On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions

Ramanathan Subramanian
Advanced Digital Sciences Center (ADSC), Singapore
Subramanian.R@adsc.com.sg

Yan Yan
Dept. of Information Engg. and Computer Science, University of Trento, Italy
yan@disi.unitn.it

Jacopo Staiano
Dept. of Information Engg. and Computer Science, University of Trento, Italy
staiano@disi.unitn.it

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

Nicu Sebe
Dept. of Information Engg. and Computer Science, University of Trento, Italy
sebe@disi.unitn.it

ABSTRACT

Correlates between ***social attention*** and ***personality traits*** have been widely acknowledged in social psychology studies. Head pose has commonly been employed as a proxy for determining the social attention direction in small group interactions. However, the impact of head pose estimation errors on personality estimates has not been studied to our knowledge.

In this work, we consider the unstructured and dynamic *cocktail party* scenario where the scene is captured by multiple, large field-of-view cameras. Head pose estimation is a challenging task under these conditions owing to the uninhibited motion of persons (due to which facial appearance varies owing to perspective and scale changes), and the low resolution of captured faces. Based on proxemic and social attention features computed from position and head pose annotations, we first demonstrate that social attention features are excellent predictors of the ***Extraversion*** and ***Neuroticism*** personality traits. We then repeat classification experiments with behavioral features computed from automated estimates– obtained experimental results show that while prediction performance for both traits is affected by head pose estimation errors, the impact is more adverse for Extraversion.

1. INTRODUCTION

Correlates between ***social attention*** and ***personality traits*** have been widely investigated and acknowledged in social psychology literature. Social attention patterns, which denote how individuals distribute their focus-of-attention (FoA) during social interactions, have been found to be excellent predictors of personality traits that manifest at the individual level (*e.g.*, Extraversion) and at the group level (*e.g.*, Dominance). The fact that extraverts influence the attentional patterns of peers and garner more social attention in small group meetings is demonstrated in [18]. Analogously, dominant people are found to pay more attention to the other person while speaking, and less attention while listening in dyadic interactions [8].

Given that attention direction can be estimated from the hierarchical combination of body pose, head pose and point-of-gaze [15], social attention is usually determined based on the direction of head pose or eye-gaze in behavioral studies. When the targets (persons) of interest are captured at low resolutions which precludes computation of the point-of-gaze, head pose is used as the proxy for attention in [25, 27, 3]. However, when faces can be captured at higher resolution employing near-field cameras, determining the point-of-gaze improves social attention estimates as shown in [26].

Typical social attention-based personality prediction frameworks adopt the following methodology– (1) employing head pose estimates to determine the FoA for each target using Hidden Markov Model (HMM) [25] or Gaussian mixture model (GMM)-based [3] approaches, (2) computing social attention features such as *attention received* from and *attention given* to each target, and (3) predicting target personality labels using a classifier to which these attention features are input. As a result, head pose estimation errors in step (1) will impact computation of social attention features in step (2) and subsequently, the personality classification performance. Since most head pose estimation methods provide only a coarse measure of the head orientation, which renders determination of the FoA-target difficult, a number of studies [13, 18] evaluate their personality hypotheses using human-annotated social attention features with automated features only used for comparison. However, no previous work has attempted to study the effects of head pose estimation errors on personality classification performance.

This work attempts to quantify the impact of head pose estimation errors on personality prediction accuracy for the

ICMI '13 Sydney, Australia

unstructured and dynamic social scenario (cocktail party) considered in [28]. Unlike structured, meeting room scenes studied in [13, 18, 3], the *party* setting studied in [28] allows for unconstrained motion of targets– therefore, personality traits can also be modeled in terms of ***proxemics***, which relates to the use of the physical space by targets, in addition to social attention cues. We first demonstrate that social attention features computed from head pose annotations are effective predictors of the Extraversion and Neuroticism traits. We then repeat personality classification experiments with attention features computed using two head pose estimation approaches [16, 9], in order to examine the impact of head pose estimation accuracy on personality prediction. Our analysis reveals that, even with the use of basic social attention features, inaccurate head pose estimation can reduce personality classification performance by over 12% under the considered conditions, emphasizing the need for further research in this domain.

The paper is organized as follows. A description of the dataset and behavioral features used, classification procedure adopted and personality classification results achieved with head pose annotations are presented in Section 2. A brief overview of the head pose estimation approaches employed in this study, and classification results achieved with automated head pose estimates are presented in Section 3. We summarize our findings and conclude in Section 4.

2. SOCIAL ATTENTION FEATURES FOR PERSONALITY PREDICTION

We considered the *cocktail party* data presented in [28] for our analysis. Unlike structured settings considered in most personality studies, involving round table meetings with seated subjects, the authors of [28] consider an unstructured and informal social setting. Two distinguishing factors of the *cocktail party* dataset are that (a) subjects freely move around while being monitored by multiple, distant large field-of-view cameras, which makes head pose estimation challenging due to unrestricted body movements and the low resolution of captured faces (see Fig.1), and (b) the social interactions are hedonistic and experiential in nature, representative of non-goal oriented groups, where the behavior exhibited by participants is more likely to be in accordance with their real personality. In contrast, meetings represent goal oriented groups, where participants interact with an explicit goal, and may need to assume certain roles to achieve the same (as in [14]).

