b)$  as the row-vectorization of  $\mathbf{K}_b$ . The same notation stands for all other matrices involved. Moreover, we define  $\mathbf{k}_{b, kc} = [(\theta_{kt}^b)_c]_{t=1}^T$ , as the  $T$ -dimensional vector composed of the coordinate  $c$  of the angle labels of person  $k$  over time. Likewise,  $\mathbf{k}_{b, v}$  stands for the  $KT(d_b + 1)$ -dimensional row vectorization of the rest of the matrix  $\mathbf{K}_b$ , that is the part corresponding to the visual features  $\mathbf{V}^b$ . With these notations, the vector  $\mathbf{k}_b$  rewrites:

$$\mathbf{k}_b = [\mathbf{k}_{b, 11}^{\top} \dots \mathbf{k}_{b, KC}^{\top} \mathbf{k}_{b, v}^{\top}]^{\top},$$

and analogously for all other row-vectorized matrices.

Each of the  $KC$  variables for body and head,  $\mathbf{k}_{b, kc}$  and  $\mathbf{k}_{h, kc}$ , are independently solved:

$$\mathbf{k}_{b, kc} = (\mathbf{L}_h \mathbf{L}_b - \mathbf{I}_T)^{-1} (\mathbf{L}_h \bar{\mathbf{k}}_{b, kc} + \bar{\mathbf{k}}_{h, kc}), \quad (14)$$

$$\mathbf{k}_{h, kc} = \mathbf{L}_b \mathbf{k}_{b, kc} - \bar{\mathbf{k}}_{h, kc}, \quad (15)$$

where all matrices and vectors are precisely defined in the supplementary material. The vector  $\bar{\mathbf{k}}_{b, kc}$  is computed from  $\tilde{\mathbf{J}}_{b, kc}^r$ ,  $\mathbf{m}_{b, kc}^r$  and  $\mathbf{j}_{b, kc}^{r+1}$  and thus depends on the iteration (analogously for the head). Importantly, the matrices  $\mathbf{L}_b$  and  $\mathbf{L}_h$ , computed from  $\mathbf{L}$  and the regularization parameters, do not depend on the iteration. Therefore, the inversion can be computed before the iterative procedure of the C-ADMM starts, leading to an efficient algorithm. The system involving  $\mathbf{k}_{b, v}$  and  $\mathbf{k}_{h, v}$  is simpler since there is no

---

**Algorithm 1** The coupled alternating-directions method of multipliers solving for the derived matrix completion for multimodal HBPE.

---

- 1: **Input:**  
Observation  $\widetilde{\mathbf{J}}_{\mathbf{h}}, \widetilde{\mathbf{J}}_{\mathbf{b}}$  and Laplacian  $\mathbf{T}_{\mathbf{h}}, \mathbf{T}_{\mathbf{b}}$  matrices.  
Regularization parameters.
  - 2: **Output:**  
Optimized  $\mathbf{J}_{\mathbf{h}}, \mathbf{J}_{\mathbf{b}}$ .
  - 3: Randomly initialize  $\mathbf{M}_{\mathbf{b}}, \mathbf{M}_{\mathbf{h}}, \mathbf{K}_{\mathbf{b}}, \mathbf{K}_{\mathbf{h}}$ ;
  - 4: **Repeat**  
Solve for  $\mathbf{J}_{\mathbf{h}}, \mathbf{J}_{\mathbf{b}}$  in (9) using (5);  
Solve for  $\mathbf{K}_{\mathbf{h}}, \mathbf{K}_{\mathbf{b}}$  in (10) using (14)-(17);  
Solve for  $\mathbf{M}_{\mathbf{h}}, \mathbf{M}_{\mathbf{b}}$  using (11)-(12);  
**Until Convergence**
- 

coupling between body and head features and there is no temporal regularization:

