PERSON-SPECIFIC DOMAIN ADAPTATION WITH APPLICATIONS TO HETEROGENEOUS FACE RECOGNITION

Yao-Hung Tsai[1,2], *Hung-Ming Hsu*[1,2], *Cheng-An Hou*[2], *and Yu-Chiang Frank Wang*[2]

[1]Dept. Electrical Engineering, National Taiwan University, Taipei, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

ABSTRACT

Heterogeneous face recognition (HFR) is a practical yet challenging task in which gallery and probe face images are collected in terms of different modalities or features (e.g., sketch vs. photo). In this paper, we present a person-specific domain adaptation framework for HFR. By utilizing the subjects not of interest (i.e., those not to be recognized), we first derive a common feature space using their cross-domain face images, with the goal of eliminating differences between image modalities. To generalize our feature space for representing and recognizing the subjects of interest, we advocate the construction of person-specific domain adaptation model in this space, so that the classifiers (trained by the gallery images) are able to achieve satisfactory recognition performance. In our experiments, we consider sketch-to-photo and near-infrared (NIR) to visible spectrum (VIS) face recognition problems for evaluating the effectiveness of our method.

Index Terms— Domain adaptation, heterogeneous face recognition

1. INTRODUCTION

Traditional pattern recognition problems typically deal with training and test data collected from the same feature domain. However, when these data are collected from different feature domains or exhibiting distinct feature distributions, the classifiers learned from training data cannot be easily generalized to recognize test data. Thus, how to solve such *cross-domain* recognition problems becomes a very challenging task.

Heterogeneous face recognition (HFR) is an example of cross-domain recognition problems (e.g., [1, 2, 3]). HFR is an emerging task in biometrics, since real-world face images to be recognized are typically acquired in different modalities (e.g., photos in visible spectrums (VIS), near-infrared (NIR), or even sketches). For example, the face image of a suspect might be captured by a camera in NIR mode at nights. In order to perform recognition, one needs to match such probe (test) images to the gallery (training) ones which are in visible spectrums (i.e., photos). Since standard matching or recognition algorithms are not expected to perform well for the above cases, several approaches have been proposed for determining a common/joint feature space using cross-domain face images, so that gallery and query images can be projected onto the derived space for recognition. For example, Yi *et al.* [4] and Sharma *et al.* [5] applied canonical correlation analysis (CCA) and Partial Least Squares (PLS) frameworks to observe such feature spaces, respectively. To further improve the representation ability for such feature spaces, techniques of dictionary learning were also integrated into the above learning process [6, 7]. Aiming at introducing additional data separation capability, Klare and Jain [8, 3] proposed to learn discriminative projections using multiple visual features observed from cross-domain data.

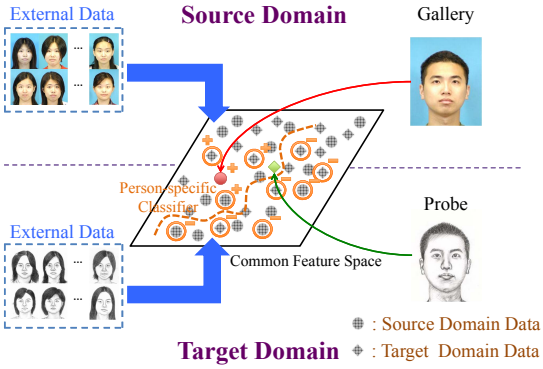


Fig. 1. Illustration of person-specific domain adaptation for HFR. Note that external face images (in gray) are collected from the subjects not of interest, while the gallery and query images of the subjects to be recognized are in green and orange, respectively.

Although the aforemention approaches have reported promising recognition results, their direct use of *external* cross-domain face images (i.e., those from the subjects not of interest) might not be preferable. For example, one cannot expect that the common feature space observed from the face images of females will generalize well to those of males. Another concern is that, most prior work on HFR require the external data *pairs* and/or their *label* information when relating cross-domain data. In this paper, as depicted in Figure 1, we advocate the learning of person-specific domain adaptation model for HFR, which observes common fea-