Two video sequences of 20 and 30 minutes duration respectively and involving a total of 13 targets, constitute the *cocktail party* data. Four cameras, fitted at the corners of a $4.8\text{m} \times 6\text{m}$ room, capture the targets at all times[1]. Since no audio is available as part of the data, social attention direction has to be inferred exclusively from head pose. Extraversion and Neuroticism personality scores for the 13 targets are obtained prior to the video recordings through the Big Five Marker Scale [21] self-reported questionnaires. The Extraversion scores range from 29–62 ($\mu = 43.5, \sigma = 9.8$), while Neuroticism scores are between 34 to 57 ($\mu = 42, \sigma = 6.5$).

2.1 Features from manual annotations

To have a consistent framework for personality classification with both manual and automated approaches, we adopted the following procedure. For extraction of social attention features from manual annotations, we employed an annotation tool to specify each target's position and head pose in every video frame as shown in Fig. 1. A cuboid-based 3D human figure model was used to denote target position and head pose– employing camera calibration information, position/head pose in all camera views were available upon specification in any of the views. Finer adjustments were possible by re-annotating the most informative camera view(s) where necessary.

Angular width of the head pose frustum was set to 45° so that the 360° head pan range is segmented into 8 intervals. Based on the assumption that valid FoA-targets in the *party* scenario would be the interacting peers, only near frontal head tilts (tilt $\in [-20°, 20°]$) were considered as valid. Therefore, the annotation tool was essentially used to mark head pan, and the pose measurement for any target exhibiting a pronounced upward/downward head tilt was invalidated, as denoted by the black pose cone in Fig.1(b).

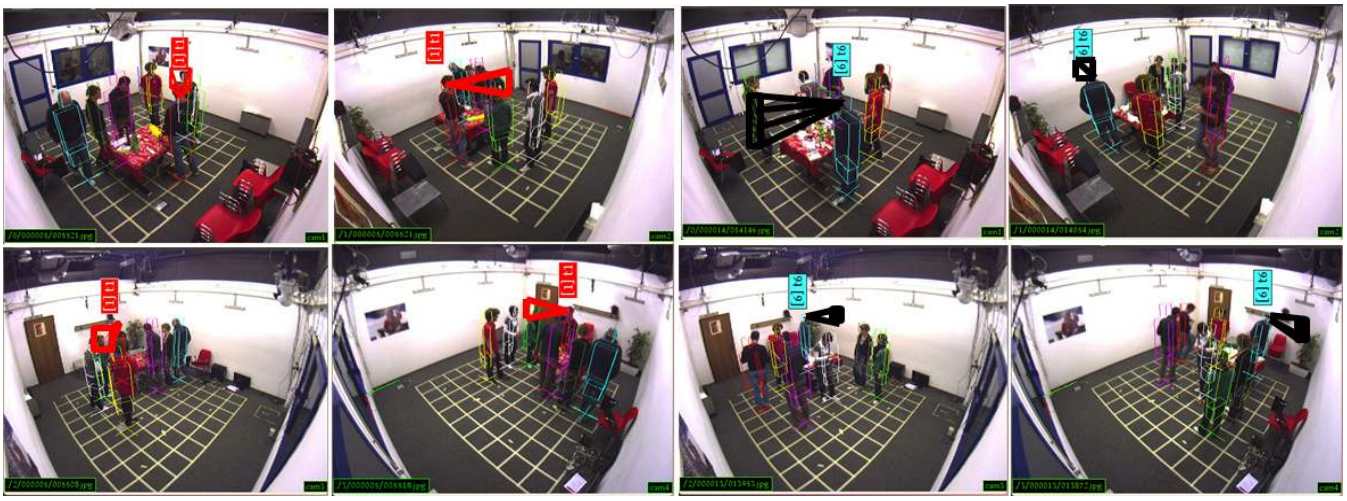
Once the target positions and head pose directions were known (either through annotations or automated estimates), a completely automated procedure was followed to compute requisite social attention features for personality classification. To determine the FoA-target from head pose, we adopted the unsupervised cognitive approach described in [20], while using generic HMM parameters. If the states and observations of a HMM are respectively defined by the FoA-targets and the corresponding head pose angles, an unsupervised approach is proposed in [20] to predict the expected viewing angle (head pan in our case) if the camera geometry, target and FoA-target locations are known. If two FoA-targets correspond to similar viewing angles, then the closer one is assumed to be the FoA-target.

For personality prediction from behavioral features, we adopted the *thin slices* approach [2] which is based on the observation that humans can accurately judge others upon observing very short sequences of expressive behaviors. We also considered different thin slices of 30 sec, 1 min, 3 min and 5 min durations– past studies [4, 18] have reported a steady increase in assessment accuracy with increasing time slice lengths. Aggregating position and head pose information over non-overlapping 30 sec time windows, we computed mean (μ) and standard deviation (σ) of the following quantities as behavioral features over the time slice duration[2]:

- **Proxemic features**: described in terms of (i) the minimum distance (Dist) maintained by the target from the others; (ii) velocity (Vel) or variation in the target's position over the considered time window, and (iii) Relationships-based (Rel) features– as per [12], the distances that humans maintain from others are indicative of the social relationships they hold. Employing identical thresholds as in [28], we calculated the number of *intimate*, *personal-close*, *personal-far* and *public* relationships for each target.
- **Social attention features**: From the FoA-target estimates, we computed the duration proportion over a time window for which each target gave attention to *any* of the others (AG), *i.e.*, we were interested only in whether the target looked at any of the others and did

[1]We did not use the data from three additional pan-tilt-zoom cameras in this work.

[2]Each 30 sec time window provides one sample for the computation of μ, σ statistics over the time slice duration.