$$(\lambda_{\mathbf{b}} + \phi_{\mathbf{b}})\mathbf{k}_{\mathbf{b},\mathbf{v}} = \lambda_{\mathbf{b}}\widetilde{\mathbf{j}}_{\mathbf{b},\mathbf{v}} + \mathbf{m}_{\mathbf{b},\mathbf{v}}^r + \phi_{\mathbf{b}}\mathbf{j}_{\mathbf{b},\mathbf{v}}^{r+1} \quad (16)$$

$$(\lambda_{\mathbf{h}} + \phi_{\mathbf{h}})\mathbf{k}_{\mathbf{h},\mathbf{v}} = \lambda_{\mathbf{h}}\widetilde{\mathbf{j}}_{\mathbf{h},\mathbf{v}} + \mathbf{m}_{\mathbf{h},\mathbf{v}}^r + \phi_{\mathbf{h}}\mathbf{j}_{\mathbf{h},\mathbf{v}}^{r+1}. \quad (17)$$

Equations (14)-(17) provide the critical point sought in (10) and conclude the derivation of the proposed C-ADMM. As it is classically the case, the complexity bottle-neck of the matrix-completion solver is in the computation of the singular-value decomposition required in (5). The complete C-ADMM algorithm is shown in Algorithm 1.

## 4. EXPERIMENTS

In this section, we show the results of our experimental evaluation conducted on the novel SALSA (Synergistic social Scene Analysis) dataset [2]. Then, we provide a detailed description of our experimental setup and show the results of our extensive evaluation, demonstrating the effectiveness of the proposed MC-HBPE method. Finally, we perform further experiments on the analysis of social interaction in SALSA, showing that the computed head and body orientations are accurate and robust so to efficiently detect F-formations and discover social attention attractors.

### 4.1 SALSA

The SALSA dataset records a two-part social event involving 18 participants. While the first half consists on a poster presentation session (see Fig.1), the second half consists on a cocktail party over food and beverages. The scene of SALSA is captured by a distributed camera network and by sociometric badges worn by the participants. Specifically, visual data are recorded by four synchronized static RGB cameras operating at 15 FPS. The sociometric badges are equipped with a microphone and an infrared (IR) beam and detector. Importantly, these wearable devices are battery-powered and store the data on a USB card without the need of any wired connection, thus guaranteeing a natural social interplay. SALSA also comprises manual annotations every 3 seconds of position, head and body orientation of each target as well as F-formations.

### 4.2 Experimental setup

In order to extract the visual features for body and head,  $\mathbf{v}_{tk}^{\mathbf{b}}$  and  $\mathbf{v}_{tk}^{\mathbf{h}}$ , the participants can be tracked on the visual stream with multi-target tracking algorithms with state-of-the-art occlusion handling such as [26]. The estimates of

the ground positions are used together with the camera calibration to retrieve the bounding box of the head and body of all participants for each camera view. We used those frames in which all participants are seen from all four cameras. Inspired by [9], we choose to describe each bounding box with a Histogram of Oriented Gradient (HOG) descriptors [12]. More precisely, the head and body images are first normalized to  $20 \times 20$  and  $80 \times 60$  pixel images, respectively, from which we extract HOG in non-overlapping cells of  $4 \times 4$  pixels. Finally, the descriptors for all the four views are concatenated and the dimensionality of the head and body descriptors is reduced using Principal Component Analysis, keeping 90% of the variance. In all, the head and body visual features extracted from the four cameras are first normalized ( $\mathbf{v}_{tk}^{\mathbf{b}}$  and  $\mathbf{v}_{tk}^{\mathbf{h}}$ ) are roughly 100-dimensional.

As outlined in Section 3, the infrared and auditory information from the sociometric badges are processed in order to automatically extract body and head pose labels, respectively. Indeed, the different nature of the infrared and auditory signals carries a rich representation of the badge's owner with respect to the other participants. On one side, if the badge of person  $k$  detects the infrared beam of person  $l$ , most likely  $k$ 's body orientation points to  $l$ , since the badge settles on its owner's chest. On the other side, if the auditory signal of  $k$ 's badge is highly correlated with the one of  $l$ , it is likely that  $l$  is speaking towards  $k$ , and that therefore the head of  $l$  points towards  $k$ . It is worth noticing that, whilst visual features are extracted in a continuous manner, infrared and auditory labels are sparse. Finally, inspired by [9], we discretize the angular space to  $C = 8$  sectors (corresponding to an angular resolution of  $45^\circ$ ) for both head and body labels.

