Evaluating Multi-task Learning for Multi-view Head-pose Classification in Interactive Environments

Yan Yan[1], Ramanathan Subramanian[2], Elisa Ricci[3,4], Oswald Lanz[4], Nicu Sebe[1]
[1]University of Trento, Italy [2]Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore
[3]Department of Electrical and Information Engineering, University of Perugia, Italy [4]Fondazione Bruno Kessler, Trento, Italy

***Abstract*—Social attention behavior offers vital cues towards inferring one's personality traits from interactive settings such as *round-table meetings* and *cocktail parties*. Head orientation is typically employed as a proxy for determining the social attention direction when faces are captured at low-resolution. Recently, multi-task learning has been proposed to robustly compute head pose under *perspective* and *scale*-based facial appearance variations when multiple, distant and large field-of-view cameras are employed for visual analysis in smart-room applications. In this paper, we evaluate the effectiveness of an SVM-based MTL (SVM+MTL) framework with various facial descriptors (KL, HOG, LBP, *etc.*). The KL+HOG feature combination is found to produce the best classification performance, with SVM+MTL outperforming classical SVM irrespective of the feature used.**

I. Introduction

Social attention behavior, characterizing how a person attends to the immediate peers and surroundings during a social interaction, have been shown to be an extremely effective behavioral cue for decoding his/her personality traits in social phycology. For example, *extraverts* who tend to be gregarious and friendly, have been found to attract significant social attention from their peers in group meetings [1]. Similarly, *dominant* or aggressive persons are found to give more social attention to the other person while speaking, but less attention while listening in dyadic interactions [2].

Typically, two types of interactive settings have been studied by social psychologists to derive observations such as the above. *Round-table meetings* where participants are assessed as they discuss/enact a pre-defined situation as they are seated, or informal gatherings such as *cocktail parties*, where persons are free to move around and act spontaneously. Apart from high-level differences between these two scenarios (*e.g.*, a pre-defined agenda for meetings results in participants explicitly or implicitly assuming certain roles, which may preclude expression of their typical behavior, while parties involve hedonistic interactions allowing more freedom to express one's behavioral traits), they also differ with respect to the amount of observable visual information. Given the relatively stationary nature of meeting scenes, it is possible to closely observe participants' social attention patterns through webcams placed directly in their front, while the fact that they can move around freely in parties only allows for behavioral monitoring using distant, large field-of-view cameras installed on the walls of a smart-room setting.

Consequently, cues employed to determine one's social attention direction also vary for the two scenarios– when a high-resolution image of the face can be captured using near-field webcams (Fig. 1(a)), computing the point-of-gaze on top of the head orientation can improve social attention estimates as observed in [3]. On the other hand, when only low-resolution face images can be captured with far-field cameras, head pose is considered as a reliable proxy for determining the direction of social attention [4], [5]. Nevertheless, when only 50×50 or lower resolution faces are observable as in Fig. 1(b), estimating head pose, even with multiple cameras, is challenging due to the following reasons: (1) Blurred appearance of faces means discriminative features and classifiers need to be employed to obtain even a coarse head pose estimate (we consider eight class head pan classification in this paper, with each class denoting a 45° pan range), and (2) Facial appearance of a person (or target) with the same relative 3D head-pose, but at two different spatial locations varies considerably due to *perspective* and *scale* changes. As the target moves, the face can appear larger/smaller when it is closer to/farther away from the camera, and face parts can become occluded/visible due to the target's relative position with respect to the camera.

A number of methods have been recently proposed to estimate the head pose of moving targets, for which acquiring extensive training data is impractical or prohibitively expensive. Head orientation is determined based on the location of the face in the unfolded spherical texture map in [6]. A domain adaptation approach where patch weights denoting their saliency for pose classification is learnt from *source* examples corresponding to stationary targets, and adapted to the *target* scenario involving mobile targets using few *target* examples is proposed in [7]. A third approach proposed in [8] utilizes weakly-labeled data for model training by automatically obtaining head pose labels from walking direction labels, and trains multiple region-specific classifiers upon spatially segmenting the scene into multiple regions employing unsupervised spectral clustering.