(a) (b)

Figure 1: Exemplar head pose annotations using the annotation tool in the four camera views shown two-by-two. (a) The position and the head pose direction of each target are denoted by a 3D stick figure model and a frustum of 45° angular width respectively. (b) Head pose of the blue target is invalidated owing to a pronounced downward head-tilt.

not consider which specific FoA-target was attended to (AG is scalar). Whenever the target head pose was invalidated, or when the head pose direction did not correspond to a valid FoA-target, AG was set to zero for that time instant. Similarly, the duration proportion for which each target received attention from *at least one* of the peers over a time window (AR) was computed. AG (attention given) and AR (attention received) are collectively termed as attentional (Att) features.

2.2 Personality prediction from behavioral features

Firstly, we analyzed the power of behavioral features (excepting Rel) to predict the observed Extraversion, Neuroticism scores by (1) computing the partial correlations between these entities, and (2) performing a series of backward linear analyses to determine the effectiveness of considered behavioral predictors and the set of best predictors for each personality trait.

Table 1 lists the obtained Spearman correlation coefficients for behavioral features extracted over 1 min time slices and corresponding p values. For Extraversion, only the negative correlation with AR_σ is found to be significant. Concerning Neuroticism, the correlations with velocity features are close to being marginally significant. Table 2 presents the regression analyses results, and the variables removed from the best model for each of the considered time slices are also listed. In general, the linear model for Extraversion (maximum R^2 of 0.9) is found to be more effective than the model for Neuroticism (maximum R^2 of 0.31) in terms of predictive power. Also, while the best Extraversion models for the various time slice durations consistently include some attentional features, the best models for Neuroticism always include distance or velocity features. Finally, models with a significant F-statistic are observed with 1 min and 5 min time slices for Extraversion, and for 5 min time slices for Neuroticism.

The obtained results lead to some interesting observations– the importance of attentional features and in particular, of the AR feature, has been noted and discussed in previous studies [18]. The significant negative correlation between AR_σ and Extraversion lends support to this finding, and conforms to the expectation that extraverts consistently attract social attention. On the other hand, past studies (*e.g.*, [7]) have observed that neurotic persons prefer large social distances and tend to avoid eye contact, and distance features are found to best predict neurotics in [28]. While distance features constitute the best Neuroticism predictor model for three time slice durations considered in Table 2, the best prediction (with 5 min time slices) and strongest correlations (even if weakly significant) are observed with velocity features. Weak negative correlation with Vel_μ and positive correlation with Vel_σ suggest that neurotics mostly remain stationary, and tend to move around in spurts. Finally, in contrast to AR_σ which is consistently a good predictor of Extraversion, AG_μ is generally seen to be a good predictor of Neuroticism, indicating that the amount of attention given can be useful for characterizing neurotics.

2.3 Classification experiments

Upon noting that the considered proxemic and social attention features are effective predictors of Extraversion and Neuroticism, we performed a classification experiment on dichotomizing the raw personality scores into 'high' or 'low' class labels. As in [28], raw scores were thresholded based on the median value, resulting in 7 'high' and 6 'low' scores for Extraversion and 8 'high' and 5 'low' scores for Neuroticism. To evaluate classification performance given the small number of samples and the unbalanced class distribution, a leave-pair-out cross validation (LPOCV) procedure was adopted. In each run, a model was trained employing time slice-based behavioral features derived from all-but-two individuals. This model was used to predict personality label for each time slice of the remaining two, with one each denoting a positive and negative class example

Table 1: Partial Spearman correlation coefficients (ρ) and corresponding significance levels (p) considering behavioral features computed over 1 min time slices.

Trait		$Dist_\mu$	$Dist_\sigma$	Vel_μ	Vel_σ	AG_μ	AG_σ	AR_μ	AR_σ
Extraversion	ρ	0.647	-0.401	-0.607	0.427	-0.611	-0.171	-0.528	**-0.842**
	p	0.165	0.431	0.201	0.398	0.197	0.756	0.282	**0.036**
Neuroticism	ρ	-0.484	0.490	**-0.702**	**0.685**	-0.248	-0.358	-0.114	-0.351
	p	0.331	0.324	**0.120**	**0.133**	0.636	0.486	0.830	0.495

Table 2: Backward linear regression analyses with behavioral features computed over the various slices. R^2 denotes coefficient of determination, while p denotes significance level.

Trait		30 sec	1 min	3 min	5 min
Extraversion	Variables removed	all except AG_μ	$Dist_\sigma$	$Dist_\mu$, Vel_μ, Vel_σ, AR_μ	$Dist_\mu$, $Dist_\sigma$, Vel_μ, AG_σ, AR_μ
	R^2	0.267	**0.904**	0.545	**0.8**
	F	1.825	6.761	2.392	11.699
	p	0.211	**0.026**	0.137	**0.002**
Neuroticism	Variables removed	all except $Dist_\mu$, $Dist_\sigma$, AG_μ	AG_σ, AR_μ, AR_σ	Vel_μ, Vel_σ, AG_σ, AR_μ, AR_σ	all except Vel_σ
	R^2	0.199	0.307	0.264	**0.31**
	F	0.747	0.621	1.078	4.984
	p	0.55	0.69	0.406	**0.0473**

respectively. LPOCV is found to provide a classification performance estimate which is significantly less biased than popular cross-validation approaches like leave-one-out cross validation (LOOCV) [1].