### 4.3 Head and body pose estimation

To experimentally validate the technical core of the present study, *i.e.* the MC-HBPE, we considered those frames of SALSA in which all the participants are in the field of view of the four cameras. In practice, we use 340 frames of the poster session (17 minutes approximately), resulting in a data set comprising 6120 head/body samples. We split this set into two equal parts, and use only the annotations of the first part for training: 10% of manual annotations and 90% of weak annotations obtained with the sociometric badges, when available. In essence, we have 5% of manual annotations for head and body pose, and therefore the labels are sparse. We report the classification accuracy of the test samples, *i.e.* the second half of the data set. In all our experiments, and for both our approach and the baseline methods, the regularization parameters are set with cross-validation.

Figure 4 shows the performance obtained with our method when the training labels come from different modalities. At the light of the results, we can clearly state that using multimodal labels significantly improves the accuracy of both head and body pose estimates. Interestingly, when considering only ground truth and auditory labels (*i.e.* no additional labels for body pose), not only the accuracy of the head pose estimates improves, but also the accuracy of body pose estimates. This is an important and remarkable effect of the head/body coupling term introduced in the proposed formulation. We notice the analogous effect with the performance rise of, not only body pose estimates, but also head pose estimates, when adding the infrared labels (*i.e.* with no additional labels for head pose). It is worth noticing that



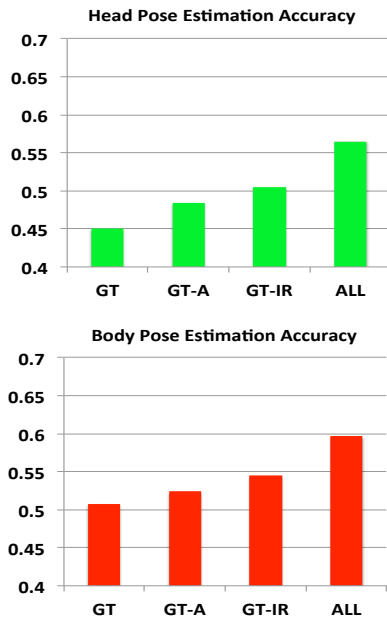


Figure 4: Accuracy of the head (top) and body (bottom) pose estimates obtained with the proposed MC-HBPE approach using visual features and labels from different modalities: ground truth annotations (GT), GT and audio data (GT-A), GT and infrared data (GT-IR), all modalities (ALL).

infrared labels have a significantly more positive effect than auditory labels. We ascribe this fact to the higher accuracy of the infrared labels (87%) compared to the auditory labels (70%). when compared to the manual annotations. Importantly, we highlight the synergistic effect of jointly using auditory, infrared and visual data to robustly solve the head and body pose estimation problem by means of matrix completion techniques. Finally, the implementation of the proposed methodology needs roughly 0.5 seconds to estimate the head and body poses of one person.

Table 1 reports the classification accuracy of several methods. First of all, the proposed MC-HBPE method considering different configurations regarding the head/body coupling and the temporal consistency (from top to bottom): (i) only with head-body coupling  $\tau_b, \tau_h = 0$  in (7); (ii) only with temporal consistency  $\lambda_c$  equal to 0 in (7) and (iii) with both types of constraints. The comparison with the state-of-the-art is particularly difficult in our case, since there is no existing methodology for head and body pose estimation using multimodal data. That is why we chose to compare with the state-of-the-art on *visual-only* HBPE [9]. Notably, the conversational flavor of SALSA is challenging for [9], since when analyzing FCG the direction of motion cannot be considered as a proxy for the body pose, as done in [9]. Moreover, given that the proposed learning methodology belongs to the class of transductive learning approaches, and [9] does not, we perform additional comparisons with Transductive Support Vector Machines (TSVM) [23]. It is worth noting that neglecting the coupling and the temporal consistency terms in MC-HBPE, *i.e.*  $\tau_b, \lambda_c, \tau_h = 0$  in (7), is similar to the work of Goldberg *et al.* [21]. Therefore, MC-HBPE could be seen as a generalization of the transductive methodology presented in [21]. Results in Table 1 demonstrate the ad-

Table 1: Classification accuracy of different methods for HBPE.

