MATCHING CROSS-RESOLUTION FACE IMAGES USING CO-TRANSFER LEARNING

Himanshu S. Bhatt[*]*, Richa Singh and Mayank Vatsa*

IIIT-Delhi, India

Nalini Ratha

IBM T. J. Watson Research Center, USA

ABSTRACT

Face recognition systems, trained in controlled environment, often fail to efficiently match low resolution images with high resolution images. In this research, a *co-transfer learning* framework is proposed in which knowledge learnt in controlled high resolution environment is transferred for matching low resolution probe images with high resolution gallery. The proposed framework seamlessly combines *transfer learning* and *co-training* to perform knowledge transfer by updating classifier's decision boundary with low resolution probe instances. Experiments are performed on the CMU Multi-PIE and SCface database with gallery images of size 72×72 and size of probe images varying from 48×48 to 16×16. The results show that, in terms of rank-1 identification accuracy, the proposed algorithm outperforms existing approaches by at least 5%.

Index Terms— Low resolution face recognition, Transfer learning, Co-training, SVM

1. INTRODUCTION

With advancements in technology, surveillance cameras now have a profound presence and are widely used for security and law enforcement applications. These cameras are primarily designed to have a wide coverage from a fixed location and may yield very low resolution face images. The need to identify individuals from such low resolution images has emerged as a new covariate in face recognition. It involves matching low resolution probe images (obtained from surveillance cameras) with high resolution gallery images captured during enrollment, as shown in Fig. 1(a) and (b). The difference in information content between high resolution and low resolution images degrades the performance of existing face recognition algorithms.

One approach to match cross resolution images, i.e. low resolution probe with high resolution gallery, is to downsample high resolution images to the level of low resolution images before matching. However, information useful for face recognition such as texture, edges, and other high frequency information is compromised while downsampling the images. Another widely used approach is to enhance the low resolution face images using super-resolution techniques [1], [2] and then match with high resolution images. Super-resolution techniques are intended for reconstructing a high resolution view from low resolution image(s) and are not optimized for face recognition applications. Though there are few techniques that incorporate face recognition with super-resolution [3], they remain susceptible to environmental variations and introduce distortions as shown in Fig. 1(c). Recently, few approaches are proposed to map features from images with different resolutions into a unified space to minimize the difference between low-resolution and high-resolution images [4], [5].

[*]Research is supported through IBM PhD Fellowship.

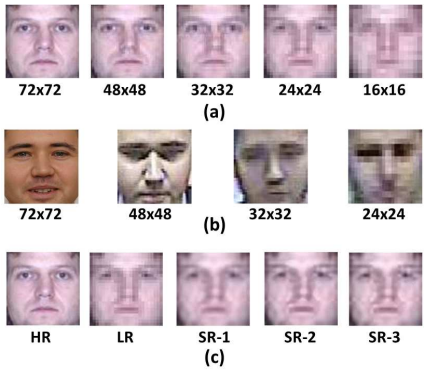

Fig. 1. Images at different resolutions from the (a) CMU Multi-PIE, (b) SCface database, and (c) comparing images enhanced using super-resolution techniques [1], [2].

Another related challenge pertains to training in controlled environment with high resolution (HR) images and testing in uncontrolled environment with low resolution (LR) images. The conditions in which a system is trained are referred to as source domain where the availability of large training data helps the system to efficiently learn the task. The conditions in which the system operates are referred to as target domain. In the source domain, high resolution probe images are matched with high resolution gallery whereas in the target domain, low resolution probe images are matched with high resolution gallery. To address this challenge, knowledge learned in the source domain is transferred to perform efficient matching in the target domain. Zhao and Hoi [6] proposed a framework for online transfer learning, where labeled training instances in the target domain are available incrementally. However, their approach requires labeled instances for the supervised learning task and obtaining labeled instances may be expensive, time consuming, and requires human effort. In this research, co-training [7] is utilized to facilitate transfer learning with unlabeled probe instances. The main contribution of this research is a *co-transfer learning framework that integrates transfer learning with co-training to efficiently match low resolution (LR) probe with high resolution (HR) gallery images*.

2. CO-TRANSFER LEARNING FRAMEWORK

Consider a scenario where there is large labeled data pertaining to high resolution images but only a few labeled instances of low resolution i are available from the target domain. The source domain classifier is well trained on large labeled data; however, the target domain classifier is trained only on a few labeled examples. Since the classifier in target domain has not seen adequate data during training, it has to learn the decision boundary from unlabeled probe instances

in the target domain. For this task, both the classifiers are combined in an ensemble to subsequently transfer the knowledge from source to target domain classifier. Further, two ensembles trained on separate views (here, it represents two feature extractors) provide pseudo labels to probe instances in the target domain. Within each ensemble, decision boundary of the target domain classifier can be updated in online manner with pseudo-labeled instances obtained during testing.

