COARSE TO FINE TRAINING FOR LOW-RESOLUTION HETEROGENEOUS FACE RECOGNITION

Sivaram Prasad Mudunuri and Soma Biswas
{sivaramm, somabiswas}@iisc.ac.in

Department of Electrical Engineering, Indian Institute of Science, Bangalore

ABSTRACT

Recently, near-infrared (NIR) images are being increasingly used for recognizing facial images across illumination variations and in low-light conditions. In surveillance scenarios, the captured NIR may have low-resolution which results in significant loss of discriminative information along with uncontrolled pose. In this work, we address the challenging task of matching these low-resolution (LR) uncontrolled NIR images with high-resolution (HR) controlled visible (VIS) images usually present in the database. Since the probe and gallery images differ significantly in terms of pose, resolution and spectral properties, we employ a two-stage approach. First, the images are transformed into a common space using metric learning such that the images of the same subject are pushed closer and those of different subjects are pushed apart. We then define an objective function which can simultaneously push both LR NIR and HR VIS samples towards the centroids of the HR VIS samples. We show that the approach is general and can be used for other data like RGB-D and also for matching across pose. Extensive experiments conducted on five datasets shows the effectiveness of our approach.

Index Terms— Heterogeneous face recognition, low-resolution, super-resolution, RGB-D, pose.

1. INTRODUCTION

In real world surveillance scenarios, considering the increased security threats at night time, it is often required to recognize faces captured under low light conditions and uncontrolled pose. Recently, surveillance cameras that can operate in NIR mode are being installed to address the low light conditions. The resolution of these face images are usually low because of the large distance between the cameras and the subjects, which leads to poor recognition performance since the LR images carries less discriminative information. On the other hand, the gallery images are usually captured with HR VIS cameras under frontal pose and good illumination conditions during enrolment. Matching the LR NIR probe images under uncontrolled pose with the HR VIS gallery faces under frontal pose and good illuminations (Fig. 1) is a challenging task [1],[2],[3]. Though there is a significant amount of work addressing the heterogeneous face recognition in literature [4],[5], limitted attention has been given to address the cross-resolution heterogeneous face recognition [1],[6].

Ghosh *et al.* [1] propose a framework that perform hierarchical fusion at both feature and score levels to match LR NIR with HR VIS faces. A re-ranking based approach that extract the features of probe and gallery faces using the relationship with reference image set is proposed in [6]. A coupled kernel embedding (CKE) scheme is designed in [9] to extract robust features from LR faces. Moutafis *et al.* [10] propose a framework to jointly learn two semi-coupled

Fig. 1. Cross-resolution heterogeneous face recognition: Few samples of HR VIS frontal gallery faces (first row) and LR NIR non frontal probe faces are shown (second row) [7] [8].

basis for LR and HR faces by maintaining the class discriminability. Maeng *et al.* [11], propose a feature extraction based method to address the NIR face recognition where the probe images are taken from a large distance.

In this work, we propose a coarse-to-fine training approach to address this task. In the course training step, we employ discriminative metric learning method to transform the LR NIR and HR VIS face images to a common discriminative space so that the samples of similar subjects move closer and that of dissimilar subjects move apart. In the second fine training stage, we design a coupled cross-domain objective function to compute two projection matrices to simultaneously pull both the LR NIR and HR VIS samples to the corresponding class centroids computed from HR VIS samples in the common space. We demonstrate that the proposed objective function improves the recognition performance since it encourages samples of both the modalities to move towards the means of HR VIS which further reduces the inter-class variations.

The contributions of the paper are as follows: 1) We propose an objective function to map the LR NIR and HR VIS samples to the class centroids of the HR VIS faces. 2) We show that the approach is general and is applicable to other data like RGB-D and for matching across pose. 3) Extensive experiments performed on modified CASIA NIR-VIS 2.0 database [7], Surveillance Camera (SC) face database [8], Texas 3D Face Recognition database [12],[13], VAP RGB-D-T database [14] and RGB-D object database [15] illustrates the effectiveness of proposed approach.