Method	$C$	$T_b, T_h$	Body	Head
MC-HBPE	✓		0.515	0.465
	✓	✓	0.548	0.515
Chen <i>et al.</i> [9]			0.524	0.475
TSVM [23]			0.481	0.429
Golberg <i>et al.</i> [21]			0.479	0.439



Figure 5: Frame example: (left) the original image of one of the cameras, (right) the head-green and body-red pose estimates plot in bird view.

vantageous performance obtained when enforcing temporal consistency or coupling the head and body pose estimates. Importantly, the effect of both regularizations is impressive, increasing by approximately 20% the classifiers' accuracy. Our understanding, is that this is the main reason for which the complete method we presented outperforms the base-lines up to a large extent. In order to show the complexity of the task, we show the original image of one of the cameras and the estimated poses plot in a bird's eye view in Figure 5. Finally, we evaluate the robustness of the approach to the amount of training and of manually annotated data in Figure 6, where we remark the 40% accuracy obtained with only 1% of manually annotated data (plus the noisy labels from the sociometric badges).

#### 4.4 F-formation detection

One of the prominent applications of robust estimation of the head and body poses when analyzing FCGs is the detection of F-formations. We evaluate three state-of-the-art approaches for F-formation detection, namely: the Hough voting method (HVFF-lin) in [11], its multi-scale extension (HVFF-ms) [33] and the graph cut approach in [34]. We chose these methods because their implementation is publicly available<sup>2</sup>. All these approaches compute the F-formation independently frame by frame, from the targets' position and pose. The rationale behind the Hough voting methods is that each target generates a set of samples in front of him/her, which are candidates of the F-formation centers. By aggregating these samples into a discrete number of cells, F-formations are detected by finding local maxima in the discretized space. Conversely, in the graph-cut algorithm [34], an optimization problem is solved, discovering directly the coordinates of the F-formations.

Previous works [22] have acknowledged the importance of accurate body pose estimates for detecting F-formations. However, one of the inherent difficulties within the analysis of FCGs in crowded indoor environments is the adverse im-

<sup>2</sup><http://profs.sci.univr.it/~cristanm/ssp/>

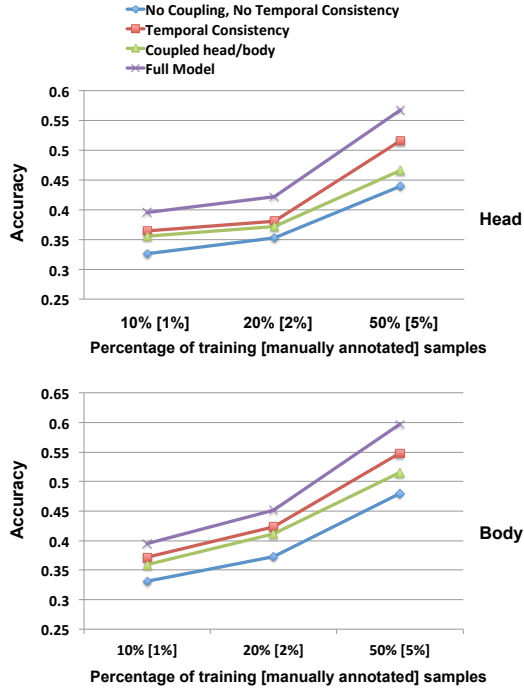


Figure 6: Classification accuracy for different percentage of training samples and of manually annotated samples (in square brackets).

pect of highly occluded body images, which empower strong noise and outliers in the estimated body poses. Classically, the head pose is used as a surrogate to the body pose, and the F-formation are detected from the head pose. In Section 4.3, we demonstrate that the proposed MC-HBPE approach can be effectively used to accurately infer the body pose, thus making possible to use the body pose for F-formation detection and therefore improving the quality of the analysis of FCGs.