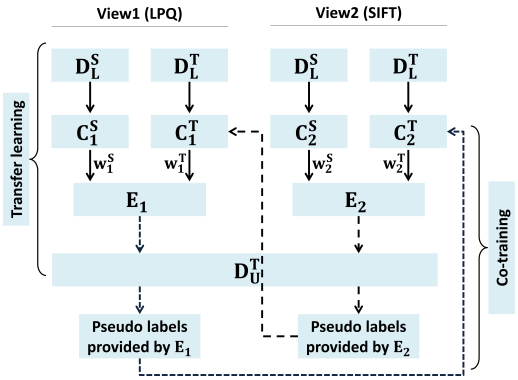

Fig. 2. Block diagram for the co-transfer learning framework.

As shown in Fig. 2, source domain classifier, C_j^S where $j = 1, 2$ represents the views (features), is trained using sufficient HR labeled data denoted by D_L^S=$\{(\mathbf{u}_1^S, z_1), (\mathbf{u}_2^S, z_2), ..., (\mathbf{u}_n^S, z_n)\}$. Every i^{th} instance, $\mathbf{u}_i$ has two views $\{x_{i,1}, x_{i,2}\}$ for the label $z_i \in \{-1, +1\}$; here $x_{i,1}$ and $x_{i,2}$ represent the input vector obtained from two separate views. Target domain classifier (C_j^T) is initially trained on a few labeled training instances in the target domain represented as D_L^T=$\{(\mathbf{u}_1^T, z_1), (\mathbf{u}_2^T, z_2), ..., (\mathbf{u}_m^T, z_m)\}$. Here, n and m are number of training instances in the source and target domains respectively. A set of r unlabeled probe instances in the target domain is represented as D_U^T=$\{(\mathbf{u'}_1^T), (\mathbf{u'}_2^T), ..., (\mathbf{u'}_r^T)\}$. Next, an ensemble prediction function, denoted as E_j, is constructed for each view. E_j is a weighted combination of source domain classifier, C_j^S, and target domain classifier, C_j^T, with w_j^S and w_j^T as the two weights. For the i^{th} unlabeled probe instance in the j^{th} view, ensemble function E_j predicts the label, $E_j(x_{i,j}) \rightarrow y_{i,j}$. In an ensemble, knowledge is transferred by updating the decision boundary of target domain classifier C_j^T using only the new incremental data as proposed in [8]. For the i^{th} instance in the target domain $\mathbf{u'}_i$, class label is predicted by the ensemble as given in Eq. 1.

$$y_{i,j} = w_{i,j}^S \, C_j^S(\mathbf{u'}_i) + w_{i,j}^T \, C_j^T(\mathbf{u'}_i) \tag{1}$$

where $w_{i,j}^S$ and $w_{i,j}^T$ are the weights for the source and target domain classifiers at the i^{th} instance. Initially, w_j^S and w_j^T are set to 0.5 such that each classifier contributes equally within an ensemble. Gradually, the two weights are adjusted to emphasize the contribution from the updated target domain classifier in an ensemble. As proposed by Zhao and Hoi [6], the two weights are updated dynamically as shown in Eqs. 2 and 3.

$$w_{i+1,j}^S = \frac{w_{i,j}^S g_i(C_j^S(\mathbf{u'}_i))}{w_{i,j}^S g_i(C_j^S(\mathbf{u'}_i)) + w_{i,j}^T g_i(C_j^T(\mathbf{u'}_i))} \tag{2}$$

$$w_{i+1,j}^T = \frac{w_{i,j}^T g_i(C_j^T(\mathbf{u'}_i))}{w_{i,j}^S g_i(C_j^S(\mathbf{u'}_i)) + w_{i,j}^T g_i(C_j^T(\mathbf{u'}_i))} \tag{3}$$

where $w_{i+1,j}^S$ and $w_{i+1,j}^T$ are the updated weights, g_i is defined as:

$$g_i(s) = \exp\{-\eta \, l(y_i, \hat{y}_i)\}, \tag{4}$$

$\eta = 0.5$, $l(y, \hat{y}) = (y - \hat{y})^2$ is the square loss function, y is the predicted label, and $\hat{y}$ is the pseudo label provided by co-training (explained later).

