Jointly Estimating Interactions and Head, Body Pose of Interactors from Distant Social Scenes

Ramanathan Subramanian[1], Jagannadan Varadarajan[1], Elisa Ricci[2,3],
Oswald Lanz[2], Stefan Winkler[1]
[1]Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore,
[2]Fondazione Bruno Kessler, Trento, Italy. [3]University of Perugia, Italy
Subramanian.R,vjagan,Stefan.Winkler@adsc.com.sg, elricci,lanz@fbk.eu

ABSTRACT

We present **joint** estimation of *F-formations* and *head*, *body pose* of interactors in a social scene captured by surveillance cameras. Unlike prior works that have focused on (a) discovering F-formations based on head pose and position cues, or (b) jointly learned head and body pose of individuals based on anatomic constraints, we exploit positional and pose cues characterizing interactors and interactions to jointly infer both (a) and (b). We show how the joint inference framework benefits both F-formation and head, body pose estimation accuracy via experiments on two social datasets.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; I.5.4 [**Pattern Recognition Applications**]: Computer vision

General Terms

Algorithms, Measurement, Human Factors

Keywords

F-formations; Head and Body Pose; Joint Estimation; Social Scenes

1. INTRODUCTION

Following considerable research progress in the areas of computer vision and multimodal analysis, examination of complex phenomena such as *social interactions* is now possible. Social interactions are commonplace in our daily lives and have been extensively studied by psychologists in a variety of contexts. Social interactions provide a wealth of information concerning individual and group behavior, and while most automated social interaction analysis methods have focused on round-table meetings [8], recent works have examined unstructured meeting scenes [11] (*e.g.*, cocktail party) involving free-standing conversational groups (FCGs). FCGs naturally emerge in social settings, and individuals constituting an FCG are characterized by mutual scene locations and head/body orientations resulting in distinct spatial patterns known as F-formations.

MM'15, October 26–30, 2015, Brisbane, Australia.
 ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806343.

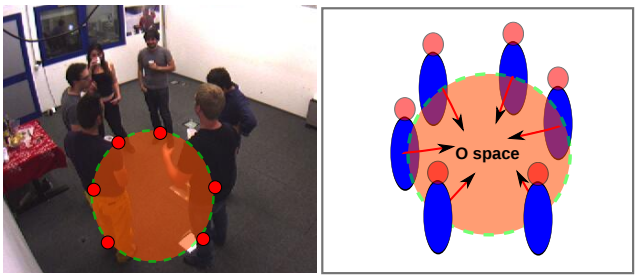

Figure 1: (Left) Social scene from the *Cocktail-Party* dataset [11]. We *jointly* estimate F-formations and the head, body pose of targets exploiting the interation-interactor relationship in terms of positional and pose constraints. Red circles denote targets' feet positions and O-space corresponding to the F-formation is shown via the circular region connecting target positions. (Right) Body pose of interactors serves as our primary cue for determining the O-space center and the corresponding F-formation.

F-formations (FFs) are primarily detected using visual cues [4, 9] and FF detection from a distant social scene captured by surveillance cameras is challenging as seen in Fig.1. Fig.1 (left) shows an FCG comprising six targets– despite being a small group, most targets are partially or severely occluded and are captured at low-resolution, making head and body pose estimation difficult. An F-formation is defined via the O-space (Fig.1 (right)), which is the smallest empty convex space encompassed by the interactors. The body orientation of targets better defines the O-space geometrically[1] as compared to head pose which can frequently change during conversations. Still, head pose is typically used for FF detection [4, 7, 9] due to severe body occlusions.

We present the first work to exploit the *social context* to *jointly* estimate targets' head and body pose and F-formations in a social scene. Prior pose estimation (PE) works have jointly learned head and body pose of individuals based on anatomic constraints [2, 3], while FF detection works have mainly used position and head pose cues [4, 9]. Differently, we exploit positional and pose constraints governing the *synergetic* interaction-interactor relationship in

[1]FF members typically orient their bodies towards the O-space center.

social scenes to jointly infer both FFs and pose of targets. Specifically, FFs are characterized by mutual locations and head, body orientations of interactors, and conversely, interactors are constrained in terms of the head and body pose they can exhibit, motivating the need for joint learning. Our multimodal framework[2] (i) exploits both annotated and unlabeled examples to learn the range of possible joint head-body orientations, (ii) employs positional and pose-based constraints relating interactors to discover FFs, and (iii) progressively refines pose estimates of interactors based on the gained FF knowledge and vice-versa.