In addition to the individual social and proxemic features described above, we also employed feature combinations by concatenating the individual feature vectors for classification ('All' denotes use of the entire set of features). As before, we repeated the classification experiments with features corresponding to various time slice durations. Three different classification strategies were used, namely linear Support Vector Machines (Lin SVM) [5], Random Forests (Rand For) [24], and Locally Weighted Naive Bayes (LW-NB) [11]. Linear SVM is a non-probabilistic binary classifier, which constructs a hyperplane such that its distance from the nearest data point from the positive/negative class is maximum. Random Forests construct multiple decision trees from the training data, and classify based on the output of the tree majority. Also, they do not require specification of a feature space as for SVMs. Locally Weighted Naive-Bayes is a lazy classification algorithm using the Naive-Bayes (NB) classifier where a test instance's class is predicted from an NB model built using a weighted set of training instances in its neighbourhood.

Table 3 outlines the classification results in terms of 2D tuples denoting the mean accuracy and F-measure (harmonic mean of precision and recall). Due to the unbalanced class distribution, the F-measure is a better indicator of classification performance as compared to accuracy. For Extraversion, attentional features and their combinations evidently produce the best classification performance. In contrast, minimum distance and velocity features are ineffective. Relationship-based (Rel) features produce moderate classification performance on their own, and the best performance when used in conjunction with attention features. For Neuroticism on the other hand, minimum distance features acquire good predictive power, while velocity features are still ineffective. Attention features also produce impressive performance on their own, and their use in combination with Dist features generates the best result. Also, using all behavioral features is not more beneficial than using a subset of features for characterizing either personality trait.

A second interesting aspect concerning time slices is that the best classification results with Dist, Vel and Dist+Vel features are obtained for smaller time slice durations, while the best results with Rel and Att features are obtained with larger time slices. When all behavioral features are combined, the best results for Extraversion and Neuroticism are obtained with 5 min and 30 sec time slices respectively. Regarding classifiers, locally weighted Naive Bayes cumulatively achieves better classification as compared to random forests and linear SVM.

It can also be seen that the classification results are in general agreement with the correlation and linear regression analyses, and the trends reported in [28]. In our earlier discussion, we noted that Att features are good predictors of Extraversion, while Dist, Vel and Att features characterize Neuroticism. Classification experiments also corroborate these findings, except that the predictive power of velocity features is not demonstrated by the observed results. As attentional features are not considered in [28], the best classification results are obtained with Rel features for Extraversion and Dist features for Neuroticism. Finally, even though the observed accuracies and F measures are lower than those obtained in [28] on the same dataset[3], the best F-measure

[3]The features used in [28] are automatically obtained, while features used in Table 3 are derived from manual annotations.

scores are very still comparable (0.66 vs 0.63 for Extraversion and 0.69 vs 0.65 for Neuroticism). Here, it must be noted that we use LPOCV instead of LOOCV adopted in [28], and LOOCV provides significantly more optimistic performance estimates than LPOCV for unbalanced data.

3. PERSONALITY CLASSIFICATION WITH AUTOMATED HEAD POSE ESTIMATES

We noted in the previous section that social attention features, in the form of attention given and received, provide vital cues to the characterization of the Extraversion and Neuroticism personality traits. Computing social attention direction through the proxy of head pose for the *cocktail party* scenario is challenging for two reasons– (1) The fact that only distant (even if multiple) large field-of-view cameras can be used implies that targets' faces can only be captured at low resolution (typically less than 50×50 pixels). (2) As targets move, their facial appearance in the multiple cameras varies with position, owing to changing perspective and scale. As a target moves, the face can appear larger/smaller and face parts can become occluded/visible due to the target's relative position with respect to the camera (Fig. 2(a)).

As stated earlier, the motivation of this work is to study the impact of head pose estimation errors on personality prediction under these conditions. In this section, we consider two automated head pose estimation (HPE) approaches, the dynamic head location and pose estimation (DHLP) approach [16] and multi-task SVM method (SVM+ MTL) [9], and compare personality classification results obtained using automatically computed features against those obtained through manual annotations. Descriptions of the two head pose estimation approaches are as follows:

3.1 Dynamic head location and pose (DHLP)

In [16], a simple and efficient approach to jointly estimate head location and orientation, following a Bayesian approach to object tracking is described. The method employs a Bayesian likelihood model, $p(z/x)$, where z denotes observed measurement (target appearance), and x denotes target state (composed of ground coordinates + head pan and tilt), to infer the target position and orientation. Based on an initialization phase, a coarse 3D color+shape model of the target is derived. Assuming a first order Markov process for dynamics, a particle filter is used to propagate state hypotheses (particles) for each target over time. In the prediction step, a new set of particles x is sampled for each target given its current state, while in the verification step the appearance likelihood $p(z|x)$ is assigned to each hypothesis x based on a matching score between the target model and associated appearance z observed in the new images. The re-sampling step withdraws hypothesis with low likelihood from further propagation. The particle with the highest likelihood is selected in order to update the state of each target. The particle filter output is shown in Fig. 2(b). While this simplistic approach is fast and can be employed to determine the location and head pose of multiple targets in real-time, it is not designed to perform accurate head pose estimation.

3.2 Multi-task Support Vector Machine

Having seen that the target facial appearance varies with position, one solution that explicitly accounts for such variation while estimating headpose involves the use of multi-task learning (MTL) [10]. If we divide the physical space into discrete regions, one can expect some similarity among the facial appearances in the different regions and some region-specific appearance differences owing to perspective and scale. Given structured data that can be easily categorized into groups, MTL involves learning the data in each group as an individual task, as well as the *relationships* between the tasks. This, in general, leads to a better model than a learner that does not account for task relationships.