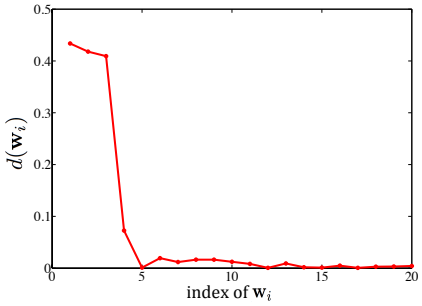
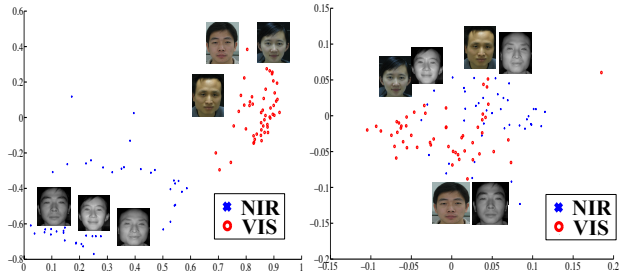


Fig. 2. Distances $d(\mathbf{w}_i)$ between projected cross-domain data using different $\mathbf{w}_i$ (by (1)) on the CASIA NIR-VIS 2.0 dataset.



(a) (b)

Fig. 3. Cross-domain face images of the CASIA NIR-VIS 2.0 dataset projected onto the subspaces spanned by the top two eigenvectors determined by (a) PCA and (b) DiCA, respectively. Note that the data variation due to domain changes is disregarded in our DiCA subspace, which makes NIR-to-VIS recognition more applicable.

ture spaces by eliminating image domain differences, while person-specific classifiers can be learned in this space using the gallery images. It is worth noting that, we not only disregard the data correspondence constraint when collecting external data, we do not require any label information when associating different image modalities. We will detail our proposed method in the following section.

2. OUR PROPOSED METHOD

2.1. DiCA for Domain Adaptation

Recall that cross-domain classification deal with training and test data collected from different domains or with distinct distributions. Transfer learning [9] is recently advanced by researchers in related fields, with the goal to transfer knowledge learned from one domain to the other. Among various scenarios of transfer learning, *domain adaption* focuses on solving the same learning task across different domains.

As observed in [10], face images of different subjects but in the *same* modality are highly correlated, while those of the same subjects but across different modalities typically exhibit significantly larger variations. In this paper, we advance *Domain-independent Component Analysis* (DiCA) [11] for associating face images across different modalities, aiming at describing the distribution of cross-domain face data in the same feature subspace. As noted earlier, we require external cross-domain data to construct a subspace for relating different image domains. Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_{n_{Ext}}] \in \mathbb{R}^{d \times n_{Ext}}$ as the external data matrix in a d-dimensional space, in which each instance $\mathbf{x}_i$ is collected from either the source and target domain. We have $n_{Ext} = n_s + n_t$ indicating the total number of external images (where n_s and n_t are the image numbers in the source D_s and target domains D_t, respectively).

It can be seen that, no data pair constraint nor label information is required for external data $\mathbf{X}$. With the centering matrix $\mathbf{C} = \mathbf{I} - \frac{1}{n_{Ext}}\mathbf{1}$, the covariance matrix of $\mathbf{X}$ can be formulated as $\mathbf{XCX}^\top$. Next, we derive an orthogonal projection matrix $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_k] \in \mathbb{R}^{d \times k}$ for $\mathbf{X}$ ($k < d$):

$$\max_{\mathbf{W}^\top \mathbf{W} = \mathbf{I}} tr(\mathbf{W}^\top \mathbf{XCX}^\top \mathbf{W}), \quad (1)$$

It can be seen that the above optimization process effectively performs Principal Component Analysis (PCA) on the collected cross-domain external data. In other words, one can apply standard eigen-decomposition techniques and reformulate the above problem as $\mathbf{XCX}^\top \mathbf{W} = \mathbf{WV}$, where $\mathbf{V} = diag(v_1, v_2, ..., v_k) \in \mathbb{R}^{k \times k}$ is a diagonal matrix, in which the diagonal elements indicate the associated eigenvalues in a descending order.

Based on the observation of [10] (and later verified by Figures 2 and 3), the first few dominant eigenvectors derived by $\mathbf{X}$ would correspond to domain differences instead of describing the data distribution itself in either domain. Therefore, once all k PCA eigenvectors for $\mathbf{X}$ are obtained, our next step is to identify the most dominant ones among these eigenvectors, which correspond to the domain variations instead of data distributions. In our work, we advance the distance measurement based on *Maximum Mean Discrepancy* (MMD) [12, 13] in the resulting PCA subspace, so that the similarity between instances collected from different domains at each subspace dimension can be determined accordingly.