2. PROPOSED APPROACH

In this section, we provide details of the proposed framework for handling cross-resolution heterogeneous face recognition. Since the probe and gallery images differ significantly in terms of pose, resolution and spectral characteristics, we propose a two-stage training approach. During the coarse training stage, both the LR NIR and HR VIS images (features) are mapped into a common space so that the samples of the same subject move closer and that of different subjects move apart. During the fine training stage, we design a coupled cross-domain objective function which can simultaneously

move both the LR NIR and HR VIS samples towards the centroids of HR VIS samples in the common space. Now, we discuss the two-stages in details.

Coarse Training Stage: Let $\mathbf{X} = [x_1, x_2, \ldots, x_N] \in \mathcal{R}^{d\times N}$ and $\mathbf{Y} = [y_1, y_2, \ldots, y_N] \in \mathcal{R}^{d\times N}$ be the LR NIR and HR VIS data respectively. Here N is the total number of training samples. Let us consider that the data samples belongs to C subjects with their labels $\ell = [\ell_1, \ell_2, \ldots, \ell_N]$. Our first step is to map these samples into a low dimensional discriminative space. The distance between any two samples can be formulated as [16],[17]:

$$\delta(x_i, y_j) = (x_i - y_j)^T \mathbf{M}(x_i - y_j) \tag{1}$$

Here, $\mathbf{M} \in R^{d\times d}$ is the Mahalanobis metric. In this task, we employ cross view quadratic discriminant analysis (XQDA) approach developed in [16] to model $\mathbf{M}$. Though we have samples from C classes, we can always make it a two class problem by considering similar pairs ($\ell_i = \ell_j$) and dissimilar pars ($\ell_i \neq \ell_j$). This two class classification problem can be formulated through log-likelihood ratio test as

$$\delta(\phi_{ij}) = log\left(\frac{\frac{1}{\sqrt{2\pi|\mathbf{\Sigma}_{\ell_{ij}=0}|}} exp\left(-\frac{1}{2}(\phi_{ij})^T \mathbf{\Sigma}^{-1}_{\ell_{ij}=0}\phi_{ij}\right)}{\frac{1}{\sqrt{2\pi|\mathbf{\Sigma}_{\ell_{ij}=1}|}} exp\left(-\frac{1}{2}(\phi_{ij})^T \mathbf{\Sigma}^{-1}_{\ell_{ij}=1}\phi_{ij}\right)}\right) \tag{2}$$

where, $\ell_{ij} = 1$, if the samples x_i and y_j have same label, and is 0 if they are different. $\phi_{ij} = x_i - y_j$ is the difference vector of the samples. The matrices $\mathbf{\Sigma}_{\ell_{ij}=1}$ and $\mathbf{\Sigma}_{\ell_{ij}=0}$ represent the covariance matrices of similar pairs and dissimilar pairs respectively.

Now, expanding (2) by applying logarithm and solving it in least square manner [17], we can compute the Mahalanobis matrix as $\mathbf{M} = \mathbf{\Sigma}^{-1}_{\ell_{ij}=1} - \mathbf{\Sigma}^{-1}_{\ell_{ij}=0}$. Since the classification task is usually preferred in lower dimensional space, we define a dimensionality reduction scheme on the above formulation to simultaneously transform the original features into lower dimension space by maintaining the discriminative power. Now, (1) can be re-written as:

$$\delta(x_i, y_j) = (x_i - y_j)^T \mathbf{U}\mathbf{M}\mathbf{U}^T(x_i - y_j) \tag{3}$$