Contributions: (i) We present a multimodal framework for jointly estimating targets' head, body orientations and F-formations in social scenes. We show the benefit of joint learning via experiments on two social datasets. (ii) Different from prior works, we primarily use body orientation for estimating F-formations. Precise body pose estimates are computed via coupled head-body pose learning, and knowledge of F-formations.

2. SOCIAL SCENE ANALYSIS

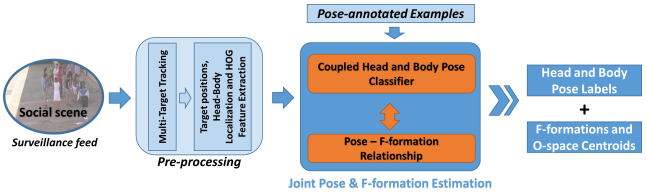

Figure 2: Our social scene analysis framework.

Overview: An overview of our social scene analysis (SSA) method is presented in Fig.2. Given the surveillance video of a social gathering, we apply multi-target tracking to estimate the feet positions of targets in the scene. As in [10], a 3-D head-plus-shoulder model is registered to each target via shape matching with color information provided by the tracker to output head coordinates. The body region is then determined as the portion lying between the head and feet coordinates. Head and body images are respectively normalized to 20×20 and 80×60 pixels, and HoG features are extracted over 4×4 cells, resulting in 775 and 2480-dimensional head and body descriptors. These features and target positions are input to our SSA algorithm to output (a) head, body pose along with FF membership for each target, and (b) O-space centroids for the detected FFs.

2.1 Joint pose and FF learning

Given a T-frame surveillance video of a social scene with K persons (targets), we assume to have the head and body crops for each target k at each frame t. In other words, for each target k, we have a set of samples $\mathcal{S}_k = \{\mathbf{x}^B_{k,t}, \mathbf{x}^H_{k,t}\}_{t=1}^T$, where $\mathbf{x}^B_{k,t} \in I\!R^{D_B}, \mathbf{x}^H_{k,t} \in I\!R^{D_H}$ are the head and body HOG descriptors of dimensionality D_B and D_H respectively. For each frame t, we also have the tracker-output ground position for each target k, $k = 1, \ldots, K$, *i.e.*, over the video length, for each target k we have a set $\mathcal{P}_k = \{\mathbf{p}_{k,t}\}_{t=1}^T$ with $\mathbf{p}_{k,t} = (p^x_{k,t}, p^y_{k,t})$ respectively denoting (x, y) feet positions.

[2]Even though our approach is vision-based, we employ multiple cues such as target positions, their head, body pose and FF membership in our model.

In addition to social scene features, we also assume to have training data from an (independent) annotated dataset, $\mathcal{T}_B = \{(\hat{\mathbf{x}}^B_i, y^B_i\}_{i=1}^{N_B}$, $\mathcal{T}_H = \{(\hat{\mathbf{x}}^H_i, y^H_i\}_{i=1}^{N_H}$, where $\hat{\mathbf{x}}^B_i \in I\!R^{D_B}$, $\hat{\mathbf{x}}^H_i \in I\!R^{D_H}$ denote HOG body and head descriptors, while $y^B_i \in \{0,1\}^{C_B}$, $y^H_i \in \{0,1\}^{C_H}$ are the corresponding labels (*i.e.*, $y^B_i = [0, 0, \ldots, 1, \ldots, 0]$, $y^H_i = [0, 0, \ldots, 1, \ldots, 0]$ where the position of the non-zero element indicates the class label of the training sample). In this work, we quantize the set of possible head and body pose directions into $C_H = C_B = 8$ possible classes, each denoting a 45^o head/body *pan* range.