Now, for the ith dimension in the PCA subspace (i.e., projected by $\mathbf{w}_i$), we measure the distance between the means of the projected data from source and target domains:

$$d(\mathbf{w}_i) = \| \frac{1}{n_s} \sum_{x_j \in D_s} \mathbf{w}_i^\top \mathbf{x}_j - \frac{1}{n_t} \sum_{x_k \in D_t} \mathbf{w}_i^\top \mathbf{x}_k \|. \quad (2)$$

With $d(\mathbf{w}_i)$ calculated for each dimension, one can identify the dominant dimensions which correspond to the domain changes. Take Figure 2 for example, the separation between projected cross-domain data in the first four dimensions of the resulting PCA subspace is clearly larger than that of the remaining dimensions, and thus the first four dimensions of this subspace mainly describe the domain variations (instead of the distribution of face images). In other words, we can simply apply a threshold T for identifying/disregarding the eigenvectors $\mathbf{w}_i$ which are with the largest $d(\mathbf{w}_i)$ values.

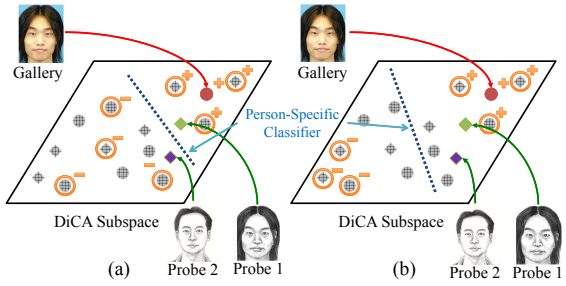


Fig. 4. Learning person-specific classifiers using projected data from subjects not of interest as negative samples. (a) Negative samples are randomly selected (i.e., our strategy), and (b) projected data which are farther away from gallery (and thus positive) ones are selected as negative samples. Comparing these two sub-figures, it can be seen that the strategy of (b) would introduce more false alarms (e.g., Probe 2) and thus decrease the recognition performance.

On the other hand, we consider the remaining eigenvectors $\mathbf{w}_i$ (with smaller $d(\mathbf{w}_i)$) as *Domain-independent Components* (DiC), since they correspond to the data distribution itself (instead of the domain changes). We derive DiC by:

$$DiC(\mathbf{X}) = \{\mathbf{w}_j | \mathbf{w}_j \subseteq columns(\mathbf{W}), d(\mathbf{w}_j) \leq T\},$$
$$\mathbf{W}_{DiCA} = [\hat{\mathbf{w}}_1, ..., \hat{\mathbf{w}}_{\tilde{k}}] \in \mathbb{R}^{d \times \tilde{k}} \; and \; \hat{\mathbf{w}}_i \in DiC(\mathbf{X}), \forall i, \quad (3)$$

where $\tilde{k}$ is the number of eigenvectors $\mathbf{w}_i$ whose corresponding $d(\mathbf{w}_i)$ are smaller than threshold T. The collection of such eigenvectors results in the final projection matrix $\mathbf{W}_{DiCA}$, which will be applied to transform source or target domain data into the DiCA subspace for domain adaptation purposes (see Figure 3 for example).

2.2. Learning of Person-Specific Classifiers

From Section 2.1, it can be seen that our DiCA is able to eliminate cross-modality variations for face images. Although one can simply project gallery and probe images (of the subjects of interest) into the DiCA subspace and perform recognition directly, it is worth repeating that our DiCA model is derived from *external* face images. In other words, it is not clear whether the observed subspace is able to sufficiently represent the face images of the subjects to be recognized.

Since we need to generalize our DiCA subspace for recognizing the subjects of interest, we propose to construct *person-specific* domain adaptation models, so that improved representation and discriminating capabilities can be introduced. Inspired by [14], we consider the gallery images of the subjects to be recognized (projected into the DiCA subspace) as positive samples, and we randomly sample projected external data from either modality (i.e., columns of $\mathbf{W}_{DiCA}^{\top}\mathbf{X}$) as the negative instances $\mathbf{X}_{neg} \in \mathbb{R}^{\tilde{k} \times n}$, where n indicates the sample number. As illustrated in Figure 4(a), the negative samples for learning person-specific classifiers are randomly selected from external (i.e., *unseen*) face images. Once the positive and negative samples are obtained for each subject, linear SVM classifiers will be trained for recognizing projected probe images. These SVM classifiers can thus be viewed as person-specific classifiers. It is worth noting that, we do not intentionally sample the projected external data which are particularly farther away from the project gallery images. This is because that, if doing so, the generalization of the resulting SVM will be poor (as depicted in Figure 4(b)).