A multi-task learning approach to robustly handle face appearance changes with target position was recently proposed in [9]. Based on the idea that there would some similarity in facial appearance for adjacent spatial regions as well as region-specific appearance differences, this method seeks to learn optimal partitioning of the physical space based on camera geometry-based and pose-based face appearance relationships, so that the facial appearance for a given pose is similar across each of the final partitions. Use of multi-task learning which learns relationships between a set of similar tasks is shown to produce superior classification performance than classical SVM and other state-of-the-art pose classification methods.

This paper seeks to provide the reader with information which is additional and complementary to that presented in [9]. The suitability and specificity of various descriptors for different computer vision tasks in well-known. While the superiority of multi-task learning against competing approaches

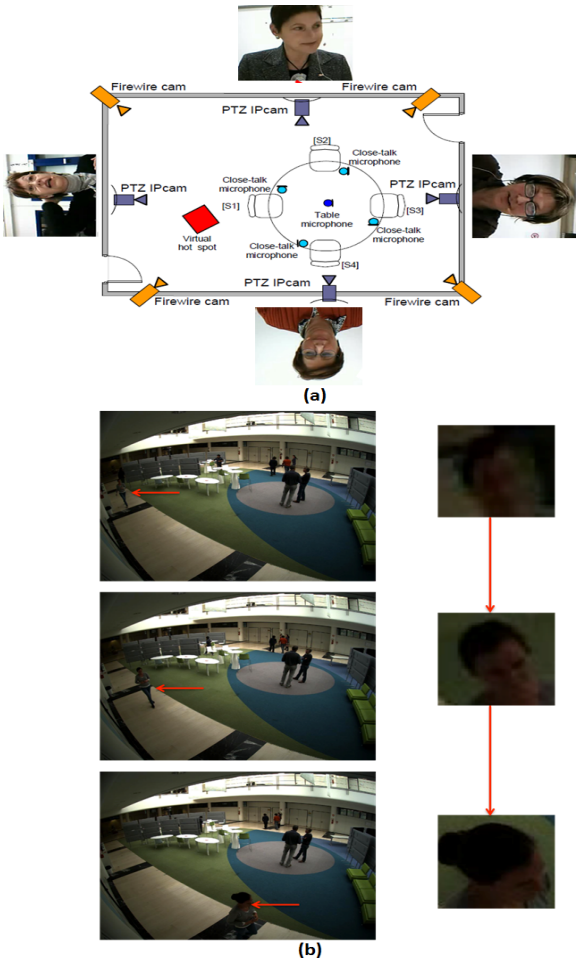

Fig. 1. Social attention analysis from *meeting* vs *party* scenes. (a) Overview of the sensing framework employed in [1] for analyzing *round-table meetings*. Webcams S1-S4 placed in front of each participant capture high resolution images from which head pose and point-of-gaze can be estimated. (b) On the other hand, face images acquired from far-field cameras monitoring *cocktail party* scenes are blurred, and facial appearance can vary for identical 3D head pose depending on the target position due to perspective and scale changes.

is demonstrated using HOG features in [9], we evaluate the performance of an SVM version of MTL (SVM+MTL) with a gamut of features including HOG, LBP, KL, skin color and their combinations. Our experimental results confirm the superiority of SVM+MTL over classical (or single-task) SVM irrespective of the descriptor used, with the best classification results obtained for the KL+HOG feature combination. Also, as demonstrated in [9], superior classification is achieved with multi-view features as compared to single-view features, and effectiveness of the MTL framework is most pronounced when only few training examples are available.

The paper is organized as follows. We review related work in Section II. Section III describes the SVM+MTL framework, while Section IV outlines the application of SVM+MTL for estimating the head pose of moving targets. Experimental results are discussed in Section V, while Section VI presents concluding remarks.

II. Related Work

This section presents an overview of previous works in the areas of (i) head-pose classification with low-resolution images and mobile targets, and (ii) multi-task learning.

A. Head-pose classification with low-resolution images

While head pose estimation from high-resolution images has been studied extensively over a decade [10], determining head orientation from surveillance images has been attempted only recently. Pose classification from crowded surveillance videos is presented in [11], where a Kullback-Leibler distance-based facial appearance descriptor (KL) is found to be more robust than explicit skin and hair segmentation for low-resolution images. The array-of-covariances (ARCO) descriptor is proposed in [12], and is found to be effective for representing objects at low-resolution and robust to scale and illumination changes. However, both these works address pose classification from monocular views, which is often insufficient for studying people's behavior in open spaces.