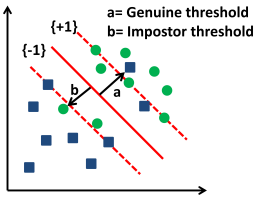

Fig. 3. Illustrates the confidence of prediction.

The transfer learning approach requires labeled instances in the target domain. Obtaining labeled training instances is a difficult and expensive task, however, large number of unlabeled instances are available as probe. Therefore, in this research, unlabeled probe instances are leveraged to transfer the knowledge learned in the source domain. Co-training, proposed by Blum [7], is used to provide pseudo labels to unlabeled probe instances available sequentially in the target domain. It assumes the availability of two ensemble functions E_1 and E_2 trained on separate views where each ensemble function has sufficient (better than random) accuracy. If one ensemble confidently predicts genuine label for an instance while the other ensemble predicts impostor label with low confidence, then this particular instance is used as a re-training sample for the second ensemble or vice-versa. In this research, confidence of prediction for an instance on the j^{th} view, denoted by α_j, is measured as the distance from the decision boundary as shown in Fig 3. A genuine threshold is computed as the distance of the farthest impostor point that is erroneously classified as genuine. For predicting an instance to be confident enough to lie in genuine class, the distance from the decision hyperplane should be greater than the corresponding threshold (P_j). Similar procedure is repeated for impostors as well. In this manner, unlabeled probe instances are transformed into pseudo labeled training data and the decision boundary of target domain classifier is updated in online manner [8]. The proposed *co-transfer learning* framework is summarized in Algorithm 1.

The proposed algorithm is particularly useful in recognizing cross-resolution face images. In this research, the proposed *co-transfer learning* framework is applied on support vector machine (SVM) classifiers. Further, two feature extractors, 1) Local Phase Quantization (LPQ) [9], and 2) Scale Invariant Feature Transform (SIFT) [10] are used to represent two separate views of face images across different resolutions. These feature extractors are resilient to scale changes and can be effectively used for matching face images with varying resolutions. Face image is tessellated into non-overlapping facial regions, features are extracted, and the distance is computed for each facial region using χ^2 distance. For a gallery-probe pair, the distances between local facial regions are vectorized and provided as input to the ensemble. SVMs in each ensemble are trained using the approach proposed by Phillips [11] and the final performance is computed by combining responses from both the ensembles.

Algorithm 1 Co-transfer learning framework

Input: Initial labeled training data D_L^S in the source domain, a few labeled instances D_L^T in the target domain. Unlabeled probe instances D_U^T (available sequentially).
Iterate: j= 1 **to** 2 (number of views)
Process: Train classifiers C_j^S and C_j^T on j^{th} view of D_L^S and D_L^T respectively to construct ensemble E_j. Compute confidence thresholds P_j for each view.
for $i = 1$ **to** r (number of probe instances) **do**
  Predict labels: $E_j(x_{i,j}) \rightarrow y_{i,j}$; α_j represents confidence of prediction
  if $\alpha_1 > P_1$ & $\alpha_2 < P_2$ **then**
    Update C_2^T with pseudo-labeled instance $\{x_{i,2}, y_{i,1}\}$ & recompute w_2^S and w_2^T.
  end if.
  if $\alpha_1 < P_1$ & $\alpha_2 > P_2$ **then**
    Update C_1^T with pseudo-labeled instance $\{x_{i,1}, y_{i,2}\}$ & recompute w_1^S and w_1^T.
  end if.
end for.
end iterate.
Output: Updated classifiers C_j^T and weights w_j^S, w_j^T.

3. EXPERIMENTAL EVALUATION

To evaluate the efficacy of the proposed framework, a joint transfer-and-test strategy is used which allows the data used in model adaptation to be concurrently used for performance evaluation. Further, two databases, (1) CMU Multi-PIE[1] and (2) SCface[2] are used for performance evaluation. For experiments on the CMU Multi-PIE database, images pertaining to 337 individuals with frontal pose and neutral expression are selected. Classifiers in the source domain are trained on high resolution images of 100 subjects and classifiers in the target domain are trained on low resolution probe and high resolution gallery images of 40 (from the 100) subjects. Performance is evaluated on images of the remaining 237 individuals with four different resolutions of probe images, as shown in Fig. 1(a). For each subject, one high resolution image is kept in the gallery and one low resolution image is used as probe. Images at a particular resolution are obtained by downsampling the original images to the required resolution (varying from 72×72 to 16×16) using bi-cubic interpolation. The second database is the SCface database that comprises images corresponding to 130 individuals captured in uncontrolled indoor environment using five video surveillance cameras placed at three different distances. For experiments on the SCface database, classifiers in the source domain are trained on high resolution images corresponding to 50 subjects and classifiers in the target domain are trained on low resolution probe and high resolution gallery images corresponding to 20 (from the 50) subjects. The performance is evaluated on images corresponding to 80 individuals with three resolutions of probe images corresponding to three different distances, as shown in Fig. 1(b). For each subject, one high resolution image is kept in gallery and five images corresponding to five different cameras are used as probe. In all experiments, resolution of gallery is set to 72×72. The experiments resemble real world scenario where ample high resolution images are available in the source domain. However, only a few low resolution probe and the corresponding high resolution gallery images are available for training in the target domain. Performance is reported with 10 times repeated random sub-sampling for non-overlapping training-testing partitions.