We implement the validation protocol defined in [11], and report F-formation precision, recall and F1-measure in Table 2. At 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  $T = 2/3$ . Table 2 shows the performance measures using body (first three rows) and head (fourth and fifth rows) pose. We compare the estimates obtained with MC-HBPE to the body pose ground truth and the body estimates computed with [9]. Regarding the head pose, in addition to our approach, we also used the transductive SVM approach in [23] to emulate previous works on F-formation detection that used head pose estimates. The results clearly confirm that accurate inference of the body pose using the MC-HBPE framework is advantageous for the detection of F-formations.

#### 4.5 Retrieving Social Attention Attractors

Automatically discovering attentional patterns from multimodal data is an unexplored area of research. Interestingly, retrieving the person who holds the attention of the other members in the group can be helpful for other higher-level tasks. In the short-term it can be used to describe who is leading the conversation. In the mid-term, it can describe the role of the person during the interaction. And in

Table 2: Performance of state-of-the-art approaches on F-formation detection from different estimations of the head and body pose. Each table cell contains precision, recall and F1-measure, in this order.

	HVFF lin [11]	HVFF ms [33]	GC [34]
Our-B	0.62 / 0.61 / 0.62	0.66 / 0.64 / 0.66	0.69 / 0.66 / 0.67
[9]-B	0.56 / 0.58 / 0.58	0.59 / 0.62 / 0.61	0.62 / 0.61 / 0.62
GT-B	0.66 / 0.65 / 0.65	0.72 / 0.69 / 0.71	0.74 / 0.72 / 0.73
Our-H	0.61 / 0.59 / 0.60	0.64 / 0.63 / 0.63	0.65 / 0.64 / 0.64
[23]-H	0.58 / 0.58 / 0.58	0.59 / 0.58 / 0.59	0.62 / 0.59 / 0.61
GT-H	0.65 / 0.63 / 0.64	0.70 / 0.67 / 0.69	0.72 / 0.70 / 0.71

the long-term it is related to personality. In this particular study we are interested in interactions that last several minutes, thus we focus on role analysis. In practice, we analyze the scientific poster session of SALSA to automatically find out who are the poster presenters.

In order to do that we introduce the first computational model for the *social attention holding power* (SAHP), which is defined as “the ability to direct favorable attention to the self” [20]. Intuitively, in our scenario the people with high SAHP are likely to be the sought poster presenters, or someone interacting with the presentation *e.g.* asking a question. Within the context of FCGs, SAHP can be computed as the percentage of the F-formation members that are looking towards a particular person.

We consider that if person  $k$  is looking to person  $l$ , the estimated head pose of person  $k$  at time  $t$  ( $\hat{\theta}_{tk}^h$ ) lies within  $\theta^{\max}$  degrees of the direction from  $k$  to  $l$  at time  $t$ . This provides us a binary vector  $\mathbf{A}_{kl}$  of size  $T$ , in which  $(\mathbf{A}_{kl})_t$  is one if  $k$  is looking at  $l$  at time  $t$  and zero otherwise. However this binary matrix is insufficient, since we are interested in the ability of *holding* the attention. Therefore, we consider that holding someone’s attention for more than  $T_N = 5$  seconds is positive, but holding it for less than the same amount, meaning that the attention was not held long enough, is negative. These allows us to split the attentional periods into two groups  $\mathcal{T}_{kl}^+ \subset \{t | (\mathbf{A}_{kl})_t = 1\}$  the temporal frames with an ongoing long held attention (positive SAHP) and  $\mathcal{T}_{kl}^- \subset \{t | (\mathbf{A}_{kl})_t = 1\}$  the temporal frames with an ongoing short held attention (negative SAHP). Formally, the social attention holding power of person  $l$  is computed as:

$$\text{SAHP}_l = \sum_{t=1}^T \frac{1}{|\mathcal{G}_{tl}|} \sum_{k \in \mathcal{G}_{tl}} (\mathbb{I}(t \in \mathcal{T}_{kl}^+) - \mathbb{I}(t \in \mathcal{T}_{kl}^-)), \quad (18)$$

where  $\mathbb{I}$  is the indicator function and  $\mathcal{G}_{tl}$  is the F-formation to which person  $l$  belongs at time  $t$ .