Works that have attempted pose classification fusing information from multiple views are [13], [6], [14], [15]. A particle filter is combined with two neural networks for pan and tilt classification in [13]. Also a HOG-based confidence measure is used to determine the relevant views for classification. In [14], multi-class SVMs are employed to calculate a probability distribution for head-pose in each view, and these results are fused to produce a more precise pose estimate. Nevertheless, both these works attempt to determine head-orientation as a person rotates in place, and position-induced appearance variations are not considered.

B. Head pose classification with moving targets

One of the first works to attempt head pose estimation with moving persons is [6], where face texture is mapped onto a spherical head model, and head-orientation determined based on the face location in the unfolded spherical texture map. However, many camera views are required to produce an accurate texture map, and nine cameras are used in this work. A transfer learning framework to compute head pose of moving persons along with a dataset of head orientation measurements for mobile targets is proposed in [7]. Upon learning the saliency of face patches for pose classification from many examples corresponding to stationary targets, these weights are *adapted* through online learning involving a few examples with moving targets. The adaptation procedure involves modulating the original patch saliency using a reliability score, which quantifies distortion in appearance of the face patch as a function of target position and the camera geometry. Upon *a priori* splitting the space into a number of regions, classifiers are independently learnt for each region given the dependence of patch reliability on target position– however, a pre-defined division of space and learning of independent classifiers is not necessarily optimal as shown in [9].

A scene-adaptive head pose estimator framework that does not explicitly require any training data annotation is proposed in [8]. First, labeled head pose data is automatically generated from the output of a person tracker denoting walking direction, and in order to tackle appearance variation with target position, the scene space is automatically segmented into multiple regions employing unsupervised spectral clustering, which generates clusters with high intra-cluster and low inter-cluster similarity. Finally, independent, region-wise pose classifiers are

Fig. 2. SVM+MTL-based head pose classification for mobile targets: In the pre-processing stage, a particle filter tracker incorporating multi-view geometry information is employed to reliably localize targets' faces and extract face crops for each view. Also, the tracker allows for determining which region in space the test sample corresponds to, so that the appropriate region-based pose classifier can be invoked. Features extracted from the multi-view face appearance images are fed to the MTL module for training/classification. Classification is attempted with a number of pixel and patch descriptors in this work.

learnt to achieve improved classification accuracy– again, joint learning of classifiers is shown to be more beneficial in [9]. In [9], a multi-task learning (MTL) framework for head pose classification of mobile targets is proposed. Joint learning of facial appearance across scene regions as well as region-specific appearance variations using a graph-guided MTL framework is shown to produce optimal pose classification performance. The learning algorithm is guided by two graphs *a priori* defining appearance similarity between (1) scene regions based on camera geometry, and (2) head pose classes, to flexibly learn the optimal space partitioning (set of related tasks). In this paper, we evaluate the effectiveness of an alternative MTL formulation based on SVMs (SVM+MTL), which assumes that all tasks are related.

C. Multi-task learning

Multi-task learning (MTL) represents a specific instance of learning with structured data (LWSD), where the training data is a union of t *related* groups. The difference between LWSD and MTL is that the group membership of the test sample needs to be *known a-priori* for MTL, while it need not be provided for LWSD. When the test sample's group membership is known (or can be determined), MTL assumes learning of the t data groups to be equivalent to learning t related tasks, and seeks to learn a decision function that minimizes the expected loss for each task.

MTL seeks to simultaneously learn the commonality [16] as well as the differences between the t tasks which leads to a better model for the main task (union of t tasks) as compared to a learner that does not consider task relationships. Convex multi-task feature learning imposing a trace norm regularization is proposed in [17]. Single-task (standard) SVM is extended through a regularization framework in [18], where the classifier comprises a common decision function and a task-specific correction function. However, the decision and correcting functions are in the same feature space. This model is improved to a more flexible form where the decision and correction spaces are different in the SVM+MTL framework proposed in [19]. We novelly apply the SVM+MTL framework for head-pose estimation, and a detailed algorithm description is provided in the following section.