Unlike [16] which outputs the absolute head orientation value, we employed MTL to build a head pose *classifier*, which assigns the target head pose to one of 9 classes (8 classes denoting a quantized $360^\circ/8 = 45^\circ$ head pan, and an 'invalid' head pose class). We divided the *party* room (ground plane) into 4 quadrants, and learnt an MTL classifier for each quadrant. The MTL-based head pose classification (SVM+ MTL) framework consists of three steps (1) Tracking and head localization (2) SVM+ Multi-task learning and (3) Classification. The particle filter tracker employed in [16] is also used in this framework. Then, a 30 cm$\times$30 cm$\times$20 cm-sized dense 3D grid (with 1cm resolution) of hypothetic head locations is placed around the estimated 3D head-position provided by the particle filter. Assuming a spherical model of the head, a contour likelihood is then computed for each grid point by projecting a 3D sphere onto each camera view employing calibration information. The grid point with the highest likelihood sum is determined as the face location (Fig. 2(c)). Upon localization, the face is cropped and resized to 20×20 pixels in each view, and 81 bin HoG [6] descriptors extracted from 4×4 patches in the 4-view face appearance image (Fig. 2(a)) are input to the learning module. In the test phase, depending on the target location corresponding to a test instance (derived from the tracker), the corresponding MTL classifier is invoked to output the head pose class. A brief overview of SVM+ MTL is provided below.

SVM+ MTL: An MTL framework similar to SVMs, applicable when the training data is the union of $t \geq 1$ related groups is presented in [10]. The SVM decision vector $\mathbf{w}$ is decomposed into $w + w_r,\ r \in (1, 2, ..., t)$ where w, w_r respectively model the commonality between groups and group specifics. The decision function is $f_r(x_i) = (w + w_r)^T \phi(x_i)$, where ϕ denotes training feature space. The optimization problem is:

$$\begin{aligned} &\min_{w, w_1, \ldots, w_t} \tfrac{1}{2} w^T w + \tfrac{\beta}{2} \sum_{r=1}^{t} w_r^T w_r + C \sum_{r=1}^{t} \sum_{i=1}^{l_r} \xi_{ir} \\ &s.t. \quad y_{ir}(w^T \phi(x_i) + w_r^T \phi(x_i)) \geq 1 - \xi_{ir}, \\ &\qquad \xi_{ir} \geq 0,\ i = 1, \ldots, l_r,\ r = 1, \ldots, t \end{aligned} \tag{1}$$

Here, all w_r's and the common w are learnt simultaneously. β regularizes relative weights of w and w_r's. ξ_{ir}'s are slack variables for the t data groups, each comprising l_r training samples. y_{ir}'s are training labels while C regulates proportion of nonseparable samples. With relevant quantities as defined in Eq.(1), SVM+ MTL [19] is formulated as:

$$\min_{w, w_1, \ldots, w_t, b, d_1, \ldots, d_t} \tfrac{1}{2} w^T w + \tfrac{\beta}{2} \sum_{r=1}^{t} w_r^T w_r + C \sum_{r=1}^{t} \sum_{i=1}^{l_r} \xi_{ir}$$

$$\begin{aligned} &s.t. \quad y_{ir}(w^T \phi(x_i) + b + w_r^T \phi_r(x_i) + d_r) \geq 1 - \xi_{ir}, \\ &\xi_{ir} \geq 0,\ i = 1, \ldots, l_r,\ r = 1, \ldots, t \end{aligned}$$

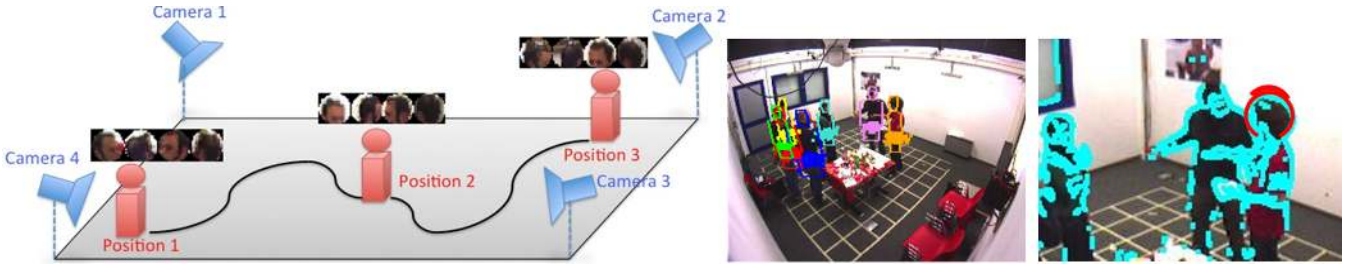
The goal of SVM+MTL is to find t decision functions $f_r(x) = w^T \phi(x) + b + w_r^T \phi_r(x) + d_r, r = 1, \ldots, t$. Therefore,

Table 3: Classification results obtained with the different classifiers and behavioral features extracted from *manual annotations* for different time slices. The two numbers corresponding to a given feature and classifier denote the 2D tuple specifying mean accuracy and F-measure respectively. Values in bold font denote *best* performance with a particular feature (over all time slices).