We define matrices $\mathbf{X}^B_k \in I\!R^{D_B \times K} = [\mathbf{x}^B_{k,1}, \ldots, \mathbf{x}^B_{k,T}]$ and $\mathbf{X}^H_k \in I\!R^{D_H \times K} = [\mathbf{x}^H_{k,1}, \ldots, \mathbf{x}^H_{k,T}]$ obtained by concatenating head and body features for target k, and define global matrices $\mathbf{X}^B \in I\!R^{D_B \times KT} = [\mathbf{X}^B_1, \ldots, \mathbf{X}^B_K]$ and $\mathbf{X}^H \in I\!R^{D_H \times KT} = [\mathbf{X}^H_1, \ldots, \mathbf{X}^H_K]$ collating features for all targets. Similarly, the matrix $\mathbf{P}$ is defined as $\mathbf{P} = [\mathbf{P}_1, \ldots \mathbf{P}_T]$, where $\mathbf{P}_t = [\mathbf{p}_{1,t}, \ldots, \mathbf{p}_{K,t}]$, *i.e.*, $\mathbf{P} \in I\!R^{2 \times KT}$ contains feet positions of all targets over the video length. We also consider matrices $\hat{\mathbf{X}}^B \in I\!R^{D_B \times N_B} = [\hat{\mathbf{x}}^B_1, \ldots, \hat{\mathbf{x}}^B_{N_B}]$ and $\mathbf{Y}^B = [y^B_1, \ldots, y^B_{N_B}]$, obtained by concatenating head, body features and corresponding labels from the annotated data. Similarly, $\hat{\mathbf{X}}^H \in I\!R^{D_H \times N_H} = [\hat{\mathbf{x}}^H_1, \ldots, \hat{\mathbf{x}}^H_{N_H}]$ and $\mathbf{Y}^H = [y^H_1, \ldots, y^H_{N_H}]$.

We propose a joint framework to infer the head and body pose of all the targets in the social scene along with FFs. More formally, we propose to jointly learn head and body pose classifiers $f^H : D_H \rightarrow I\!R^{C_H}$ and $f^B : D_B \rightarrow I\!R^{C_B}$, and respectively parametrized by matrices $\mathbf{\Theta}_B \in I\!R^{C_B \times D_B}$ and $\mathbf{\Theta}_H \in I\!R^{C_H \times D_H}$, and indirectly infer F-formations based on shared cluster memberships. We propose to solve the following optimization problem:

$$\min_{\mathbf{\Theta}_B, \mathbf{\Theta}_H, \mathbf{\Theta}_F} \quad \mathcal{L}_{\mathbf{\Theta}_B, \mathbf{\Theta}_H}(\hat{\mathbf{X}}^B, \hat{\mathbf{X}}^H, \mathbf{Y}^B, \mathbf{Y}^H) \qquad (1)$$
$$+\lambda_U \mathcal{U}_{\mathbf{\Theta}_B, \mathbf{\Theta}_H}(\mathbf{X}^B, \mathbf{X}^H) + \lambda_F \mathcal{F}_{\mathbf{\Theta}_B, \mathbf{\Theta}_F}(\mathbf{X}^B, \mathbf{P})$$

where $\mathbf{\Theta}_F$ denotes FF parameters. The objective function is the sum of three terms. The first term $\mathcal{L}(\cdot)$ leverages annotated data to learn both the head and body pose classifiers. Formally, we minimize the training error on labeled head and body samples, while regularizing the classifiers to reflect the distribution of the training examples. The second term $\mathcal{U}(\cdot)$ exploits unlabeled examples gathered from the social scene to improve PE performance. The last term models the relationship between targets' body pose and FFs. Intuitively, if the precise body orientation of interactors is known, F-formations can be detected. Conversely, if FFs are known, body pose of interactors can be enforced/refined. The parameters λ_U and λ_F regulates the importance of the three terms. We now describe loss functions $\mathcal{L}(\cdot)$, $\mathcal{U}(\cdot)$ and $\mathcal{F}(\cdot)$.