III. The SVM+MTL framework

An MTL framework extending support vector machines (SVMs) to the situation where the training set $\mathcal{T}$ is the the union of task specific sets $\mathcal{T}_r = \{\boldsymbol{x}_{ir}, y_{ir}\}_{i=1}^{l_r}$ is presented in [18]. For each task the learned weights vector is decomposed as $\boldsymbol{w} + \boldsymbol{w}_r,\ r \in (1, 2, ..., t)$ where $\boldsymbol{w}, \boldsymbol{w}_r$ respectively model the commonality between tasks and task specific components. Following [18], the SVM+MTL framework is proposed [19]. The associated optimization problem is formulated as follows:

$$\begin{aligned}\min_{\boldsymbol{w},\boldsymbol{w}_1,\ldots,\boldsymbol{w}_t,b,d_1,\ldots,d_t} \ & \tfrac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \tfrac{\beta}{2}\sum_{r=1}^{t}\boldsymbol{w}_r^T\boldsymbol{w}_r + C\sum_{r=1}^{t}\sum_{i=1}^{l_r}\xi_{ir} \\ \text{s.t.} \ & y_{ir}(\boldsymbol{w}^T\phi(\boldsymbol{x}_{ir}) + b + \boldsymbol{w}_r^T\phi_r(\boldsymbol{x}_{ir}) + d_r) \geq 1 - \xi_{ir}, \\ & \xi_{ir} \geq 0,\ i = 1, \ldots, l_r,\ r = 1, \ldots, t\end{aligned}$$

Here, all $\boldsymbol{w}_r$'s and the common $\boldsymbol{w}$ are learnt simultaneously. β regularizes the relative weights of $\boldsymbol{w}$ and $\boldsymbol{w}_r$'s. ξ_{ir}'s are slack variables measuring the errors $\boldsymbol{w}_r$'s make on the t data groups, each comprising l_r training samples. y_{ir}'s denote training labels while C regulates the complexity and proportion of nonseparable samples.

The goal of SVM+MTL is to find t decision functions $f_r(\boldsymbol{x}) = \boldsymbol{w}^T\phi(\boldsymbol{x}) + b + \boldsymbol{w}_r^T\phi_r(\boldsymbol{x}) + d_r, r = 1, \ldots, t$. Each decision function f_r comprises two parts: (a) the common weights vector $\boldsymbol{w}$ with bias term b, and the group-specific correction function $\boldsymbol{w}_r$ with bias term d_r. SVM+MTL [19] improves over regularized MTL [18] on two counts: (i) In regularized-MTL, the common and the task specific functions share the same feature space (ϕ), while they may be different in SVM+MTL and (ii) SVM+MTL considers a more general form of the decision function with bias terms (b, d_r).

Introducing the Lagrangian multipliers $\boldsymbol{\alpha}, \boldsymbol{\mu}$, the dual objective function is:

$$\begin{aligned}L(\boldsymbol{\alpha}, \boldsymbol{\mu}) = \ & \tfrac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \tfrac{\beta}{2}\sum_{r=1}^{t}\boldsymbol{w}_r^T\boldsymbol{w}_r + C\sum_{r=1}^{t}\sum_{i=1}^{l_r}\xi_{ir} \\ & + \sum_{r=1}^{t}\sum_{i=1}^{l_r}\alpha_{ir}[1 - \xi_{ir} - y_{ir}(\boldsymbol{w}^T\phi(\boldsymbol{x}_{ir}) + b + \boldsymbol{w}_r^T\phi_r(\boldsymbol{x}_{ir}) + d_r)] \\ & - \sum_{r=1}^{t}\sum_{i=1}^{l_r}\mu_{ir}\xi_{ir}\end{aligned}$$

To eliminate the primal variables from the dual form, we set the Lagrangian derivative with respect to $\boldsymbol{w}, \boldsymbol{w}_r, b, d_r$ to zero. The dual form of the optimization problem then becomes:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\mu}} L(\boldsymbol{\alpha},\boldsymbol{\mu}) = \sum_{i=1}^{l}\alpha_i - \tfrac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_{ir}y_{jr}\phi(\boldsymbol{x}_{ir})\phi(\boldsymbol{x}_{jr}) - \tfrac{1}{2\beta}\sum_{r=1}^{t}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_{ir}y_{jr}\phi_r(\boldsymbol{x}_{ir})\phi_r(\boldsymbol{x}_{jr})$$
$$s.t. \quad \sum_{i=1}^{l_r}\alpha_i y_{ir} = 0,\ r = 1,\ldots,t,\ \alpha_i + \mu_i = C, \quad i = 1,\ldots,l_r\ \alpha_i \geq 0, \mu_i \geq 0$$

