ABSTRACT

This work proposes a boosting-based transfer learning approach for head-pose classification from multiple, low-resolution views. Head-pose classification performance is adversely affected when the source (training) and target (test) data arise from different distributions (due to change in face appearance, lighting, etc.). Under such conditions, we employ Xferboost, a Logitboost-based transfer learning framework that integrates knowledge from a few labeled target samples with the source model to effectively minimize misclassifications on the target data. Experiments confirm that the Xferboost framework can improve classification performance by up to 6%, when knowledge is transferred between the CLEAR and FBK four-view headpose datasets.

Index Terms— Multi-view headpose classification, low-resolution, Xferboost, boosting-based transfer learning

1. INTRODUCTION

Despite much progress in head-pose estimation and tracking (see [1] for a detailed survey), most methods and benchmarking datasets [2, 3, 4] focus on determining pose from high-resolution imagery. However, recent works have actively attempted head-pose recovery from surveillance videos [5, 6, 7, 8] where faces are blurred and are at low resolution. These approaches classify head-pose to one of many discrete classes denoting a range of orientations.

This paper deals with head-pose classification as a person is imaged with multiple, large field-of-view cameras in a closed setting. Also, we seek to adapt existing models derived from available data to new situations through transfer learning. Figure 1 shows multiple instances of persons captured in two distinct settings. Images on the left are from the CLEAR07 [9] head-pose dataset, which contains around 27000 4-view images with pose ground-truth. Likewise, 4-view images from the FBK dataset captured under different camera, illumination and environmental settings are shown on the right. Head-pose comprises pan and tilt1, which denote out-of-plane horizontal and vertical head rotation, and examples with downward, frontal and upward head-tilt are respectively shown in the top, middle and bottom rows of Figure 1.

The rest of this paper discusses how models learnt from extensive source (CLEAR) data can be adapted to effectively work on novel target (FBK) data. As a preliminary step, we divided the CLEAR, FBK data into three parts corresponding to downward ([−90°, −20°]), frontal ([−20°, 20°]) and upward ([20°, 90°]) tilt, and then attempted eight-class head-pan classification (Figure 2) fixing the head-tilt. Apart from simplifying the pose-labeling problem to pan classification, fixing the head-tilt range allowed us to explore the adaptation problem under realistic settings. E.g., how labeled head-pan examples acquired from boardroom meeting scenes (where head-tilt is typically frontal) can be utilized to determine what attracts people’s attention in a supermarket/museum setting (where downward/upward head-tilt is generally expected).

Secondly, we trained a state-of-the-art ARCO descriptor [7] model for each of the source subsets, and tested these models on the different source and target subsets. Training and test data sizes for the considered source/target subsets are specified in Table 1. Table 2 lists classification accuracies obtained with the ARCO models for these combinations. It is evident from the tabulated results that while high accuracies are observed when training and test set distributions are the same (i.e., same dataset, similar head-tilt), a significant drop in performance is observed even as the training and test set distributions vary. This is the case when (i) the face appearance changes due to head-tilt differences (even for the same dataset) and (ii) training and test data attributes vary (as for CLEAR and FBK).

To counter this problem, we propose Xferboost, a boosting-

1We ignore roll (in-plane head rotation) here
based transfer learning approach. Transfer learning [10, 11] allows for adaptation of models learnt from available source data to novel target data, using additional knowledge from a few labeled target samples. Xferboost integrates Tradaboost [11], which ‘tunes’ the source model to the target data by assigning greater importance to target samples, with the Logitboost learner employed by ARCO. Experimental results reveal that this tuning is more effective than simply learning a model with many source and few target samples, and can improve classification performance by more than 6%.

In summary, this paper represents one of the first works to explore a transfer-learning approach for multi-view head-pose classification. The next section evaluates related work to motivate the proposed approach, while Section 3 discusses Xferboost in detail. Experimental results are presented in Section 4 and we end with concluding remarks in Section 5.

2. RELATED WORK

We now review related work in (i) head-pose classification from low-resolution images and (ii) transfer learning.

2.1. Pose classification from low-resolution views

Recent and popular head-pose classification algorithms that work on low-resolution images are [8, 7]. In [8], a Kullback-Leibler (KL) distance-based facial appearance descriptor is found to be effective for pose classification on the i-LIDS dataset comprising footage of an underground scene. In [7], array-of-covariance (ARCO) descriptors, robust to scale/lighting variations and occlusions, produce 11% better classification on i-LIDS as compared to [8]. Combining a dynamic Bayesian network with Gaussian mixture-cum-Hidden Markov models, a visual focus-of-attention (VFOA) estimation algorithm for multiple subjects moving in front of a surveillance camera is proposed in [5]. However, all these works address head-pose estimation from a single view.