Training loss: The first term $\mathcal{L}(\cdot)$ in Eq(1) implements the traditional trade-off between minimization of the empirical error on labeled data and regularization. More formally, considering the head and body training data, we define $\mathcal{L}(\cdot) = \mathcal{L}^H(\cdot) + \mathcal{L}^B(\cdot)$ and:

$$\mathcal{L}^\triangle = \|\mathbf{Y}_\triangle - \mathbf{\Theta}_\triangle \hat{\mathbf{X}}_\triangle\|^2 + \lambda_r \|\mathbf{\Theta}_\triangle\|^2$$
$$+\lambda_g \sum_{(i,j) \in \mathcal{G}_\triangle} \gamma^\triangle_{ij} \|\boldsymbol{\theta}^{c_i}_\triangle - \boldsymbol{\theta}^{c_j}_\triangle\|_2$$

where $\triangle = \{B, H\}$. The last term enforces that similar classifiers, corresponding to columns $\boldsymbol{\theta}^{c_i}_\triangle$ of the matrix $\mathbf{\Theta}_\triangle$, are obtained for neighboring poses. This is achieved by defining an appropriate pose graph $\mathcal{G}_\triangle$ such that $\gamma^\triangle_{ij} = 1$

if the classes c_i and c_j correspond to similar head/body orientations. The parameters λ_g and λ_r regulate the trade-off between loss and regularization.

Unsupervised loss: Since one would expect unlabeled data from the social scene to be consistent with the distribution of the labeled data (this is the assumption in semi-supervised manifold learning), we propose to integrate unlabeled data information by adopting a graph-based regularization term as typically done in semi-supervised methods [12]. The assumption is that the manifold in which data are embedded can be approximated by a weighted discrete graph whose vertices are identified with (labeled and unlabeled) training examples. The proposed unsupervised data term $\mathcal{U}(\cdot)$ in Eq(1) is defined as $\mathcal{U}(\cdot) = \mathcal{U}^B(\cdot)+\mathcal{U}^H(\cdot)+\mathcal{U}^P(\cdot)$, where:

$$\mathcal{U}^{\triangle} = \mathbf{\Theta}'_{\triangle}\mathbf{L}_{\triangle}\mathbf{\Theta}_{\triangle}$$

where $\triangle = \{B, H\}$ and $\mathbf{L}_{\triangle}$ is the Laplacian matrix defined as $\mathbf{L}_{\triangle} \in I\!R^{M_{\triangle}\times M_{\triangle}}$, $\mathbf{L}_{\triangle} = \mathbf{D}_{\triangle} - \mathbf{A}_{\triangle}$, where $\mathbf{D}_{\triangle} = \text{diag}(d_i)$, $i = 1, \ldots, M_{\triangle}$, $d_i = \sum_{j=1}^{M_{\triangle}} [A_{\triangle}]_{ij}$ and $M_{\triangle} = N_{\triangle} + KT$ is the total number of labeled and unlabeled data. The $M_{\triangle} \times M_{\triangle}$ adjacency matrix $\mathbf{A}_{\triangle}$ is defined such that $[A_{\triangle}]_{ij} = 1$ if the i-th sample is one of the k-nearest neighbors of j and zero otherwise. The term $\mathcal{U}^P$ couples head and body pose estimates on unlabeled data based on human anatomic constraints (*e.g.*, the body and head cannot be oppositely oriented) and it is defined as follows:

$$\begin{aligned}\mathcal{U}^P &= \sum_{k=1}^{K}\sum_{t=1}^{T} \|f^B(\mathbf{x}^B_{k,t}) - f^H(\mathbf{x}^H_{k,t})\|^2 \\ &= \|\mathbf{\Theta}_H\mathbf{X}_H - \mathbf{\Theta}_B\mathbf{X}_B\|^2\end{aligned} \tag{2}$$

