This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2868173, IEEE Transactions on Information Forensics and Security

Dictionary Alignment with Re-ranking for Low Resolution NIR-VIS Face Recognition

Sivaram Prasad Mudunuri*, Shashanka Venkataramanan*, and Soma Biswas

***Abstract*—Recently, near-infrared (NIR) images are increasingly being captured for recognizing faces in low-light/night-time conditions. Matching these images against the controlled high-resolution visible facial images usually present in the database is a challenging task. In surveillance scenarios, the NIR images can have very low-resolution and also non-frontal pose which makes the problem even more challenging. In this work, we propose an orthogonal dictionary alignment approach for addressing this problem. We also propose a re-ranking approach to further improve the recognition performance for each probe by combining the rank list given by the proposed algorithm with that given by another complementary feature/algorithm. Finally, we have also collected our own database HPR (Heterogeneous face recognition across Pose and Resolution) which has facial images captured from two surveillance quality NIR cameras and one high-resolution visible camera, with significant variations in head pose and resolution. Extensive experiments on the modified CASIA NIR-VIS 2.0 database, Surveillance Camera face (SCface) database and our HPR database shows the effectiveness of the proposed approaches and the collected database.**

***Index Terms*—heterogeneous face recognition, re-ranking, low-resolution, head pose.**

I. Introduction

To facilitate face recognition in low-light or night-time conditions, near-infrared (NIR) facial images are becoming quite common. In contrast, as most of the enrolled images in the database are controlled visible (VIS) images, this results in the heterogeneous face recognition problem. Several approaches have been proposed to address this task [36][39], but most of them deal with good resolution images. Recently, for security applications, more and more surveillance cameras are being installed at several places. The NIR images captured by these surveillance cameras usually have low-resolution in addition to considerable variations in pose (Fig. 1). This makes the problem even more challenging, since now the low-resolution (LR), uncontrolled NIR probe images needs to be matched against the high-resolution (HR) controlled VIS gallery images. But this scenario is relatively less explored in literature [15][41].

The contribution of this work is three-fold. First, we propose a dictionary alignment based approach to address this challenging problem. Since one-to-one correspondence between the HR VIS and LR NIR images may not be available during training, we propose to learn two orthogonal dictionaries for the two domains independently. In order to be able to match the coefficients computed from independently learnt dictionaries, we then learn the correspondence between the atoms of the two dictionaries. Finally, metric learning is used to make the coefficients discriminative since the final objective is recognition. The proposed framework can also be useful for addressing other cross-domain matching problems like photo vs sketch, RGB vs. thermal images, etc.

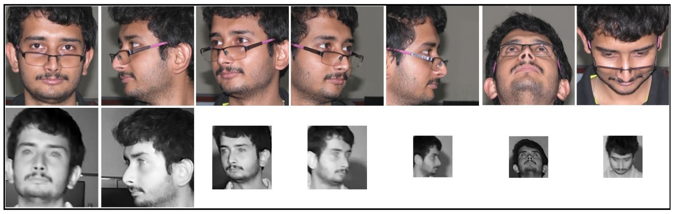

Fig. 1: Sample images from our HPR database. Column 1-7: HR VIS (Row 1) and NIR faces (Row 2) at 7 different poses (pose00-pose06). Column (1,2), (3,4), (5,6,7), Row 2 shows images at 8ft., 12ft. and 16ft. respectively from the camera.

It has been observed that given a probe and its retrieved results by any algorithm, the ranked scores and the information about the neighbors can potentially be utilized to further improve the retrieval performance. Our second contribution is a novel re-ranking algorithm, which takes the ranks (scores) given by the proposed approach and any other approach to further improve the recognition performance. We design our re-ranking approach by analyzing strongly similar, dissimilar and the neutral gallery images given by the two techniques. The proposed re-ranking approach is general and can help to improve the performance of any two baseline algorithms. Extensive experiments conducted on the modified CASIA NIR-VIS 2.0 database [23], Surveillance Camera face (SCface) database [9] and the proposed HPR database demonstrates the usefulness of the proposed dictionary alignment approach and the re-ranking approach.

In order to advance the state-of-the-art in this area, we feel that a systematic analysis of the effect of variations in pose and distance between the subject and the camera on the recognition performance will be useful. Most of the available datasets for this problem does not suit this purpose. In this work, to facilitate this study, we introduce a new database, termed as HPR (Heterogeneous face recognition across Pose and Resolution), which is our third contribution. It consists of 200 subjects captured using two different NIR surveillance cameras and one HR VIS camera.