Here, $\mathbf{U} \in R^{d\times k}$ is a kernel matrix that will map the d dimensional point into k dimensional space where, $k < d$. (3) cannot be directly solved since the matrix $\mathbf{U}$ is associated with the inverse of covariance matrices. Instead, by imposing the Gaussian mean distribution assumptions, we can get the $\mathbf{U}$ matrix by analyzing variances of similar pairs and dissimilar pairs by solving the equation below which is similar to the Generalized Rayleigh Quotient:

$$J(u) = \frac{u^T \mathbf{\Sigma}^{-1}_{\ell_{ij}=0} u}{u^T \mathbf{\Sigma}^{-1}_{\ell_{ij}=1} u} \tag{4}$$

Here u is the direction at which the ratio between the variances of dissimilar and similar classes are maximum. This equation can be solved by the generalized eigen value decomposition problem [16] in which, the largest eigenvalue of $(\mathbf{\Sigma}^{-1}_{\ell_{ij}=1}\mathbf{\Sigma}_{\ell_{ij}=0})$ is the value of $J(u)$ and the corresponding direction is the first vector u_1 in the matrix $\mathbf{U}$. The other directions $u_2, u_3 \ldots u_k$ are the (k-1) directions of subsequent highest (k-1) eigen values respectively. We found that, having the value of k which is equal to the number of eigen values that are greater than 1 works well for all our experiments.

Fine Training Stage: Once the matrices $\mathbf{U}$ and $\mathbf{M}$ are solved, we can transform the original data points $\mathbf{X}$ and $\mathbf{Y}$ into common space and let us denote them as $\hat{\mathbf{X}} \in R^{k\times N}$ and $\hat{\mathbf{Y}} \in R^{k\times N}$. Since the HR VIS images are more discriminative than LR NIR, we design a novel objective function which can simultaneously transform both the data domains into the corresponding class centroids of HR VIS data.

Let $\mathbf{S} \in R^{k\times C}$ be the matrix in which each column represents the centroid of the individual class. Since we assume that, there are C number of subjects, we have C centroid vectors. Let $\mathbf{L} \in R^{N\times C}$ denote the label indicator matrix in which each row indicates the label of the corresponding sample. $\mathbf{L}_{ij}$ = 1 if the i^{th} sample in $\hat{\mathbf{X}}$ (and $\hat{\mathbf{Y}}$) belongs to j^{th} class. Otherwise, $\mathbf{L}_{ij}$ = 0. The proposed objective function can be formulated as below:

$$\begin{aligned}\min_{\mathbf{W_1},\mathbf{W_2}} \left\| \hat{\mathbf{X}}^T\mathbf{W_1}\mathbf{S} - \mathbf{L}\right\|_2 + \left\| \hat{\mathbf{Y}}^T\mathbf{W_2}\mathbf{S} - \mathbf{L}\right\|_2 + \\ \lambda_1 \left\| \hat{\mathbf{X}}^T\mathbf{W_1}\right\|_2 + \lambda_2 \left\| \hat{\mathbf{Y}}^T\mathbf{W_2}\right\|_2 + \lambda_3 \| \mathbf{W_1}\|_2 + \\ \lambda_4 \| \mathbf{W_2}\|_2 + \lambda_5 \| \mathbf{W_1}\mathbf{S}\|_2 + \lambda_6 \| \mathbf{W_2}\mathbf{S}\|_2 . \\ \text{such that, } \hat{\mathbf{Y}}^T\mathbf{W_2} = \hat{\mathbf{X}}^T\mathbf{W_1}.\end{aligned} \tag{5}$$

Here, $\mathbf{W_1} \in R^{k\times k}$ and $\mathbf{W_2} \in R^{k\times k}$ are the respective projection matrices that transforms both LR NIR and HR VIS samples towards the centroids $\mathbf{S}$. The objective function is inspired from [18], but here, we develop the formulation from the cross-domain perspective instead of a single domain. The first term in (5) can also be thought of as a domain adaptation task since we transform a domian of LR faces (say source distribution) which has poor representation towards the means of HR faces (target distribution) that have better representation.