Finally, we note that the above strategy is not limited to the use of DiCA for domain adaptation, and can be extended to other domain adaptation approaches which are based on the derivation of common feature spaces.

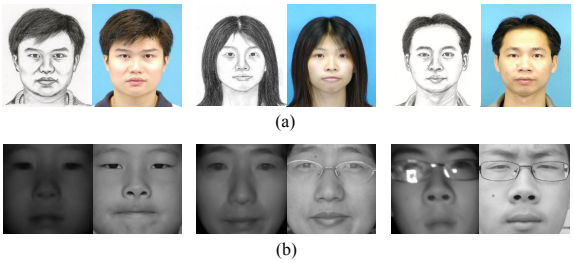


Fig. 5. (a) Example sketch-photo image pairs of the CUFS dataset, and (b) example NIR-VIS image pairs of the CASIA 2.0 dataset. Note that image differences between the same subject but across different domains in CASIA is more significant than that in CUFS, and thus HFR with CASIA is expected to be more challenging.

3. EXPERIMENTS

3.1. Sketch-to-Photo Face Recognition

We first address the problem of sketch-to-photo face recognition. We apply the setting of [5, 7] and take a subset of CUFS database [15] for evaluation. This dataset contains sketch/photo pairs of 188 persons, and each sketch/photo image is of size 200×155 pixels (see examples in Figure 5(a)). When performing recognition, 100 subjects are randomly selected to be recognized, and the image pairs of the remaining 88 persons are considered as external data.

To select the threshold T for DiCA, we observe the $d(\mathbf{w}_i)$ values of the eigenvectors derived from external data (as depicted in Figure 2) and set $T = 0.1$. When training person-specific linear SVMs in our DiCA subspace, we randomly choose $n = 20$ projected external face images as negative samples $\mathbf{X}_{neg}$ (and the projected gallery images of the subjects of interest as positive ones). We found that the recognition performance is not sensitive to the selection of C for SVMs and thus set $C = 0.1$. We repeat the above process 10 times, and report the average recognition performance.

For evaluations and comparisons, we first apply the use of nearest neighbor (NN) classifiers as the baseline approach (i.e., no domain adaptation). We further consider several com-

Table 1. Performance comparisons of sketch-to-photo recognition using the CUFS dataset.

NN	PLS [5]	Bilnear [16]	CCA [17]
83.5	93.6	94.2	94.6

Yang *et al.* [18]	Li *et al.* [10]	SCDL [6]	Ours
95.4	95.1	95.2	**97.5**

Table 2. Performance comparisons of NIR-to-VIS recognition using the CASIA dataset.

NN	SCDL [6]	CCA [17]	Li *et al.* [10]	Ours
3.9	8.9	10.1	23.7	**26.8**

mon feature space based approaches: PLS [5], Bilnear [16], CCA [17], SCDL [6], the approaches of Yang *et al.* [18] and Li *et al.* [10]. Table 1 lists the recognition rates of different methods, using the same setting for training/testing data as noted above. From Table 1, it can be seen that our proposed approach significantly outperformed others.

As commented in Section 2.2, our proposed strategy for person-specific domain adaptation can be applied to other common feature space based approaches. We particularly consider the methods of CCA and SCDL and train our person-specific SVMs in the corresponding subspaces. Compared to 94.6% (CCA) and 95.2% (SCDL) in Table 1, we observed that the recognition rates for these two approaches were increased to 96.8% and 97.1%, respectively. This confirms the effectiveness of our proposed person-specific domain adaptation strategy. It is worth noting that, our proposed method still achieved the highest recognition rate of 97.5%, and thus it can be verified that our DiCA exhibits better domain adaptation ability for cross-modality face images.

3.2. NIR-to-VIS Face Recognition

We next consider the dataset of CASIA NIR-VIS 2.0 [10] for NIR-to-VIS face recognition, which contains NIR and VIS face images of a total of 725 persons (1-22 VIS face images and 5-50 NIR face images available for per subject). We note that, since there exist four different types of facial variations (i.e., pose, expression, eyeglasses, and scale), the recognition task is very challenging (and thus lower recognition rates will be expected). Example images are shown in Figure 5(b), in which each cropped face image is of size 128×128 pixels.