		Extraversion			Neuroticism		
Slice Dur	**Feature**	**Lin SVM**	**Rand For**	**LW-NB**	**Lin SVM**	**Rand For**	**LW-NB**
30 sec	Dist	0.30,0.24	**0.47,0.46**	0.31,0.28	0.55,0.49	0.51,0.49	0.62,0.55
	Vel	0.44,0.33	**0.48,0.47**	0.46,0.42	0.52,0.44	**0.49,0.47**	0.47,0.31
	Rel	0.42,0.40	0.49,0.48	0.47,0.44	0.48,0.32	0.49,0.45	0.47,0.33
	Att	0.54,0.52	0.47,0.47	0.52,0.48	0.52,0.41	0.49,0.48	0.53,0.47
	Dist+Vel	0.35,0.30	0.45,0.44	0.36,0.32	0.56,0.5	0.53,0.51	**0.65,0.57**
	Dist+Rel	0.39,0.36	0.46,0.46	0.45,0.43	0.58,0.44	**0.55,0.52**	0.58,0.44
	Rel+Att	0.47,0.46	0.46,0.46	0.48,0.44	0.51,0.39	0.50,0.47	0.49,0.38
	Dist+Att	0.45,0.44	0.47,0.47	0.40,0.34	0.57,0.45	0.52,0.50	0.59,0.51
	All	0.37,0.33	0.43,0.42	0.35,0.32	0.58,0.55	0.55,0.53	**0.66,0.59**
1 min	Dist	0.30,0.24	0.43,0.42	0.37,0.35	0.57,0.50	0.49,0.46	**0.60,0.58**
	Vel	0.41,0.30	0.48,0.47	0.49,0.46	0.49,0.38	0.49,0.45	0.47,0.33
	Rel	0.43,0.41	0.50,0.48	0.47,0.42	0.48,0.32	0.48,0.44	0.47,0.34
	Att	0.53,0.51	0.49,0.48	0.52,0.50	0.54,0.47	0.52,0.51	0.55,0.49
	Dist+Vel	0.34,0.28	**0.47,0.46**	0.41,0.37	0.54,0.48	0.51,0.50	0.53,0.44
	Dist+Rel	0.43,0.41	**0.48,0.48**	0.47,0.45	0.57,0.43	0.53,0.48	0.58,0.43
	Rel+Att	0.47,0.46	0.47,0.46	0.46,0.43	0.52,0.41	0.49,0.46	0.48,0.39
	Dist+Att	0.46,0.45	0.49,0.49	0.47,0.44	0.59,0.53	0.55,0.53	0.61,0.55
	All	0.38,0.34	0.46,0.45	0.4,0.36	0.58,0.55	0.57,0.55	0.58,0.51
3 min	Dist	0.31,0.27	0.4,0.39	0.38,0.35	0.52,0.45	0.39,0.37	0.55,0.50
	Vel	0.45,0.38	0.45,0.44	0.44,0.42	0.43,0.33	0.47,0.45	0.38,0.34
	Rel	0.45,0.44	0.46,0.44	0.50,0.48	0.42,0.32	0.44,0.38	**0.55,0.53**
	Att	0.60,0.59	0.49,0.48	0.58,0.55	0.52,0.47	0.55,0.54	0.57,0.56
	Dist+Vel	0.36,0.33	0.41,0.39	0.38,0.36	0.46,0.40	0.39,0.36	0.52,0.49
	Dist+Rel	0.38,0.33	0.46,0.45	0.39,0.37	0.55,0.42	0.52,0.46	0.52,0.41
	Rel+Att	0.52,0.51	0.49,0.47	0.54,0.53	0.54,0.48	0.52,0.48	**0.57,0.54**
	Dist+Att	0.51,0.50	0.50,0.49	0.50,0.49	0.58,0.42	0.53,0.52	0.55,0.49
	All	0.42,0.4	0.41,0.39	0.45,0.42	0.52,0.50	0.48,0.44	0.54,0.51
5 min	Dist	0.25,0.20	0.47,0.44	0.39,0.36	0.60,0.55	0.45,0.41	0.47,0.45
	Vel	0.39,0.34	0.35,0.33	0.48,0.45	0.49,0.41	0.45,0.41	0.48,0.43
	Rel	0.34,0.32	0.52,0.47	**0.56,0.53**	0.46,0.33	0.41,0.35	0.46,0.41
	Att	0.53,0.51	0.48,0.45	**0.63,0.61**	**0.64,0.61**	0.51,0.45	0.64,0.60
	Dist+Vel	0.36,0.32	0.42,0.40	0.44,0.41	0.51,0.46	0.46,0.41	0.47,0.45
	Dist+Rel	0.33,0.29	0.46,0.45	0.39,0.37	0.55,0.41	0.52,0.46	0.57,0.43
	Rel+Att	0.40,0.39	0.49,0.47	**0.64,0.63**	0.57,0.53	0.51,0.44	0.53,0.48
	Dist+Att	0.43,0.41	0.43,0.42	**0.55,0.54**	**0.68,0.65**	0.57,0.53	0.64,0.60
	All	0.35,0.32	0.49,0.47	**0.63,0.62**	0.60,0.58	0.49,0.42	0.50,0.48

Table 4: *Best* accuracy, F-measure obtained using position, head pose *estimates*. Time slice duration and classifier corresponding to the best result are specified in braces. Wherever attentional features are employed, the superior result obtained through either of the HPE approaches is denoted in bold.