For the easiness of our presentation, let us denote all the regularization terms in (5) that are associated with λ_m (m = 1 to 6) as $\vartheta(\mathbf{W_1}, \mathbf{W_2})$. Then, (5) can be re-written as:

$$\begin{aligned}\min_{\mathbf{W_1},\mathbf{W_2}} \left\| \hat{\mathbf{X}}^T\mathbf{W_1}\mathbf{S} - \mathbf{L}\right\|_2 + \left\| \hat{\mathbf{Y}}^T\mathbf{W_2}\mathbf{S} - \mathbf{L}\right\|_2 + \\ \gamma \left\| \hat{\mathbf{Y}}^T\mathbf{W_2} - \hat{\mathbf{X}}^T\mathbf{W_1}\right\|_2 + \vartheta(\mathbf{W_1}, \mathbf{W_2}).\end{aligned} \tag{6}$$

Since the above equation is coupled in two variables $\mathbf{W_1}$ and $\mathbf{W_2}$, we can solve for one variable by fixing the other one. This can be iteratively updated until there is no significant variations with the resultant matrices between the previous and the present iterations.

Compute $\mathbf{W_1}$:
In this case, the other projection matrix $\mathbf{W_2}$ is considered fixed and by computing the derivative of the objective function (6) w.r.t $\mathbf{W_1}$ and making it equal to zero, we can get the closed form solution of $\mathbf{W_1}$ as below.

$$\mathbf{W_1} = [\hat{\mathbf{X}}\hat{\mathbf{X}}^T + \lambda_5\mathbf{I}]^{-1}[\hat{\mathbf{X}}\mathbf{L}\mathbf{S}^T + \gamma\hat{\mathbf{X}}\hat{\mathbf{Y}}^T\mathbf{W_2}][\mathbf{S}\mathbf{S}^T + (\lambda_1 + \gamma)\mathbf{I}]^{-1} \tag{7}$$

Compute $\mathbf{W_2}$:
The other projection matrix $\mathbf{W_1}$ is fixed in this case and by computing the derivative of the objective function (6) w.r.t $\mathbf{W_2}$ and making it equal to zero, we can get the closed form solution of $\mathbf{W_2}$ as below.

$$\mathbf{W_2} = [\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T + \lambda_6\mathbf{I}]^{-1}[\hat{\mathbf{Y}}\mathbf{L}\mathbf{S}^T + \gamma\hat{\mathbf{Y}}\hat{\mathbf{X}}^T\mathbf{W_1}][\mathbf{S}\mathbf{S}^T + (\lambda_2 + \gamma)\mathbf{I}]^{-1} \tag{8}$$

The matrices are initialized with random gaussian matrices and updated iteratively until there are no significant variations with $\mathbf{W_1}$ and $\mathbf{W_2}$. In most of our experiments, we found that, the function usually converges after around 15 iterations. For simplifying equations (7) and (8), we assume that $\lambda_5 = \lambda_3/(\lambda_1 + \gamma)$ and $\lambda_6 = \lambda_4/(\lambda_2 + \gamma)$.

The objective function for fine-training (5) has the following differences with [18]: 1) It can handle cross domain data, 2) It is

designed in such a way that, samples from both the domains will be pushed towards the class centroids of the better domain.

Testing Stage: Once the optimization is completed, we can directly move to the testing stage. During testing, the gallery images are HR VIS and the probe images are LR NIR faces. Both the gallery and probe features will be transformed into the common low dimensional space using the matrices obtained from discriminant analysis (eqn 3). Then, we transform them individually using their corresponding matrices $\mathbf{W_1}$ and $\mathbf{W_2}$ as learnt using our proposed objective function. For a given probe whichever gallery is the first nearest neighbour, we declare that as the probe ID. We found a noticeable improvement in the recognition rates by employing modified cosine similarity (MCS) [19] instead of cosine similarity (CS) to find the first nearest neighbour. In our work, we compute the scores after step 1 (XQDA) and step 2 (the entire framework) individually and apply score fusion to further improve the performance. We applied weighted average in our settings.