A preliminary version of this work appeared in [29]. The additional contribution of this paper as compared to its preliminary version are as follows: (1) First, a novel re-ranking approach is proposed in Section IV; (2) Second, we introduce

* Authors contributed equally. The authors are with the Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India. E-mail: {sivaramm , shashankv, somabiswas}@iisc.ac.in

1556-6013 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

a new database HPR in Section V. Three different protocols are also described illustrating the effect of head pose and distance between the camera and the subject (Section VI D); (3) Finally, extensive experiments are conducted on the modified CASIA NIR-VIS 2.0 database, SCface database and our proposed database in Section VI to demonstrate the effectiveness of our proposed approach. Extensive analysis including the effect of super-resolution, different resolutions, face hallucination and different score fusion techniques are reported in Section VIB.

II. Related Work

In this section, we provide pointers to some related approaches in literature. We also discuss some available databases for matching NIR with VIS face images later.

Re-ranking Approaches: Several re-ranking approaches have been proposed for the application of person re-identification (re-ID) [51]. Zheng *et al.* [50] propose a late fusion method at score level to address the task, while Leng *et al.* [21] propose utilizing context and content similarity information. Ye *et al.* [47] use both similarity and dissimilarity cues to develop an optimized ranking framework. A fusion based re-ranking method which exploits the diversity of high dimensional feature vector is addressed in [48]. Sarfraz *et al.* [38] incorporate fine and coarse pose information of a person to learn a discriminative embedding used for automatic re-ranking. Bai *et al.* [2] propose a framework to improve the performance of object retrieval with diffusion process. A sparse contextual activation that encodes the local distribution of an image by considering the pairwise distance in contextual space is described in [1]. A re-ranking strategy which takes into account the relation of both probe and gallery with the reference image set is described in [30]. We utilize the concepts of strongly similar, dissimilar and neutral galleries from [8] and dual rank lists from [47] to design our own novel re-ranking framework for the LR heterogeneous face recognition problem.

Heterogeneous and low-resolution face recognition: Here, we discuss some recent works related to heterogeneous and low-resolution face recognition. Canonical Correlation Analysis [12] and its variants [25][34] are proposed to match the samples across any two modalities. Generalized Multi-view Analysis [40] solves a joint, relaxed quadratic constraint to obtain a common nonlinear discriminative subspace. A compact binary face descriptor that computes the descriptors from the pixel difference vectors in each local patch of the image is described in [27][6]. Wu *et al.* [45] describe a light CNN framework which learns a compact embedding on the large-scale face data with massive noisy labels. This work does not need one-to-one paired data. An extensive survey on heterogeneous face recognition is presented in [32]. Reale *et al.* [35] propose three different methods to improve heterogeneous face recognition involving removal of extra information from pre-trained networks, utilizing an altered contrastive loss function and improving the initialization network. A deep multi-task learning approach to jointly estimate multiple heterogeneous face attributes from a single image is discussed in [10]. Modality invariant deep features are learned by orthogonal subspace decomposition and max out operations at high-level representations of CNNs in [11]. Lezama *et al.* [22] use cross-spectral hallucination and low rank embedding for this task. A coupled deep learning framework for heterogeneous face recognition involving trace norm and block diagonal constraints is proposed in [46]. Klare *et al.* [18] propose to match cross-modal faces by computing kernel similarities with a collection of prototype faces. A framework to match NIR and VIS faces using transduction is described in [52]. A single hidden-layer Gabor-based network is proposed in [31] to recognise cross-modal faces.

We now discuss some recent works on LR face recognition. Hu *et al.* [13] propose to train CNNs with the available small number of face samples by generating synthetic images. Roy *et al.* [37] propose an illumination invariant feature describing the effect of the center pixel on its neighborhood for heterogeneous face recognition. Jin *et al.* [14] propose a coupled feature learning method to simultaneously exploit discriminative information and to reduce the appearance difference. Chu *et al.* [5] investigate LR face recognition with single sample per person and propose a cluster based regularized simultaneous discriminant analysis method.