[1]http://www.multipie.org/
[2]http://www.scface.org/

Table 1. Rank-1 accuracy of different algorithms.

Probe resolution		**48×48**	**32×32**	**24×24**	**16×16**
Database	**Algorithm**				
CMU Multi-PIE, Gallery (72×72)	Ensemble1	79.4%	69.1%	61.8%	56.7%
	Ensemble2	76.3%	55.2%	59.4%	52.1%
	Fusion	86.1%	79.4%	70.3%	66.2%
	Proposed	**92.3%**	**84.1%**	**77.4%**	**72.4%**
SCface, Gallery (72×72)	Ensemble1	63.2%	58.1%	52.6%	NA
	Ensemble2	60.4%	57.8%	49.1%	NA
	Fusion	74.4%	67.4%	60.2%	NA
	Proposed	**79.4%**	**72.8%**	**66.4%**	NA

The performance of the proposed approach is compared with fusion of two ensembles trained on the initial data when no knowledge is transferred (referred to as *'fusion'*). It allows to analyze the performance gain due to transfer of knowledge. The performance is also compared with three super-resolution techniques. Super-resolution-1 is the standard bi-cubic interpolation [2], super-resolution-2 [3] is a regression based technique proposed by Kim and Kwon [1] , and super-resolution-3 [4] is a sparse representation based approach proposed by Yang *et al.* [2].[5]

Figs. 4(a) and (b) show the rank-1 identification performance of the proposed approach with probe images of different resolutions on the CMU Multi-PIE and SCface databases respectively. Table 1 compares the rank-1 accuracy of the proposed approach with the accuracy of individual ensembles as well as their fusion without transfer learning. The results show that there is an improvement of about $5 - 7\%$ with transfer of knowledge from high resolution to low resolution domain. CMC curves in Fig. 4(c) show that the proposed approach also outperforms all three super-resolution techniques by at least 11%. Super-resolution is performed with a magnification factor of three to match probe images of size 24×24 with 72×72 gallery images, as shown in Fig. 1(c). LPQ and SIFT features are extracted from enhanced images and the performance is computed using sum-rule fusion of LPQ and SIFT match scores. Fig. 5(a) shows some examples where transferring the knowledge from source to target domain classifier helped to correctly recognize low resolution probe images (correctly identified in rank-10). Examples in Fig. 5(b) show some of the cases where the proposed approach performed poorly. The poor performance can be attributed to the fact that some of the pseudo labels assigned to unlabeled probe instances may be incorrect leading to *negative transfer*. However, the effect of *negative-transfer* can be minimized by intelligently selecting the confidence threshold for the prediction. High threshold values imply conservative transfer while smaller values of the threshold lead to aggressive transfer. In the proposed framework, updating decision boundary of the target domain classifier within an ensemble allows to capture the knowledge transferred from high resolution domain. Initially, equal weights were assigned to both the classifiers in an ensemble; however with transfer of knowledge, weights for classifiers in the target domain become more prominent. For example, in experiments with the CMU Multi-PIE database, C_1^T and C_2^T were updated

[3]http://www.mpi-inf.mpg.de/ kkim/.
[4]http://www.ifp.illinois.edu/ jyang29/.
[5]For experiments with super-resolution techniques, code, pre-trained dictionary and parameters are used as provided in the online implementation.