In the addressed scenario, however, part of the attention is held by the posters themselves. Intuitively, the attention given to the poster corresponds to the poster presenter or to an attendee who just took part into the discussion (*e.g.* asking a question). Therefore, we add the attention given to the poster after participation of person  $l$ , to the previously computed  $\text{SAHP}_l$ . Moreover, since we know that there are four posters, we seek the four highest values of SAHP, and we refer to these participants as *social attention attractors*.

We performed an extensive set of experiments to systematically evaluate the robustness of the proposed computational formulation of SAHP against errors in the estimation of the head pose and the detection of F-formations. First, we computed the attention vectors  $\mathbf{A}_{kl}$  with an angular threshold of  $\theta^{\max} \in \{5^\circ, 10^\circ, 15^\circ, 20^\circ\}$ , and the positive



**Table 3: Accuracy results of the social attention attractor detector built from the proposed computational model for the SAHP.**

Parameters		Simulated Data			Real Data			
$W$	$\theta^{\max}$	5°	20°	55°	HP	F-f.	Both	Pos
1'	5°	0.92	0.77	0.74	0.71	0.90	0.71	0.73
	10°	0.92	0.82	0.76	0.72	0.92	0.73	0.70
	15°	0.94	0.87	0.78	0.75	0.90	0.73	0.70
	20°	0.95	0.87	0.82	0.75	0.90	0.75	0.70
2'	5°	0.93	0.78	0.77	0.70	0.91	0.68	0.70
	10°	0.92	0.83	0.77	0.70	0.88	0.69	0.69
	15°	0.94	0.89	0.83	0.77	0.92	0.76	0.71
	20°	0.96	0.90	0.87	0.79	0.92	0.79	0.74
4'	5°	0.96	0.81	0.82	0.70	0.89	0.68	0.71
	10°	0.96	0.88	0.76	0.76	0.91	0.72	0.67
	15°	0.96	0.94	0.85	0.79	0.96	0.78	0.69
	20°	0.98	0.93	0.88	0.79	0.98	0.77	0.76

and negative held attentional periods from them. Thirdly, we calculated the SAHP within a sliding window of length  $W \in \{1, 2, 4\}$  minutes. Finally we compare the retrieved social attention attractors with the ones obtained applying the same procedure to the ground truth data. This evaluation procedure is applied to simulated and to real data. In the former case, we corrupted the head pose ground truth values with uniform random noise of amplitude 5°, 20° and 55°. In the latter case, we evaluate the effect of the estimation procedure of the proposed MC-HBPE together with the F-formation detection state-of-the-art graph-cut method [34] (since it provided the best performance with our data). Finally, we compare our computational model to the straightforward strategy of selecting the person closest to the poster.

We report the accuracy for all combinations of the parameter values in Table 3. The two group columns correspond to the experiments with simulated and real data respectively. In the case of real data we consider three sub-cases: when only the head pose is estimated (HP), when only the F-formation is estimated (F-f.) and when both are estimated (Both). Last single column corresponds to the strategy of selecting the person closest to the poster (Pos). We first notice that, in the case of simulated data, the noise notably decreases the performance of the social attention attractor detector: in severe noise conditions, the accuracy drops down roughly to 0.75. Importantly, we also remark that part of this performance loss can be recovered if we choose the right value for the parameters. Indeed, in severe noise conditions, the right choice of parameters make the performance increase by 0.14. Moreover, the effect of the parameter choice in the robustness against noise is intuitive. On the one side, when the integration window  $W$  is longer, we expect short errors to be smoothed out. On the other side, a larger angular threshold allows to account for low-accuracy estimates of the head pose. We performed experiments with larger values of  $W$  (up to 30') and of  $\theta^{\max}$  (up to 40°), and the accuracy results follow the same trend. Regarding the real data, we remark that the effect of estimating the head pose is more critical than detecting the F-formations. In addition, the proposed model outperforms the Pos strategy probably because: (i) the person closest to the poster is not necessarily the social attractor at a given time (ii) the poster presenter is not systematically the social attractor. The results confirm this fact since the accuracy of the head estimates is lower than the accuracy of the detected