Among multi-view pose-estimation works, a robust approach to positional variations is proposed in [6], where face texture is mapped onto a spherical head model, and head-pose is determined from the face location on the unfolded texture map. Nevertheless, many cameras are required to generate an accurate texture map, while we explore a purely image-based approach for multi-view head-pose classification.

2.2. Transfer learning approaches

There are several approaches to transfer learning. Instance-transfer [11] involves reuse of source data in a related target domain assuming that certain parts of the source data are still useful in the target scenario. Feature-representation-transfer [12] involves finding a ‘good’ feature representation that reduces differences between the source and target data. Parameter-transfer [13] involves discovery of shared parameters or priors between the source and target models which can benefit from transfer learning. Transfer learning approaches have become very popular in computer vision- a transferable distance function is learned with sparse training data for action detection in [14]. We propose a transfer learning for pose classification in this work.

3. LOGITBOOST-BASED TRANSFER LEARNING

This section describes in detail, (i) the pre-processing steps involved (ii) the array of covariance descriptors (ARCO) algorithm and (iii) Xferboost, the proposed Logitboost-based transfer learning algorithm for head-pose classification.

3.1. Pre-processing

As large field-of-view cameras are used to acquire both source (CLEAR) and target (FBK) datasets, the first step involves facial appearance extraction in each of the camera views. To this end, we employ a multi-view color-based particle filter [15] which can handle multiple, moving subjects without manual initialization. Upon estimating the 3D body-centroid and height of the moving subject(s) with the tracker, particles are sampled around the 3D head-position within a search window. Assuming a spherical model of the head, a contour likelihood is computed for each particle by projecting a 3D
sphere onto each view employing camera calibration information. Finally, the sample with the highest likelihood sum is determined as the head location. Upon face localization, the face crop is resized to $20 \times 20$ resolution prior to ARCO feature computation. Facial appearance extraction process is outlined in Figure 3.

3.2. Array of covariance descriptors (ARCO)

The state-of-the-art ARCO algorithm [7] employs covariance features, robust to occlusions as well as scale and lighting variations, for head-pose classification from low-resolution images. Upon dividing the image into a number of overlapping patches, ARCO computes covariance-based patch descriptors. Subsequently, a multi-class Logitboost classifier is learnt for each patch, and the test sample is assigned a label based on majority vote of the patch-based classifiers.

The ARCO algorithm has many advantages. Firstly, covariance matrices are flexible, low-dimensional features. A requisite number of image features can be combined to generate covariance descriptors, which can effectively describe visual objects at prohibitively low resolutions. Also, each patch descriptor is only a $d \times d$ matrix, where $d$ denotes the number of image features used. This can be further reduced to a $d(d+1)/2$ dimensional vector upon projecting covariance descriptors. Subsequently, a multi-class Logitboost classifier is learnt for each patch, and the test sample is assigned a label based on majority vote of the patch-based classifiers.

The main contribution of this work is that we seek to adapt an existing model derived from many source training samples to novel target data, using additional knowledge from a few target training samples and minimizing the effort required to label target samples in the process. ARCO employs a multi-class Logitboost learner (strong classifier) $\{F_j\}$ for each image patch, comprising $l = 1..L$ weak classifiers. Given a training set $\{x_i\}$ with $N$ samples corresponding to class labels $1..J$, the Logitboost algorithm iteratively re-weights training samples most difficult to classify, through a set of weights $w_i$ and posterior probabilities, $P_j(x_i)$. Each weak learner solves a weighted-regression problem, whose goodness of fit is measured by the response value vector for the $i^{th}$ training sample, $z_i = \{z_{ij}\}_{j=1}^J$.

The Logitboost learner learns until most training samples are classified correctly. Therefore, when presented with a training set containing many source and few target samples, the model could still be source-oriented, given varying attributes of the source and target. Instead, we adopt the methodology employed in Tradaboost [11], which prioritizes misclassified target samples in the boosting framework, so that the resulting model is ‘tuned’ to the target.

Given $N + M$ training data, with $N$ source (Src) and $M$ target (Tgt) samples, where $N >> M$, the Xferboost algorithm proceeds as follows. At every step, upon normalizing $w_i$‘s, the error on target, $\epsilon_t$ ($\epsilon_t < 0.5$) is computed for the misclassified target samples. Also, $\alpha_s$ and $\alpha_t$, which are respectively the attenuating and boosting factors for misclassified source and target samples, are determined. Finally, weights of misclassified target data are boosted by a factor of $e^{\alpha_t}$, so that more target-specific information can be learned, while misclassified source weights are attenuated by a factor of $e^{-\alpha_s}$ to discourage learning of these samples. The proposed Xferboost algorithm is summarized in Algorithm 1.
mments for various poses were recorded for 16 subjects using an accelerometer, gyro, magnetometer platform. The FBK data differs from CLEAR with respect to distance of cameras from the person, relative camera positions and illumination conditions. The FBK dataset contains over 25000 examples (Table 1), out of which 50 random samples were used for training while the remainder were used for testing.