3. EXPERIMENTAL SECTION

In this section, we demonstrate the effectiveness of the proposed framework in handling LR NIR with HR VIS faces by conducting extensive experiments on CASIA NIR-VIS 2.0 database [7] and SC face database [8]. We also demonstrate the usefulness of our method for matching depth images with RGB images by conducting experiments on Texas 3D Face Recognition Database [12],[13] and VAP RGB-D-T Database [14]. To further illustrate the generalizability of our approach, we also performed experiments on RGB-D object database [15] to match objects across pose variation.

3.1. Experiments on CASIA NIR-VIS 2.0 database [7]

The database has VIS and NIR images of 725 subjects that are captured in four sessions. There are 1-22 visible and 5-50 NIR facial images per person in the database. Since all the images in the database are of high resolution (128×128), we synthesize the LR NIR images by downsampling the HR NIR images to the resolution of 20×20 and upsampling to the original resolution 128×128 using bi-cubic interpolation. The VIS images are kept as HR of original size 128×128. Now, we match the synthesized LR NIR images with HR VIS images. We follow the standard experimental protocol as given by the database (but the NIR images are of LR) in which, we repeat the recognition experiment over 10 folds and report the mean and standard deviation. The results along with the comparisons with state-of-the-art approaches are reported in Table 1.

For all the experiments in this database, we extracted VGG face deep features [27] (concatenated pool5, fc6 and fc7) and applied PCA to reduce the dimensionality to 2500. It can be observed from Table 1 that, the proposed approach is outperforming the state-of-the-approaches. As we discussed earlier, the modified cosine similarity metric is boosting the performance even further. We present the results with the final step of our framewrok (score fusion) in all our experiments unless explicitly mentioned. Please note that, all the other methods including **Ours** are evaluated based on CS metric whereas, **Ours (MCS)** is evaluated with MCS metric. We experimentally fixed the parameters: $\lambda_1 = \lambda_2 = 20$; $\lambda_3 = \lambda_4 = 0.01$, $\lambda_5 = \lambda_6 = 0.0003$, $\gamma = 10$.

Effect of Super Resolution: Since the NIR images are of low resolution, one can apply super resolution (SR) on the LR image to synthesize the high resolution image and perform matching. In our experiments, we employ state-of-the-art deep learning based super resolution technique [28] on the input LR NIR images to synthesize the corresponding HR image and apply our proposed framework. We similarly apply the other better performing baseline methods on these SR images to demonstrate the performance comparison. For this experiment, we randomly select *fold 3* of the same database [20] and performed the experiment. Rank-1 recognition rates (%) on both LR images (of three resolutions: 20×20, 30×30 and 60×60) and the corresponding SR images are reported in Table 2. We observe that, as expected, the performance of all the approaches degrades with decrease in resolution. We also observe that SR is helpful in improving the performance and thus can be used to boost the performance of any baseline algorithm and that the proposed approach significantly outperforms all the other state-of-the-art algorithms.

Table 1. Comparison of the proposed framework with the state-of-the-art on CASIA NIR-VIS 2.0 database [20]. The HR VIS and the LR NIR images have resolution 128×128 and 20×20 respectively. The experiment is repeated over the 10 folds and the rank-1 and rank-5 recognition rates with standard deviation (std) are reported.

Method	Rank-1±std (%)	Rank-5±std (%)
GMA [21]	41.56±1.72	67.71±1.50
meanCCA [22]	41.82±1.87	67.72±1.22
ClusterCCA [22]	42.64±1.61	69.00±1.22
CBFD [23]	42.47±0.97	61.92±1.08
CA-LBFL [24]	44.41±1.12	63.74±1.22
RandKCCA [25]	45.42±1.51	71.63±1.07
Dict Align [26]	58.38±1.19	80.88±1.06
Re-ranking on LR [6]	60.21±1.26	82.64±1.02
Ours	**64.71±1.40**	**85.90±1.12**
Ours (MCS)	**69.45±1.40**	**88.28±0.76**

Table 2. Illustrating the effect of applying SR technique [28] on LR NIR faces before applying the recognition method. LR: results (Rank-1 %) without applying SR method, SR: results after applying the SR method on LR NIR faces.