F-formation loss: The third term in the objective function Eq(1) models the relationship between interactors' body pose and the F-formations. Our aim is to iteratively exploit the targets' FF membership for refining body pose estimates as interactors tend to orient towards the O-space center, and conversely, to detect FFs from the body pose of interactors. We consider the interactions between **pairs of targets** and compute at each frame t, the angle β^t_{kq} of the line connecting targets q, k. We formulate FF detection as the problem of learning a set of parameters $\mathbf{\Theta_F} = \{\boldsymbol{C_t}, \boldsymbol{Z_t}\}$, where $z_{kt} \in \boldsymbol{Z}_t \in I\!R^{K\times K}$. At each frame t, we aim to learn both $\boldsymbol{C}_t$, the matrix of O-space centers and the FF membership matrix $\mathbf{Z}_t$. To learn $\boldsymbol{C}_t, \boldsymbol{Z}_t$, we define loss $\mathcal{F}_P$ as follows:

$$\begin{aligned}\mathcal{F}_P = &\sum_{t=1}^{T} \|\hat{\mathbf{P}}_t - \mathbf{C}_t\mathbf{Z}_t\|^2 \\ &+\gamma_p \sum_{t=1}^{T}\sum_{k,q=1}^{K} \delta^t_{kq}\|\eta^B_{k,t} - \beta^t_{kq}\|^2 (\mathbf{z}_{k,t})^T\mathbf{z}_{q,t}\end{aligned} \tag{3}$$

and we minimize it imposing $\mathbf{C}_t, \mathbf{Z}_t \geq 0$. In (3), γ_p is a user-defined parameter, $\mathbf{z}_{k,t}$ denotes the k-th column of the matrix $\mathbf{Z}_t$, δ^t_{kq} is an indicator function, *i.e.* $\delta^t_{kq} = I(\|\mathbf{p}_{k,t} - \mathbf{p}_{q,t}\|^2 < \tau)$ and τ is a user-defined threshold. $\eta_{k,t}$ denotes transformation from $f_B(x^B_{k,t})$ to a real-valued angle obtained by computing the weighted average vector $\sum_{i=1}^{C_B/C_H} \hat{y}_i \vec{n_{\alpha_i}}$, where $\vec{n_{\alpha_i}}$ is the unit vector corresponding to α_i. The proposed function $\mathcal{F}_P(\cdot)$ aims to jointly learn the F-formation detection parameters and the body classifier $\mathbf{\Theta}_B$. Intuitively, minimizing $\mathcal{F}_P(\cdot)$, we enforce that targets belonging to the same FF (*i.e.*, with same membership $\mathbf{z}_{k,t}$) should have the body pose almost aligned with the direction individuated by the angle β_{kq}. On the other hand, detection of FFs is influenced by the body pose of targets, *i.e.*, targets belong to the same FF if they are close-by, and their body orientation is consistent with β_{kq}.

Optimization: To solve the optimization problem (1), we adopt an alternating optimization approach solving for pose classifiers $\mathbf{\Theta}_{\triangle}$ having $\mathbf{\Theta}_F$ fixed, and solving with respect to $\mathbf{\Theta}_F$ when the pose classifiers are known.

3. EXPERIMENTAL RESULTS

3.1 Datasets

We evaluate our framework on the *CocktailParty* (CP) and *CoffeeBreak* (CB) social datasets. **CocktailParty** [11] is a 30-minute video recording of a social event involving six targets in a 30m^2 room, and recorded using four synchronized wall-mounted cameras. We only use images from Camera 1 for analysis. 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, while F-formation annotations are available for every 75^{th} frame. **CoffeeBreak** 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 assigned to one of eight classes. F-formations are annotated for two sequences of lengths 45 and 75 frames respectively. as auxiliary data, we used 1000 frames from the DPOSE dataset [5] which contains head pose measurements, while body pose is computed via walking direction as in [2].