F-formations. Remarkably, the effect of the parameters follows the same trend as in the case of simulated data. At the light of this results we conclude that, there is still room for improvement in the estimation of the head and body poses within FCGs in crowded environments. Furthermore, since robust estimation of head and body pose is useful for higher-level analysis, and no previous methodologies exist to fuse multimodal data flows emanating from a combination of distributed and wearable sensor networks, we believe it is worth pursuing further this line of research.

## 5. CONCLUSIONS AND FUTURE WORK

The focus of this research study is the analysis of free-standing conversational groups from multimodal estimates of head and body pose. We propose a framework able to (i) combine data emanating from a distributed camera network and from sociometric badges worn by each participant, (ii) fuse data from different modalities to boost the estimation of head and body pose, (iii) alleviate the annotation process thanks to the multimodal labels provided by the sociometric badges, (iv) handle noisy features and labels by means of the matrix completion framework, (v) account for the temporal consistency and head/body coupling inherent to the scenario and (vi) provide robust estimates of the head and body poses that have proven to be useful for higher-level tasks, such as the detection of F-formations and the discovery of social attention attractors.

At the light of the results obtained through the extensive experimental evaluation of the presented methodology, we believe that this study opens the door to several interesting future research directions. First of all, the links with other related tasks should be exploited in a bidirectional way. Indeed, the estimation of the head and body poses can ease the tracking of multiple people. Similarly, the recognition of F-formations and discovery of social attention attractors can help the estimation of head and body poses. Second, the spectrum of applications should be enlarged to address personality recognition. Importantly, the analysis of personality traits could require longitudinal studies (*i.e.* extend the recordings to days or weeks), and the current methodology should be adapted to this scenario. Finally, from the technical point of view, one could focus on speeding up the current MC-HBPE algorithm and develop incremental matrix completion strategies, so to adapt the proposed methodology to real-time processing.

## Acknowledgments

This work has been partially supported by the MIUR Cluster project Active Ageing at Home, the EC project ACANTO and the A\*STAR Singapore Human-Centered Cyber-physical Systems (HCCS) grant.

## 6. REFERENCES

- [1] X. Alameda-Pineda and R. Horaud. Vision-guided robot hearing. *IJRR*, 34(4-5):437–456, 2015.
- [2] 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:1506.06882*, 2015.
- [3] B. Benfold and I. Reid. Unsupervised learning of a scene-specific coarse gaze estimator. In *ICCV*, 2011.