	20×20		30×30		60×60	
Method	LR	SR	LR	SR	LR	SR
CBFD [23]	42.80	47.99	63.21	65.84	68.52	71.28
CA-LBFL [24]	46.29	48.65	63.29	65.11	68.18	72.98
RandKCCA [25]	45.60	50.09	65.42	67.95	73.83	74.96
Dict Align [26]	57.56	63.04	76.19	79.64	83.34	84.44
Ours	**63.39**	**67.34**	**78.96**	**80.95**	**84.92**	**84.76**
Ours (MCS)	**67.12**	**71.22**	**83.59**	**85.10**	**87.89**	**87.63**

Effect of the different stages of the proposed approach: Here we analyze the effectiveness of each of the stages of the proposed approach. For the probe resolution of 20×20, Table 3 gives the rank-1 and rank-5 performance after each stage. We observe that, both the coarse and the fine stages as well as the score fusion helps in boosting the performance.

Table 3. Effect of the different stages of the proposed approach on the rank-1 and rank-5 recognition rates.

Performance (%)	Coarse	Coarse + Fine	Coarse + Fine + Fusion
Rank-1	62.51±1.64	64.11±1.45	**64.71±1.40**
Rank-5	84.44 ±1.17	85.66±1.07	**85.90±1.12**

3.2. Experiments on Surveillance Camera Database [8]

To demonstrate the applicability of the proposed approach on real surveillance quality NIR images, we perform the experiment on SC face database [8]. The database has both VIS and NIR images of 130 subjects. The VIS images are of high resolution and captured under well controlled settings. The NIR images of the database are captured from three different distances and there are two surveillance quality NIR cameras at each distance. We randomly take 50 subjects for training and the remaining for testing. We fix the HR VIS images as the gallery and the surveillance quality NIR images from two cameras at each distance are taken as probe images. We repeat the experiment for three distances and present the average rank-1 recognition rate (%) in Table 4. We tried evaluating the other baseline approaches on this database but, their performance is quite low. It can be observed from the Table 4 that, the proposed approach is performing close to the state-of-the-art approach under the surveillance quality settings.

Table 4. Rank-1 recognition rates (%) of the proposed approach on SC face database [8].

Method	Rank-1 (%)
NN	17.08
Dict Align [26]	30.83
Ours	29.37
Ours (MCS)	**31.87**

3.3. Experiments on Texas 3D Face Recognition Database

To demonstrate the generalizability of the proposed approach for other applications, we conduct experiments on Texas 3D Face Recognition Dataset [12],[13] to match depth data with RGB images. The database has a total of 1149 pairs of depth and high resolution RGB facial images of 118 adult human subjects. The provided images are pose normalized, aligned and pre-processed. In our settings, following the same protocol of [29],[30], we take depth data as probe and, RGB images as gallery. We randomly select 60 subjects for training and the remaining 58 subjects are used for testing. The results are reported with SIFT descriptors [31] extracted at 14 fiducial locations as well as the LBP features [32] in Table 5. The results of all the other approaches are directly taken from [29],[30] since we followed the same protocol.

3.4. Experiments on VAP RGB-D-T Database

We also conducted experiments on VAP RGB-D-T Database [14] which has synchronized RGB, Depth and thermal images of 51 persons captured in three different settings: rotation, expression and illumination to further demonstrate the usefulness of our approach in general cross-modal applications. In our experiments, we followed the same protocol as [30], where we take 26 subjects for training and 25 for testing and then we match depth (probe) data with RGB (gallery) images. The experimental results are reported in Table 5 and the results of all the other methods are directly taken from [29],[30] since we followed the same protocol.
It can be observed from Table 5 that, the proposed approach is performing better than the state-of-the-art approaches for both the databases which illustrates the usefulness of our approach for this application.