The multi-task SVM is a quadratic programming (QP) problem. We adopt generalized sequential minimal optimization (SMO) [19] to solve this optimization problem. Based on Karush-Kuhn-Tucker (KKT) conditions, $\boldsymbol{w}$, $\boldsymbol{w}_r$ can respectively be expressed in terms of training samples as:

$$\boldsymbol{w} = \sum_{i=1}^{l}\alpha_i y_{ir}\phi(\boldsymbol{x}_{ir}),\ \boldsymbol{w}_t = \frac{1}{\beta}\sum_{i=1}^{l_r}\alpha_{ir}y_{ir}\phi_r(\boldsymbol{x}_{ir})$$

The decision function for task r is:

$$f_r(\boldsymbol{x}) = \sum_{i=1}^{l}\alpha_i y_{ir}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}) + b + \frac{1}{\beta}\sum_{i=1}^{l_r}\alpha_i y_i\phi_r(\boldsymbol{x}_i)\phi_r(\boldsymbol{x}) + d_r$$

IV. Head-pose classification under varying target position

An overview of head-pose classification usng SVM+MTL is presented in Fig. 2. Similar to previous works such as [7], [9], the proposed approach relies on a multi-view, color-based particle filter [20] to compute the position of a target and obtain an accurate estimate of the head location. A dense 3D grid (of size 30 cm×30 cm×20 cm with 1cm resolution) of hypothetic head locations is placed around the estimated 3D head-position provided by the particle filter as in [9]. Assuming a spherical model of the head, a contour likelihood is then computed for each grid point by projecting a 3D sphere onto each view employing camera calibration information. The grid point with the highest likelihood is then determined as the face location. It is worth noting that accurate face localization is crucial in the considered scenario where targets can move around freely.

Following localization, the face is cropped and resized to 20 × 20 pixels in each view. A second phase involves the computation of multi-view features that can effectively describe the facial appearance for pose classification. In each view, the face image is divided into $n_x \times n_y$ overlapping patches (we consider $n_x = n_y = 8$ and a step-size of 4 between two patches in our experiments), and corresponding patch descriptors are computed. In this work, we specifically the following features:

- Histogram of Oriented Gradients (HOG) [21] are popular descriptors for low-resolution head-pose estimation. In our experiments, 144-dimensional (16 blocks × 9 bins) HOG features are used.
- Kullback-Leibler Divergence (KL) features [11] used as similarity distance maps by indexing each pixel with respect to the mean appearance templates of different head pose classes.
- Local Binary Pattern (LBP) [22], which is a simple and efficient texture operator that labels image pixels using binary values, by thresholding them with respect to their neighborhood. In our experiments, 256-dimensional (16 cells × 16 bins) LBP histogram features are used.
- Skin color pixels which constitute an important cue for headpose estimation in low-resolution images, and have been widely used in previous works [23]. We first detect skin region in images using a Gaussian mixture-based skin color model. We divide images into 4 × 4 cells and count the number of skin pixels in each cell. Then, a 16-dimensional feature vector is constructed.

Once the features are computed for each view, the associated descriptors are concatenated and fed to a classifier for head pose estimation. In this paper, we divide the scene space into t (=4) regions and use SVM+MTL to train t region-based pose classifiers. At test time, the tracker allows for determining which region in space the test sample corresponds to, so that the appropriate region-based pose classifier can be invoked. The advantage of adopting a MTL strategy is intuitive: we expect that even if only few training samples are available for each region, the use of MTL can compensate by transferring discriminative information from one region (task) to the others. This allows for enhanced classification performance as confirmed by our experimental evaluation presented in the following section.

V. Experiments and Results

This section presents experimental results (a) to evaluate the performance of SVM+MTL against classical/single-task SVM and (b) to investigate the impact of different features and training set sizes on classification performance. For our evaluation, we consider the publicly available DPOSE database [7] (Fig. 3). This dataset has been compiled for benchmarking head-pose classification performance under target motion, and contains over 50000 4-view synchronized images of 16 moving targets with ground-truth head-pose measurements acquired using an accelerometer, gyrometer, magnetometer platform. Our task is to assign the test head pose to one of eight classes, each denoting a quantized 45^o (360/8) head-pan.