When performing HFR, we apply the setting of [10] for selecting external, gallery, and probe data. In particular, 367 (out of 725) persons are randomly selected as subjects not of interest (i.e., external data), and we randomly choose 1500 images from each image domain (which are *not* necessarily from the same persons) for constructing our DiCA subspace. As for the remaining 358 subjects of interest, we have one VIS gallery image per person, and about 6200 NIR images of these subjects as the probe images. We have $T = 0.01$ for our DiCA, and we also set $C = 0.1$. We have $n = 20$ projected

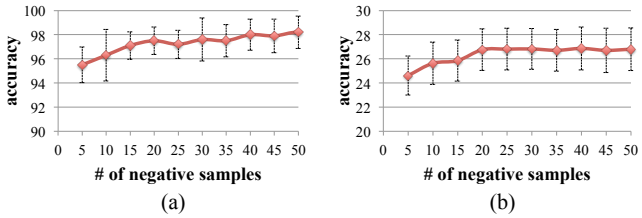


Fig. 6. Recognition performance using different numbers of randomly selected negative instances for our person-specific classifiers:(a) CUFS and (b) CASIA datasets.

external face images randomly selected as negative samples $\mathbf{X}_{neg}$ when training our classifiers.

Table 2 compares the performances of different methods: NN, CCA [17], SCDL [6], and the approach of Li *et al.* [10]. Recall that, the NIR and VIS images of the same subject in this dataset are very different, and thus lower recognition rates will be expected (compared to those of sketch-to-photo recognition). Nevertheless, from Table 2, it is clear that our proposed method obtained the highest recognition rate among all methods for NIR-to-VIS recognition.

Finally, we discuss the performance sensitivity to the number of randomly selected negative samples (i.e., n). As shown in Figure 6, we achieved comparable results for HFR tasks, when the number of negative instances (i.e., $\mathbf{X}_{neg}$ which is randomly selected from cross-domain image data of subjects not of interest) was above 20. Since adding more negative instances did not affect the performance (but would increase computation complexity), we conclude that $n = 20$ was sufficient for both HFR problems. Based on our experiments presented for both datasets, the effectiveness and robustness of our proposed person-specific domain adaptation model for HFR can be successfully verified.

4. CONCLUSIONS

We proposed a person-specific domain adaptation approach for HFR. We first apply DiCA for associating cross-modality face images using external data. In order to generalize the derived feature space to the subjects of interest, we further advocated the learning of person-specific SVM classifiers using projected gallery and external images as positive and negative samples, respectively. In addition to improved representation and discriminating capabilities for associating cross-modality face images, a major advantage of our approach is that no data correspondence or label information was needed when collecting external cross-domain data. Our experiments confirmed that we outperformed several baseline and state-of-the-art domain adaptation approaches on sketch-to-photo and NIR-to-VIS HFR problems.

Acknowledgements This work is supported in part by the Ministry of Science and Technology of Taiwan via NSC102-3111-Y-001-015 and NSC102-2221-E-001-005-MY2.

5. REFERENCES

[1] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *IEEE CVPR*, 2009.

[2] B. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Trans. PAMI*, 2011.

[3] B. Klare and A. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE PAMI*, 2012.

[4] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li, "Face matching between near infrared and visible light images," in *Advances in Biometrics*. 2007.

[5] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *IEEE CVPR*, 2011.

[6] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *IEEE CVPR*, 2012.

[7] D.-A. Huang and Y.-C. F. Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition," in *IEEE ICCV*, 2013.

[8] B. Klare and A. K. Jain, "Heterogeneous face recognition: Matching NIR to visible light images," in *IEEE ICPR*, 2010.

[9] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE TKDE*, 2010.

[10] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 Face Database," in *PBVS*, 2013.

[11] C.-A. Hou, M.-C. Yang, and Y.-C. F. Wang, "Domain adaptive self-taught learning for heterogeneous face recognition," in *ICPR*, 2014.

[12] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *IEEE ICCV*, 2013.

[13] A. Gretton, K. Borgwardt, M. Rasch, B. Schlkopf, and A. Smola, "A kernel method for the two sample problem," in *NIPS*, 2007.

[14] Q. Yin, X. Tang, and J. Sun, "An associate-predict model for face recognition," in *IEEE CVPR*, 2011.

[15] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2009.

[16] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, 2000.

[17] H. Hotelling, "Relations between two sets of variates," *Biometrika*, 1936.

[18] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE TIP*, 2010.