We compared Xferboost accuracies against the Logitboost learner fed with both source and target training data (baseline/no Xferboost condition). We analyzed the effect of varying the number of weak learners $L$ and target training set size (with 5-30 target samples/class) on classification performance (Figure 4). Each point on the graph denotes mean accuracy obtained from five independent trials (employing randomly generated target training sets) for the given condition. Also, since we used multiple views for pose classification, we compared performance considering (i) one-view accuracy or the mean accuracy obtained using only one of the 4 views and (ii) four-view accuracy, the accuracy obtained upon feeding features from all four views to the classifier.

Notice from Figure 4 that much higher accuracies are obtained employing features from all views instead of only one view, implying that multi-view information is more robust compared to single-view for head-pose classification on low-resolution images. Higher improvements in classification performance are obtained with Xferboost when a) fewer patch learners (implying less computation resources) and b) fewer target training samples are employed. Also, higher improvements are obtained with Xferboost when the source and target distributions vary significantly (e.g., CLEAR up- FBK down combination), as compared to cases where they are similar (e.g., CLEAR up- FBK up). Target-specific information is most beneficial when the source and target have minimal similarity, and transfer learning works best in such cases.

Table 3 presents the best improvements obtained with Xferboost when only 5 labeled target samples/class (0.5% of the target size) are employed for transfer learning. In the fourth and sixth columns, the Xferboost accuracies are presented along with the baseline accuracies (within parentheses). Here again, the maximum improvements with Xferboost are obtained for those cases where the source and target distributions vary significantly. Also, single-view improvements are higher as compared to employing all 4 views, suggesting that transfer learning perhaps works better when less information is available. Overall, a maximum performance gain of 6.2% is obtained with the Xferboost approach for the CLEAR down- FBK up combination.

## 5. CONCLUSIONS

The paper proposes Xferboost, a boosting-based transfer learning approach for pose classification from multiple, low-resolution views. Experimental results confirm that the effectiveness of Xferboost, which improves classification performance significantly when the source and target distributions are very different. Future work involves integrating information from multiple sources for transfer learning, and exploiting temporal constraints for efficient head-pose tracking.

## 6. ACKNOWLEDGEMENTS

This work was supported by the FIRB S-PATTERNS project. Radu L. Vieriu’s research was supported by EURODOC “Doctoral Scholarships for research performance at European level”, ID 59410.

## 7. REFERENCES


[5] K. Smith, S.O. Ba, J.M. Odobez, and D. Gatica-Perez, “Tracking the visual focus of attention for a varying
Fig. 4. Variation in classification accuracy with the number of learners (top) and target samples (bottom). Results presented for Src-Tgt combinations CLEAR down-FBK up (a,d), CLEAR frontal-FBK down (b,e) and CLEAR up-FBK up (c,f).

Table 3. Best improvements obtained with Xferboost for different combinations. L denotes number of weak learners.

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>L</th>
<th>Accuracy (1-view)</th>
<th>% gain</th>
<th>Accuracy (4-view)</th>
<th>% gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEAR down</td>
<td>FBK down</td>
<td>8</td>
<td>42.7 (41.3)</td>
<td>3.4</td>
<td>66.5 (64.5)</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>FBK frontal</td>
<td>9</td>
<td>40.8 (38.6)</td>
<td>5.8</td>
<td>65.5 (62.7)</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td>FBK up</td>
<td>9</td>
<td>51.5 (48.5)</td>
<td>6.2</td>
<td>81.2 (78.7)</td>
<td>3.2</td>
</tr>
<tr>
<td>CLEAR frontal</td>
<td>FBK down</td>
<td>10</td>
<td>40.1 (39.1)</td>
<td>2.5</td>
<td>61.9 (60.8)</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>FBK frontal</td>
<td>8</td>
<td>54.1 (52.3)</td>
<td>3.5</td>
<td>78.8 (77.1)</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>FBK up</td>
<td>8</td>
<td>63.7 (62)</td>
<td>2.7</td>
<td>87 (86)</td>
<td>1.1</td>
</tr>
<tr>
<td>CLEAR up</td>
<td>FBK down</td>
<td>10</td>
<td>40 (38.3)</td>
<td>4.5</td>
<td>59.7 (57.7)</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>FBK frontal</td>
<td>8</td>
<td>58.1 (57.3)</td>
<td>1.4</td>
<td>80.6 (80.1)</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>FBK up</td>
<td>9</td>
<td>69.3 (68.9)</td>
<td>0.7</td>
<td>88.8 (88.6)</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 3. Best improvements obtained with Xferboost for different combinations. L denotes number of weak learners.