To train the MTL classifier, we divided the room into four quadrants (Fig. 3) and for each quadrant, requisite number of training samples per class were randomly selected. Presented results denote the mean classification accuracy obtained with five such randomly chosen training sets.

The linear kernel $K(x_i, x_j) = x_i^T x_j$ was used for modeling the shared space among tasks, while the Gaussian kernel $K_G(x_i, x_j) = \exp(-|x_i - x_j|^2/2\sigma^2)$ was employed for modeling the task-specific feature space. The optimal values of parameters C, β and σ were determined upon tuning from $\{0.01, 0.1, 1, 10, 100\}$. Also, we used the 'one-vs-all' method for extending binary classification to multi-class.

The first experiment involved training the SVM-MTL with 30 images/class/region and testing with images from all regions. This is equivalent to testing the classifier on the union of t data groups. Fig.4 presents classification accuracies achieved with (i) SVM+MTL, (ii) single-task (standard) SVM and (iii) regularized MTL (rMTL) [18] employing different features. For rMTL, the linear kernel is used to map training data onto the common and task specific spaces. From the figure, we can observe the following:

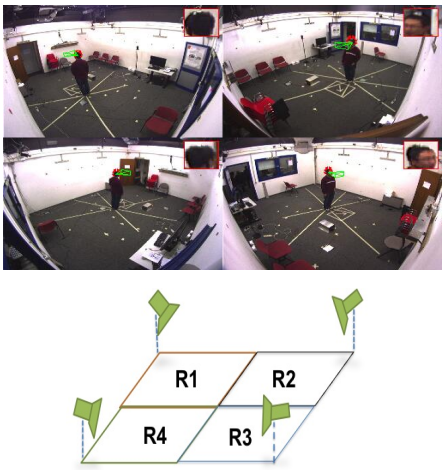

Fig. 3. Exemplar image from the DPOSE dataset (top). Green colored cone denotes head pose direction, while automatically extracted face crops are shown on the top-right insets. Considered scene space division (bottom)

- The $KL + HOG$ combination produced the highest classification accuracy. Among individual features, KL and LBP descriptors were respectively found to be most and least effective.
- SVM+MTL achieved the best classification in all-but-one cases, implying that having different common and specific feature spaces is beneficial.
- Interestingly, performance gains with rMTL and SVM+MTL were higher for those features for which standard SVM performed poorly. For example, with the HOG descriptor, classification accuracy improved from 55% to 80% with SVM+MTL. This observation demonstrates that learning the commonality as well as differences between related tasks is beneficial as compared to learning them independently, and echoes what is reported in other multi-task learning works such as [9].

Table I compares classification performance obtained with single-task SVM (STS) and multi-task SVM (MTS) when the MTS was trained with samples from all regions (30 samples/class/region $\times$ 8 classes $\times$ 4 regions), while the test set comprised images exclusively acquired from one of R1-R4. For fair comparison, we trained the STS with only images arising from the same region as the test set. This was to evaluate the performance of MTL for the r^{th} task, $r = 1, \ldots, t$. Here, MTS outperformed STS comfortably with performance improving by over 20% in some cases, thereby demonstrating the power of MTL to efficiently learn task-specific data differences. The highest classification accuracies were again achieved with $KL + HOG$ features, for which MTS consistently outperformed STS by over 10%. We also compared MTS with the state-of-the-art ARCO [12] pose classifier. As for STS, ARCO was trained exclusively with the (test) region-specific images. ARCO produced comparable classification performance only when trained with 12 dimensional covariance features incorporating pixel locations (x, y), color (R, G, B), Gabor coefficients at four orientations, gray-scale gradients I_x, I_y and gradient orientation OG.

Finally, the effects of varying the training set size and using only one of the four camera views are presented in Fig. 5, 6 respectively. We gradually increased the training set size

TABLE I. PERFORMANCE COMPARISON OF SINGLE-TASK SVM (STS), MULTI-TASK SVM (MTS) AND ARCO [12] WHEN ALL TEST IMAGES ARISE FROM ONE QUADRANT (FIG. 3(B)).