- [4] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
- [5] R. Cabral, F. De la Torre, J. Costeira, and A. Bernardino. Matrix completion for multi-label image classification. *TPAMI*, 37(1):121–135, 2014.
- [6] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, and R. A. Peterson. People-centric urban sensing. In *Int. Work. on Wireless Internet*, 2006.
- [7] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theory*, 56(5):2053–2080, 2010.
- [8] C. Canton-Ferrer, C. Segura, J. R. Casas, M. Pardas, and J. Hernando. Audiovisual head orientation estimation with particle filtering in multisensor scenarios. *EURASIP Jour. on Adv. in Sig. Proc.*, 2008.
- [9] 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.
- [10] T. Choudhury and A. Pentland. Sensing and modeling human networks using the sociometer. In *Inter. Sym. on Wearable Comp.*, 2003.
- [11] M. Cristani et al. Social interaction discovery by statistical analysis of F-formations. In *BMVC*, 2011.
- [12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
- [13] F. De la Torre et al. Guide to the Carnegie Mellon University multimodal activity (CMU-MMAC) database. Technical report, 2009.
- [14] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel. Probabilistic temporal head pose estimation using a hierarchical graphical model. In *ECCV*, 2014.
- [15] T. M. T. Do and D. Gatica-Perez. Human interaction discovery in smartphone proximity networks. *Personal and Ubiquitous Computing*, 17(3):413–431, 2013.
- [16] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.
- [17] S. Escalera, X. Bar, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce, H. J. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014: Dataset and results. In *ECCV Workshops*, 2014.
- [18] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12):1775–1787, 2009.
- [19] X. Geng and Y. Xia. Head pose estimation based on multivariate label distribution. In *CVPR*, 2014.
- [20] P. Gilbert, J. Price, and S. Allan. Social comparison, social attractiveness and evolution: How might they be related? *New Ideas in Psychology*, 13(2), 1995.
- [21] A. Goldberg, B. Recht, J. Xu, R. Nowak, and X. Zhu. Transduction with matrix completion: Three birds with one stone. In *NIPS*, 2010.
- [22] H. Hung and B. Kröse. Detecting F-formations as dominant sets. In *ICMI*, 2011.
- [23] T. Joachims. Transductive inference for text classification using SVM. In *ICML*, 1999.
- [24] V. Kalofolias, X. Bresson, M. Bronstein, and P. Vandergheynst. Matrix completion on graphs. *arXiv:1408.1717*, 2014.
- [25] N. Krahnstoeve, M.-C. Chang, and W. Ge. Gaze and body pose estimation from a distance. In *AVSS*, 2011.
- [26] O. Lanz. Approximate bayesian multibody tracking. *TPAMI*, 28:1436–1449, 2006.
- [27] F. Lingensfelder, J. Wagner, E. André, G. McKeown, and W. Curran. An event driven fusion approach for enjoyment recognition in real-time. In *ACMMM*, 2014.
- [28] A. Matic, V. Osmani, A. Maxhuni, and O. Mayora. Multi-modal mobile sensing of social interactions. In *PervasiveHealth*, 2012.
- [29] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang. Affective multimodal human-computer interaction. In *ACMMM*, 2005.
- [30] S. Petridis, B. Martinez, and M. Pantic. The mahnob laughter database. *Image and Vision Computing*, 31(2):186–202, 2013.
- [31] A. K. Rajagopal, R. Subramanian, E. Ricci, R. L. Vieri, 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.
- [32] N. Robertson and I. Reid. Estimating gaze direction from low-resolution faces in video. In *ECCV*, 2006.
- [33] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani. Multi-scale F-formation discovery for group detection. In *ICIP*, 2013.
- [34] F. Setti, C. Russell, C. Bassetti, and M. Cristani. F-formation detection: Individuating free-standing conversational groups in images. *arXiv:1409.2702*, 2014.
- [35] J. Shotton et al. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.
- [36] Y. Song, L.-P. Morency, and R. Davis. Multimodal human behavior analysis: learning correlation and interaction across modalities. In *ICMI*, 2012.
- [37] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *ICACII*, 2009.
- [38] M. Voit and R. Stiefelhagen. 3d user-perspective, voxel-based estimation of visual focus of attention in dynamic meeting scenarios. In *ICMI*, 2010.
- [39] B. Wu, S. Lyu, B.-G. Hu, and Q. Ji. Multi-label learning with missing labels for image annotation and facial action unit recognition. *Pattern Recognition*, 2015.
- [40] L. Wu, R. Jin, and A. K. Jain. Tag completion for image retrieval. *TPAMI*, 35(3):716–727, 2013.
- [41] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *ICCV*, 2013.
- [42] Y. Yan, E. Ricci, R. Subramanian, G. Liu, and N. Sebe. Multitask linear discriminant analysis for view invariant action recognition. *IEEE Transactions on Image Processing*, 23(12):5599–5611, 2014.