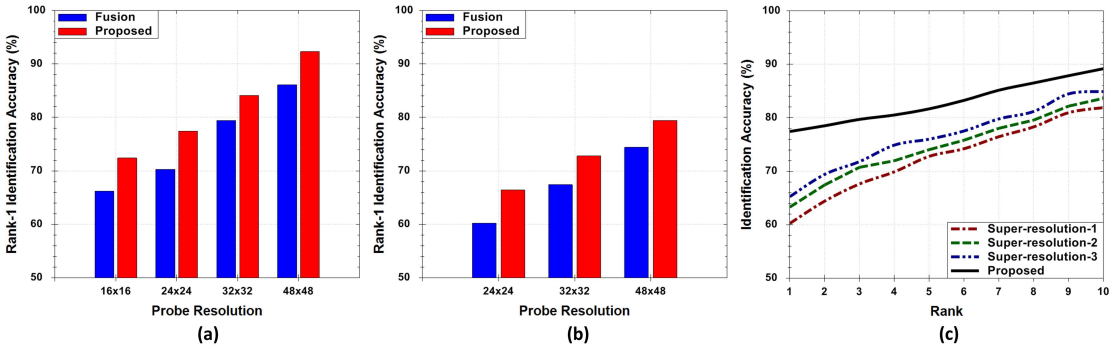

Fig. 4. Rank-1 accuracy at different probe resolution for the (a) CMU Multi-PIE database, (b) SCface database, and (c) comparison with super-resolution techniques for matching LR images (24×24) with HR images (72×72) from the CMU Multi-PIE.

on $5,184$ and $4,210$ pseudo labeled probe instances respectively. It is observed that 96.43% of the total pseudo labels were correct. The weights for classifiers in each ensemble also saturate at w_1^S=0.18, w_1^T=0.82, w_2^S=0.23, and w_2^T=0.77 at the end of co-transfer. In experiments with the SCface database, C_1^T and C_2^T were updated on $7,346$ and $5,268$ pseudo labeled probe instances respectively. It is observed that 94.43 % of the total pseudo labels were correct. The weights for classifiers in each ensemble also saturate at w_1^S=0.21, w_1^T=0.79, w_2^S=0.27, and w_2^T=0.73. These numbers correspond to the experiments with probe size 24×24 and gallery size 72×72. The experimental results suggest that the proposed approach efficiently matches cross-resolution face images by leveraging knowledge from the source domain. It also validates our assertion that co-training enables updating the decision boundary of the target domain classifiers with unlabeled probe instances as and when they arrive.

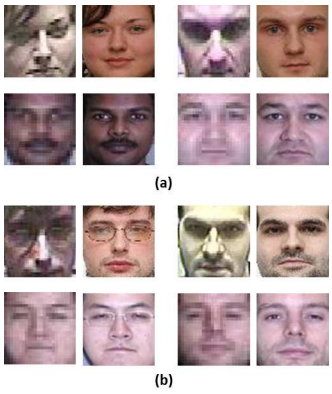

Fig. 5. Illustrating sample cases when the proposed approach (a) correctly recognizes, (b) fails to recognize. All examples are with probe size 24×24 and gallery size 72×72.

4. CONCLUSION

This paper presents a co-transfer learning framework in which knowledge from high resolution domain is transferred to perform efficient face matching in low resolution domain. An ensemble is constructed from weighted combination of classifiers in the source and target domains. In an ensemble, decision boundary of the classifier in the target domain is updated to accommodate knowledge transferred from the classifier in the source domain. Moreover, updating the weights assigned to each classifier also facilitates gradual shift of knowledge from the source to the target domain classifier. In addition, co-training is used to provide pseudo labels to unlabeled probe instances to seamlessly improve the performance by leveraging the knowledge learnt in the source domain. The proposed co-transfer learning framework provides significant improvements in performance as compared to other algorithms.

5. REFERENCES

[1] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE TPAMI*, vol. 32, no. 6, pp. 1127–1133, 2010.

[2] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE TIP*, vol. 19, no. 11, pp. 2861–2873, 2010.

[3] P. H. Hennings-Yeomans, S. Baker, and V. Bhagavatula, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *CVPR*, 2008, pp. 1–8.

[4] B. Li, H. Chang, S. Shan, and X. Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE SPL*, vol. 17, no. 1, pp. 20–23, 2010.

[5] S. Biswas, K. Bowyer, and P. Flynn, "Multidimensional scaling for matching low-resolution face images," *IEEE TPAMI*, 2012 (In press).

[6] P. Zhao and S. Hoi, "OTL: A framework of online transfer learning," in *ICML*, 2010, pp. 1231–1238.

[7] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *CoLT*, 1998, pp. 92–100.

[8] H. S. Bhatt, S. Bharadwaj, R. Singh, M. Vatsa, A. Ross, and A. Noore, "On co-training online biometric classifiers," in *IJCB*, 2011, pp. 1–6.

[9] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using local phase quantization," in *ICPR*, 2008, pp. 1–4.

[10] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[11] P. J. Phillips, "Support vector machines applied to face recognition," in *NIPS*, 1999, pp. 803–809.