	R1		R2		R3		R4	
	STS	MTS	STS	MTS	STS	MTS	STS	MTS
HOG	59.8	**78.4**	60.1	**81.6**	63.9	**78.6**	60.3	**78.5**
LBP	66.6	**71.4**	69.5	**74.9**	70.0	**71.4**	60.9	**67.5**
KL	76.6	**85.1**	77.3	**89.1**	84.8	**87.4**	73.2	**84.0**
HOG+SKIN	75.7	**84.8**	83.6	**88.2**	77.8	**85.9**	77.2	**85.1**
LBP+SKIN	72.4	**83.4**	80.9	**87.3**	78.0	**86.3**	74.4	**83.6**
KL+HOG	76.6	**86.3**	79.8	**89.1**	83.2	**89.2**	74.8	**85.2**
ARCO	86.2		89.2		87.3		89.0	

from 1-30 samples/class/region and computed classification performance on the test set comprising samples from all regions (as for the first experiment). For training set sizes of less than 5 samples/class/region, STS slightly outperformed MTS. This is because MTS requires sufficient samples to effectively learn the commonalities and differences between the tasks- learning is affected when there are too few samples. Nevertheless, when the training size was increased further, MTS comfortably outperformed STS and the performance gap increased with larger training sizes.

We also analyzed how the ARCO descriptor [12] performed with varying training set sizes. Sophisticated methods like ARCO also require enough samples to learn an effective model. *E.g.*, when 5 training samples/class/region were used, random classification accuracy was obtained with ARCO as against 75% with STL/MTL. This demonstrates that multi-task learning effectively works with small training sets. Figure 6 presents accuracies obtained with $KL + HOG$ features computed for all four views, as against features computed for only one of the four views (mean of the accuracies obtained with each of the four views is considered here). Expectedly, higher classification accuracy is obtained with the four-view features. Also, the performance gain with MTS over STS is higher when single view features are used for training- performance gains of over 10% and 20% are respectively observed with MTS with four-view and single-view features.

Summarizing the observed results, the proposed SVM+MTL framework outperforms both single-task SVMs and regularized MTL irrespective of the facial descriptor used, implying that it is important to consider both appearance similarity across regions, and region-specific appearance variations while attempting head pose classification under changing appearance due to target motion. Nevertheless, comparable accuracies are achieved with the SVM+MTL and ARCO approaches, when 12-dimensional covariance features are used. Superior classification accuracies are achieved using the MTL framework described in [9] with respect to ARCO. In this respect, a key difference between SVM-MTL and the FEGA-MTL proposed in [9] is that FEGA-MTL attempts to *learn the set of related tasks*, whereas the SVM+MTL formulation assumes that all tasks are related. If some tasks are actually unrelated, knowledge sharing can negatively impact model performance.

VI. CONCLUSION

While *Cocktail parties* allow for more naturalistic social behavior in comparison to *round-table meetings*, they are challenging for analyzing social attention behavior from visual

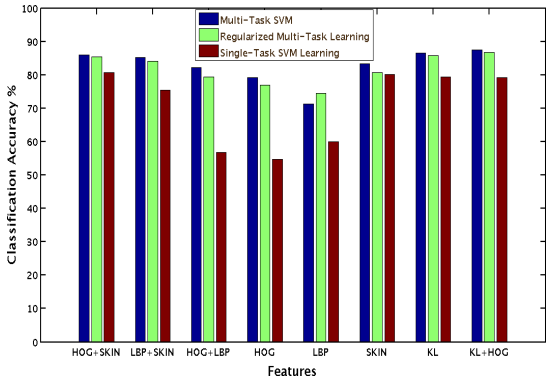

Fig. 4. Head-pose classification accuracy obtained with different methods employing various features.

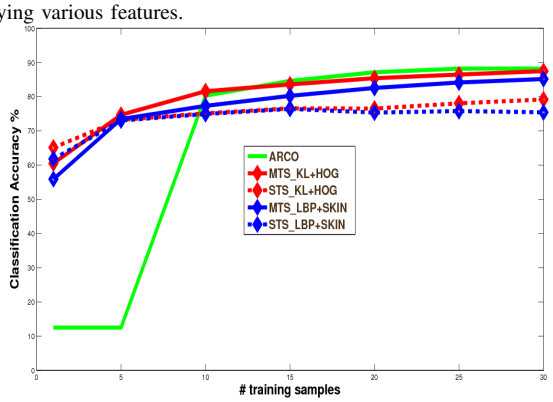

Fig. 5. Variation in classification performance with varying training set sizes.