3.5. Experiments on RGB-D Object Database

To illustrate the effectiveness of our framework for object recognition across pose variations, we conduct the experiments on visible images of RGB-D Object database [15] which has images from 51 categories. Sample images of the database with pose variations are shown in Fig. 2. We followed the same protocol (visual - visual) as [36],[37] where we match the RGB objects across pose variations. The experimental results are provided in Table 6 along with the comparisons taken from [36],[37]. We observe that the proposed framework performs favorably compared to the state-of-the-art techniques which demonstrate the usefulness of the proposed approach.

Table 5. Rank-1 recognition rates (%) of the proposed approach on Texas 3D Face Recognition dataset [12],[13] and VAP RGB-D-T database [14]. The probe images are depth and the gallery are RGB images. '-' indicates, no result is reported in [30].

	Texas 3D Face		**VAP RGB-D-T**
Method	SIFT	LBP	SIFT
Cluster CCA [22]	50.05	86.37	23.23
GMA [21]	62.41	95.46	32.16
SCDL+LSML [33],[17]	60.78	91.59	36.64
CDL+LSML [34],[17]	56.32	93.14	34.36
Deep CCA [35]	43.03	-	-
S2CDL^I [29]	64.73	97.1	37.69
S2CDL^C [30]	62.16	-	-
S2CDL^P [30]	60.39	-	-
Ours	**68.34**	**99.25**	**49.10**
Ours (MCS)	**69.29**	**99.35**	**49.14**

Fig. 2. Sample RGB images of RGB-D Object database [15] illustrating the pose variations.

Table 6. Recognition performance (%) of our approach on RGB-D object database for matching objects across pose (visual - visual).

Method	Rank-1 (%)
LSML [17]	60.1
GMA [21]	70.6
SCDL [33]	80.4
CFDL [34]	81.0
SCDL+LSML [33],[17]	81.7
CFDL+LSML [34],[17]	82.0
DPFD [37]	86.0
DPFD-LCC [37]	84.8
ADPFD [36]	91.4
Ours	**93.2**
Ours (MCS)	**93.4**

4. CONCLUSION

A novel approach for matching low resolution NIR images with non frontal pose and poor illumination conditions with high resolution VIS faces captured under controlled settings is proposed. A coupled cross domain objective function is designed to pull all the samples to their corresponding class centroids of HR VIS faces in the common space. Extensive experiments are conducted on five different databases to demonstrate the generalizability of the proposed approach in handling different challenging recognition tasks.

5. REFERENCES

[1] S. Ghosh, R. Keshari, R. Singh, and M. Vatsa, "Face identification from low resolution near-infrared images.," *ICIP*, pp. 938–942, 2016.

[2] M. Singh, S. Nagpal, N. Gupta, S. Ghosh, R. Singh, and M. Vatsa, "Cross-spectral cross-resolution video database for face recognition," *BTAS*, pp. 1–7, 2016.

[3] S. Ouyang, T. Hospedales, Y.Z. Song, X. Li, C.C. Loy, and X. Wang, "A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution," *Image and Vision Computing*, vol. 56, pp. 28–48, 2016.

[4] J. Lezama, Q. Qiu, and G. Sapiro, "Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding.," *CVPR*, pp. 6807–6816, 2017.

[5] T. de Freitas Pereira and S. Marcel, "Heterogeneous face recognition using inter-session variability modelling.," *CVPR Workshops*, pp. 111–118, 2016.

[6] S. P. Mudunuri, S. Venkataramanan, and S. Biswas, "Improved low resolution heterogeneous face recognition using re-ranking," *Indian National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, 2017.

[7] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir - vis 2.0 face database," *CVPR*, pp. 348–353, 2013.