HPE Method	Feature	Extraversion	Neuroticism
DHLP [16]	Dist	0.50, 0.50 (1 min, Rand For)	0.52, 0.50 (1 min, Rand For)
	Vel	0.51, 0.51 (30 sec, Rand For)	0.52, 0.51 (1 min, LW-NB)
	Rel	0.50, 0.50 (1 min, Rand For)	0.57, 0.50 (30 sec, Rand For)
	Att	0.50, 0.49 (5 min, Rand For)	0.55, 0.54 (1 min, Rand For)
	Dist+Vel	0.47, 0.47 (1 min, Rand For)	0.57, 0.49 (1 min, Lin SVM)
	Dist+Rel	0.48, 0.48 (1 min, Rand For)	0.59, 0.55 (3 min, Rand For)
	Rel+Att	**0.48, 0.49** (1 min, Rand For)	0.58, 0.52 (5 min, Lin SVM)
	Dist+Att	0.49, 0.47 (5 min, Rand For)	0.51,0.49 (30 sec, Rand For)
	All	0.47, 0.46 (5 min, Rand For)	0.66, 0.62 (30 sec, Lin SVM)
SVM+ MTL [9]	Dist	0.50, 0.50 (1 min, Rand For)	0.52, 0.50 (1 min, Rand For)
	Vel	0.51, 0.51 (30 sec, Rand For)	0.52, 0.51 (1 min, LW-NB)
	Rel	0.50, 0.50 (1 min, Rand For)	0.57, 0.50 (30 sec, Rand For)
	Att	**0.51, 0.49** (3 min, LW-NB)	**0.63, 0.56** (5 min, Lin SVM)
	Dist+Vel	0.47, 0.47 (1 min, Rand For)	0.57, 0.49 (1 min, Lin SVM)
	Dist+Rel	0.48, 0.48 (1 min, Rand For)	0.59, 0.55 (3 min, Rand For)
	Rel+Att	0.47, 0.48 (1 min, Rand For)	**0.62, 0.56** (5 min, Lin SVM)
	Dist+Att	**0.49, 0.49** (3 min, Rand For)	**0.52, 0.51** (1 min, Rand For)
	All	**0.49, 0.48** (5 min, Rand For)	**0.71, 0.64** (30 sec, LW-NB)



(a) (b) (c)

Figure 2: (a) When the target is free to move around, the multi-view facial appearance corresponding to an identical 3D headpose varies due to camera perspective and scale changes for the locations P1-P3. (b,c) illustrate the tracking and head localization procedure. Particle filter outputs are shown in (b), while projection of the spherical head model used for shape-likelihood estimation is shown in red in (c).

each decision function f_r comprises two parts: (a) the common decision function w with bias term b, and the group-specific correction function w_r with bias term d_r. Also, the feature space of f_r involves two spaces, the decision space ϕ and the correction space, ϕ_r. SVM+MTL improves over regularized MTL on two counts: (i) In regularized-MTL, the decision and correcting functions share the same feature space (ϕ), while they may be different in SVM+MTL, thereby providing more flexibility and (ii) SVM+MTL considers a more general form of the decision function with bias terms (b, d_r). SVM+ MTL is a quadratic programming (QP) problem. We adopted generalized sequential minimal optimization (SMO) [19] to solve this optimization problem. SMO training time is of the order of $\approx N^{(1.6-1.9)}$, where N denotes training set size.

3.3 Classification with automated features

Upon obtaining head pose estimates for all targets in the two *party* sequences using the aforementioned methods, we determined the number of frames for which the computed head pose class matched with the annotated pose class– the overall agreement with the head pose annotation was 62.1% and 40.5% with SVM+ MTL and DHLP respectively. While this difference is expected given that the SVM+ MTL approach is devised to explicitly account for appearance changes with position, the fact that head pose estimation performance is not high indicates the challenge in the *party* scenario. Certainly, more sophisticated algorithms are needed for head pose determination in such difficult conditions.

Table 4 presents the personality classification results derived from automated behavioral features. Note that since the same tracker is used in both DHLP and SVM+ MTL implementations and therefore, identical performance is achieved with Dist, Vel, Rel and the combination of these proxemic features. With most features (or combinations), classification performance obtained with automated estimates is lower than that obtained using manual annotations, especially for Extraversion. Slightly higher performance is obtained with attentional features with SVM+ MTL as compared to DHLP. Even with significant difference in head pose estimation accuracy using the two methods, the possible reason for only a marginal difference in performance could be the use of basic attentional features in our framework– we do not consider who the target is giving attention to/ receiving attention from, but only whether the target is giving/getting attention at a certain time instant. Nevertheless, the sharp dip in classification performance even with the use of these simple features reinforces why extensive research is necessary for achieving better results in this domain.

Concerning Neuroticism, results comparable to those obtained with manual annotations are obtained for some feature combinations. This is possibly because of the predictive power of proxemic features, *e.g.*, Dist, for which accurate estimates are still obtained using the tracker. However, the performance obtained with automatically computed Dist features is still poor. Overall, a best F-measure score of 0.51 and 0.64 is obtained for Extraversion and Neuroticism (as against 0.63 and 0.65 obtained with manual annotations). Consistent with the manual annotation results, in most cases, the best results with Dist, Vel features are obtained at smaller time slices, while the best results with Att features are observed with larger time slice durations. Also, unlike annotation-based results where LW-NB classifier performed best, most best results with automated features are obtained using Rand For and Lin SVM classifiers.

4. DISCUSSION AND CONCLUSION

In this work, we have studied the impact of head pose estimation errors on the computation of social attention features, and therefrom, prediction of the Extraversion and Neuroticism personality traits. We considered the *cocktail party* dataset for our analysis, which represents an unstructured social interaction scenario, where proxemic features provide vital cues in addition to social attention features for predicting targets' personalities (especially Neuroticism). The *party* scene also represents a challenging situation for head pose estimation due to unhindered motion of targets and variation in facial appearance with changing target position. In this respect, we considered two head pose estimation approaches– the DHLP, which simultaneously estimates target position and orientation, and SVM+ MTL, which learns a number of head pose classifiers corresponding to different room regions to account for position-based appearance variations, while also learning the face appearance relationships among these regions. While SVM+ MTL is found to estimate head pose better than DHLP, errors in social attention features computed using either approach adversely impact personality classification performance, especially for Extraversion. We believe that the unstructured and hedonistic *party* setting is better for analyzing personality-behavior correlates as compared to round table meetings, and more

accurate head pose estimation algorithms can immensely help in this regard.