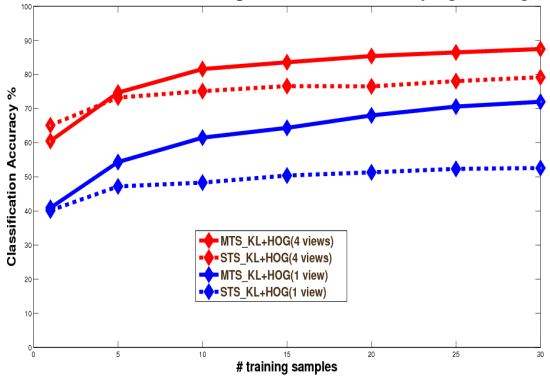

Fig. 6. Classification accuracies obtained with single and four-view features.

cues. The SVM+MTL framework is shown to robustly estimate head orientation under target motion characteristic of party scenes, which can then be used to determine social attention direction in behavioral studies. Future work involves extending SVM+MTL for flexibly learning related tasks as in [9], and exploiting pre-created models as in [7] to obviate model re-synthesis for novel data.

Acknowledgements: This work was partially supported by EIT ICT Labs SSP 12205 Activity TIK - The Interaction Toolkit, tasks T1320A-T1321A; EU FP7 xLIME and PERTE project; A*STAR Singapore under the Human Sixth Sense Program (HSSP) grant.

REFERENCES

[1] 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*, vol. 3, no. 4, pp. 443–455, 2012.

[2] 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*, vol. 45, no. 2, pp. 106–113, 1982.

[3] 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*, 2010, pp. 659–662.

[4] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," vol. 13, no. 4, pp. 928–938, 2002.

[5] R. Subramanian, Y. Yan, J. Staiano, O. Lanz, and N. Sebe, "On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions," in *ACM Int'l Conference on Multimodal Interfaces*, 2013.

[6] X. Zabulis, T. Sarmis, and A. A. Argyros, "3d head pose estimation from multiple distant views." in *British Machine Vision Conference*, 2009.

[7] 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*, 2012, pp. 652–666.

[8] I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, and A. Sugimoto, "Head direction estimation from low resolution images with scene adaptation," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1502 – 1511, 2013.

[9] 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 *Int'l Conference on Computer Vision*, 2013.

[10] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 607–626, 2009.

[11] J. Orozco, S. Gong, and T. Xiang, "Head pose classification in crowded scenes," in *British Machine Vision Conference*, 2009, pp. 1–11.

[12] D. Tosato, M. Farenzena, M. Cristani, M. Spera, and V. Murino, "Multi-class classification on riemannian manifolds for video surveillance," in *European Conference on Computer Vision*, 2010, pp. 378–391.

[13] M. Voit and R. Stiefelhagen, "A system for probabilistic joint 3d head tracking and pose estimation in low-resolution, multi-view environments," in *Computer Vision Systems*, 2009, pp. 415–424.

[14] R. Muñoz-Salinas, E. Yeguas-Bolivar, A. Saffiotti, and R. M. Carnicer, "Multi-camera head pose estimation," *Mach. Vis. Appl.*, vol. 23, no. 3, pp. 479–490, 2012.

[15] Y. Yan, R. Subramanian, O. Lanz, and N. Sebe, "Active transfer learning for multi-view head-pose classification," in *Int'l Conference on Pattern Recognition*, 2012.

[16] Y. Yan, G. Liu, E. Ricci, and N. Sebe, "Multi-task linear discriminant analysis for multi-view action recignition," in *Int'l Conference on Image Processing*, 2013.

[17] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Neural Information Processing Systems*, 2007.

[18] M. P. Theodoros Evgeniou, "Regularized multi-task learning," in *ACM International Conference on Knowledge Discovery and Data Mining*, 2004.

[19] L. Liang and V. Cherkassky, "Connection between svm+ and multi-task learning," in *International Joint Conference on Neural Networks*, 2008.

[20] O. Lanz, "Approximate bayesian multibody tracking." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1436–1449, 2006.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[22] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *International Conferencec on Computer Vision*, 2009, pp. 32–39.

[23] I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, and A. Sugimoto, "Appearance-based head pose estimation with scene-specific adaptation," in *ICCV Workshops*, 2011, pp. 1713–1720.