3.2 Quantitative evaluation

Algorithm parameters for our method and other baselines were tuned using a small validation set. To evaluate head, body pose estimation (HBPE) accuracy, we use the mean angular error (in degrees) as defined in [3]. F-formation estimation (FFE) accuracy is evaluated using F1-score as described in [4, 9]. Table 1 shows the average HBPE errors on the CP, CB datasets. Maximum error of about 72^o is obtained for both when the objective function involves only the training error (L), which is within two pose classes. Additionally incorporating unlabeled scene examples (L+U) and coupling head and body pose learning (L+U+H/B) considerably reduces HBPE error, while exploiting knowledge of FFs (L+U+H/B+FF) further minimizes the error to produce the best performance. We compare our approach with the state-of-the-art for joint HBPE [3]. Our algorithm performs significantly better than [3] on both datasets as the *social context* is taken into account, and other cues (*e.g.*, velocity direction) are ineffective when targets are mostly static and heavily occluded. FFE results are compared with the state-of-the-art in Table 2. Specifically, we compare with the linear Hough transform method (HVFF lin) [4], its non-linear (HVFF ent) [6] and multi-scale extensions [7] (HVFF ms). Evidently, we achieve the best performance on both datasets as we primarily use body orientation for detecting FFs, and refine body pose estimates via coupled HBPE.

3.3 Qualitative Results

Fig.3 presents some qualitative results with our algorithm. Specifically, we show the inferred head and body pose for each target along with the detected F-formations. First two columns in Fig.3 present results on CP. Despite being a small group, severe occlusions are observed for conversing targets making HBPE challenging. Note that most of the

Figure 3: Qualitative results on the CP (Columns 1,2) and CB (Columns 3,4) datasets: Pies on the ground located at targets' feet positions denote body pose, while arrows denote head pose. FF members are shown connected (Figure best viewed under zoom).

Table 1: Mean HBPE error with various cues.

	CP		CB	
Method	Head	Body	Head	Body
L	67.3	71.4	68.3	72.6
L + U	61.3	64.7	59.8	62.3
L + U + H/B	58.9	61.6	55.3	59.6
L + U + H/B + FF	**51.7**	**55.3**	**49.6**	**51.4**
Chen *et al.* [3]	58.3	62.7	56.1	60.2

Table 2: FFE evaluation via precison (pre), recall (rec) and F1-scores (F1).

	CP			CB		
Method	pre	rec	F1	pre	rec	F1
HVFF lin [4]	0.64	0.72	0.68	0.74	0.85	0.78
HVFF ent [6]	0.78	0.82	0.79	0.81	0.78	0.79
HVFF ms [7]	0.79	0.80	0.79	0.78	0.85	0.81
Our	0.79	0.82	**0.82**	0.82	0.84	**0.83**

head, body pose and FF estimates are still accurate, demonstrating the effectiveness of our joint learning framework. In column 1, the target coded in yellow is incorrectly left out of the FF. Also, while HPE outputs for the green and yellow targets are erroneous, their body pose is estimated correctly. This is because we link FFs with the body pose of interactors, and progressively refine both body pose and FF estimates upon gaining knowledge of the other. All estimates are correct for the frame in column 2, including for the singleton yellow target who moves away from the group. Columns 3 and 4 show exemplar results on the CB dataset, which also captures a densely crowded social gathering over a large area. Here again, we notice that FF and body pose estimates are generally more correct than targets' head pose estimates. Overall, these results demonstrate the effectiveness of our approach for HBPE and FFE on challenging social datasets.

4. CONCLUSION

Our multimodal social scene analysis framework exploits positional and pose constraints relating interactors to improve both HBPE and FFE accuracy for social scenes involving persistent and considerable occlusions. Different from prior works, we employ body pose of interactors as the primary cue for determining F-formations, and progressively refine both body pose and F-formation estimates via an alternating optimization strategy. Nevertheless, difficulties in studying small groups point to the extreme challenges posed to vision-based tracking and pose estimation while dealing with large groups– incorporating multi-sensory information can help alleviate limitations of purely vision-based analysis [1], and will be the focus of future work.

5. ACKNOWLEDGEMENT

This study is supported by the research grant for ADSC's Human-Centered Cyber-physical Systems (HCCS) Program from Singapore's Agency for Science, Technology and Research (A*STAR).

6. REFERENCES

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

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

[3] 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.

[4] 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.

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

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

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

[8] J. Staiano, B. Lepri, R. Subramanian, N. Sebe, and F. Pianesi. Automatic modeling of personality states in small group interactions. In *ACMMM*, 2011.

[9] 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.

[10] 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.

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

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