Another aspect to note is that when we perform personality prediction from behavioral features aggregated over thin time slices, even the best results obtained with annotated features are only moderately good– this is also the case with other studies such as [18]. One possible reason for this phenomenon is that human behavior may change dynamically, but dynamic behavior is correlated with a *single* personality score in most studies. To this end, it may be interesting to model personality in terms of *personality states* [23], which refer to specific behavioral episodes, so that the personality trait is modeled as a distribution over personality states. Finally, as no audio data was available for the *cocktail party* dataset, annotations were solely based on visual cues. Determining the FoA purely based on visual data can be extremely hard, especially in *party* scene where a group of persons can interact while standing close to each other. In this regard, audio cues can help predict (and annotate) FoA better [25], and similarly, the use of head pose measurement sensors [22] and multi-sensor sociometric badges [17], which provide a variety of information such as the target's speaking time duration, extent of face-to-face interaction and physical proximity to others.

5. ACKNOWLEDGEMENTS

This work was partially supported by A*STAR Singapore under the Human Sixth Sense Program (HSSP) grant and EIT ICT Labs SSP 12205 Activity TIK- The Interaction Toolkit, tasks T1320A-T1321A. The authors would like to thank Francesco Tobia (TeV group, Fondazione Bruno Kessler) for developing the head pose annotation tool.

6. REFERENCES

[1] A. Airola, T. Pahikkala, W. Waegeman, B. D. Baets, and T. Salakoski. A comparison of auc estimators in small-sample studies. *Journal of Machine Learning Research - Proceedings Track*, 8:3–13, 2010.

[2] N. Ambady and R. Rosenthal. Thin slices? of expressive behaviors as predictors of interpersonal consequences. a meta analysis. *Psychological Bulletin*, 111:156–274, 1992.

[3] S. O. Ba and J.-M. Odobez. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, 39(1):16–33, 2009.

[4] D. R. Carney, C. R. Colvin, and J. A. Hall. A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41:1054–1072, 2007.

[5] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[7] S. De Julio and K. Duffy. Neuroticism and proxemic behavior. *Perception and Motor Skills*, 45(1):51–55, 1977.

[8] J. F. Dovidio and S. L. Ellyson. Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. *Social Psychology Quarterly*, 45(2):106–113, 1982.

[9] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Int'l conference on Knowledge Discovery and Data Mining*, pages 109–117, 2004.

[10] T. Evgeniou and M. Pontil. Regularized multi-task learning. In *ACM Int'l Conference on Knowledge Discovery and Data Mining*, 2004.

[11] E. Frank, M. Hall, and B. Pfahringer. Locally weighted naive bayes. In *Uncertainty in Artificial Intelligence*, pages 249–256, 2003.

[12] E. T. Hall. *The hidden dimension*. Anchor Books, 1963.

[13] H. Hung, D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *Int'l Conference on Multimodal Interfaces*, pages 233–236, 2008.

[14] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Trans. Audio, Speech and Lang. Proc.- Special issue on multimodal processing in speech-based interactions*, 17(3):501–513, 2009.

[15] S. R. Langton, R. J. Watt, and I. Bruce. Do the eyes have it? cues to the direction of social attention. *Trends in Cognitive Science*, 4(2):50–59, 2000.

[16] O. Lanz and R. Brunelli. Dynamic head location and pose from video. In *Int'l Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 47–52, 2006.

[17] B. Lepri, J. Staiano, G. Rigato, K. Kalimeri, A. Finnerty, F. Pianesi, N. Sebe, and A. Pentland. The sociometric badges corpus: A multilevel behavioral dataset for social behavior in complex organizations. In *Int'l Conference on Social Computing*, pages 623–628, 2012.

[18] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe. Connecting meeting behavior with extraversion - a systematic study. *IEEE Transactions on Affective Computing*, 3(4):443–455, 2012.

[19] L. Liang and V. Cherkassky. Connection between svm+ and multi-task learning. In *Int'l Joint Conference on Neural Networks*, 2008.

[20] J.-M. Odobez and S. O. Ba. A Cognitive and Unsupervised MAP Adaptation Approach to the Recognition of the Focus of Attention from Head Pose. In *Int'l Conference on Multi-Media & Expo*, 2007.

[21] M. Perugini and L. Di Blas. Analyzing personality-related adjectives from an eticemic perspective: the big five marker scale (bfms) and the italian ab5c taxonomy. *Big Five Assessment*, pages 281–304, 2002.

[22] A. K. Rajagopal, R. Subramanian, R. L. Vieriu, E. Ricci, O. Lanz, K. Ramakrishnan, and N. Sebe. An adaptation framework for head-pose classification in dynamic multi-view scenarios. In *Asian conference on Computer Vision*, pages 652–666, 2012.

[23] J. Staiano, B. Lepri, R. Subramanian, N. Sebe, and F. Pianesi. Automatic modeling of personality states in small group interactions. In *ACM Int'l conference on Multimedia*, pages 989–992, 2011.

[24] L. B. Statistics and L. Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

[25] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4):928–938, 2002.

[26] R. Subramanian, J. Staiano, K. Kalimeri, N. Sebe, and F. Pianesi. Putting the pieces together: multimodal analysis of social attention in meetings. In *Int'l Conference on Multimedia*, pages 659–662, 2010.

[27] M. Voit and R. Stiefelhagen. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *Int'l Conference on Multimodal interfaces*, pages 173–180, 2008.

[28] G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space speaks: towards socially and personality aware visual surveillance. In *ACM Int'l Workshop on Multimodal Pervasive Video Analysis*, pages 37–42, 2010.