[8] M. Grgic, K. Delac, and S. Grgic, "Scface–surveillance cameras face database," *MTA*, vol. 51, no. 3, pp. 863–879, 2011.

[9] C. X. Ren, D. Q. Dai, and H. Yan, "Coupled kernel embedding for low-resolution face image recognition," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3770–3783, 2012.

[10] P. Moutafis and I. A. Kakadiaris, "Semi-coupled basis and distance metric learning for crossdomain matching: Application to low-resolution face recognition.," *IJCB*, pp. 1–8, 2014.

[11] H. Maeng, H. C. Choi, U. Park, S. W. Lee, and A. K. Jain, "Nfrad: Near-infrared face recognition at a distance," *IJCB*, pp. 1–7, 2011.

[12] S. Gupta, K.R. Castleman, M.K. Markey, and A.C. Bovik, "Texas 3d face recognition database," *IEEE Southwest Symposium on Image Analysis and Interpretation*, pp. 97–100, 2010.

[13] S. Gupta, M.K. Markey, and A.C. Bovik, "Anthropometric 3d face recognition," *International Journal of Computer Vision*, vol. 90, no. 3, pp. 331–349, 2010.

[14] O. Nikisins, K. Nasrollahi, M. Greitans, and T. B. Moeslund, "Rgb-dt based face recognition," *CVPR*, pp. 1716–1721, 2014.

[15] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," *ICRA*, vol. 67, pp. 1817 – 1824, 2011.

[16] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," *CVPR*, pp. 2197–2206, 2015.

[17] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," *CVPR*, pp. 2288–2295, 2012.

[18] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," *ICML*, pp. 2152–2161, 2015.

[19] C. Liu, "Discriminant analysis and similarity measure," *Pattern Recognition*, vol. 47, no. 1, pp. 359 –367, 2014.

[20] S. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database.," *CVPR Workshops*, pp. 348–353, 2013.

[21] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space.," *CVPR*, pp. 2160–2167, 2012.

[22] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis.," *In Artificial Intelligence and Statistics*, pp. 823–831, 2014.

[23] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 10, pp. 2041–2056, 2015.

[24] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Context-aware local binary feature learning for face recognition.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2017.

[25] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Scholkopf, "Randomized nonlinear component analysis.," *ICML*, pp. 1359–1367, 2014.

[26] S. P. Mudunuri and S. Biswas., "Dictionary alignment for low-resolution and heterogeneous face recognition.," *WACV*, pp. 1115–1123, 2017.

[27] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition.," *BMVC*, pp. 1–12, 2015.

[28] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution.," *CVPR*, pp. 1–9, 2017.

[29] N. Das, D. Mandal, and S. Biswas, "Simultaneous semi-coupled dictionary learning for matching rgbd data," *CVPR Workshops*, pp. 175–183, 2016.

[30] N. Das, D. Mandal, and S. Biswas, "Simultaneous semi-coupled dictionary learning for matching in canonical space," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3995–4004, 2017.

[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[32] T. Ojala, M. Pietikinen, and T. Menp, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[33] S. Wang, D. Zhang, Y. Liang, and Q. Pan, "Semi coupled dictionary learning with applications to image super resolution and photo sketch synthesis," *CVPR*, pp. 2216–2223, 2012.

[34] D. A. Huang and Y. C. F. Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition," *ICCV*, pp. 2496–2503, 2013.

[35] W. Wang, R. Arora, K. Livescu, and J.A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," *ICASSP*, pp. 4590–4594, 2015.

[36] S. Sanyal, D. Mandal, and S. Biswas, "Aligned discriminative pose robust descriptors for face and object recognition," *ICIP*, 2017.

[37] S. Sanyal, S.P. Mudunuri, and S. Biswas, "Discriminative pose-free descriptors for face and object matching," *Pattern Recognition*, vol. 67, pp. 353 – 365, 2017.