III. DAlign: Proposed Dictionary Alignment Approach

In this section, we present our proposed framework containing orthogonal dictionary alignment method and re-ranking approach for matching data points from two different domains. Let us consider that, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_1}\} \in \mathcal{R}^{d\times N_1}$ and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{N_2}\} \in \mathcal{R}^{d\times N_2}$ be the data in the two domains, where N_1 may not be equal to N_2. Our approach does not require one-to-one paired data for training unlike other coupled dictionary learning approaches [44]. The training stage has the following steps: 1) Learn separate orthogonal dictionaries for the two domains and compute the correspondence of the dictionary atoms; 2) Align the dictionaries to minimize the shift between the two domains and make the sparse coefficients discriminative for recognition task. We now describe each step in details below.

A. Orthogonal Dictionary Learning and Correspondence

In this work, we propose to learn domain specific orthogonal dictionaries which do not require one-to-one paired data thereby reflecting practical scenarios. In our work, during training, data in the two domains (VIS and NIR) are used to learn orthogonal dictionaries of their corresponding domains independently, thus the input data need not be paired. There are several advantages of learning orthogonal as compared to over-complete dictionaries as explained in [3]. In addition, another advantage of our approach is that since the dictionaries span corresponding space in the two domains and are orthogonal, the dictionary atoms can be aligned to minimize the shift between the two domains. The sparse coefficients computed from the aligned dictionaries can then be directly compared.

1556-6013 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

In our work, we use the orthogonal dictionary learning framework proposed by Bao *et al.* [3]. Given the data from one domain $\mathbf{X}$, we learn an orthogonal dictionary $\bar{\mathbf{D}}_x$ = $[\mathbf{A}_x$ $\mathbf{D}_x]$ such that $\bar{\mathbf{D}}_x^T$ $\bar{\mathbf{D}}_x$ = $\mathbf{I}_{d\times d}$. Here, d is the size of the input feature vector. The optimization function that can learn the orthogonal dictionary is formulated as follows:

$$\begin{aligned} &\min_{\mathbf{D}_x,\mathbf{\Lambda}} \| \mathbf{X} - [\mathbf{A}_x, \mathbf{D}_x]\,\mathbf{\Lambda}\|_2^2 + \alpha \| \mathbf{\Lambda}\|_0^2 \\ &\qquad \text{s.t. } \mathbf{D}_x^T\mathbf{D}_x = \mathbf{I}_m, \mathbf{A}_x^T\mathbf{D}_x = \mathbf{0}. \end{aligned} \tag{1}$$

Here m is the number of atoms in dictionary $\mathbf{D}_x$. The dictionary $\bar{\mathbf{D}}_x$ has two sub dictionaries, where $\mathbf{A}_x$ can be used to have a control on the required number of orthogonal dictionary atoms, and $\mathbf{D}_x$ represents the orthogonal dictionary atoms learned from the input data $\mathbf{X}$. We have set $\mathbf{A}_x$ to a null matrix so that $\bar{\mathbf{D}}_x$ = $\mathbf{D}_x$ and initialized $\mathbf{D}_x$ to DCT matrix of size $d \times d$. During the t^{th} iteration, we have $\mathbf{D}_x^t$ so that, the sparse vectors $\mathbf{\Lambda}_x^t$ can be updated as follows

$$\begin{aligned} \hat{\mathbf{\Lambda}}_x^t = \arg\min_{\mathbf{\Lambda}_x^t} &\left\| \mathbf{X} - \left[\mathbf{A}_x, \mathbf{D}_x^t\right] \mathbf{\Lambda}_x^t\right\|_2^2 + \alpha \| \mathbf{\Lambda}_x^t\|_0^2 \\ &\text{s.t. } \mathbf{D}_x^{t^T}\mathbf{D}_x^t = \mathbf{I}_m, \mathbf{A}_x^T\mathbf{D}_x^t = \mathbf{0}. \end{aligned} \tag{2}$$

The problem in (2) has a unique solution given by [3]

$$\hat{\mathbf{\Lambda}}_x^t = T_\epsilon(\bar{\mathbf{D}}_x^{t^T}\mathbf{X}) \tag{3}$$

Here $T_\epsilon(\mathbf{v})$ represents a hard threshold operation on the vector $\mathbf{v}$. The objective function to update the dictionary at t^{th} iteration can be formulated as

$$\begin{aligned} \hat{\mathbf{D}}_x^t = \arg\min_{\mathbf{D}_x^t} &\left\| \mathbf{X} - \left[\mathbf{A}_x, \mathbf{D}_x^t\right] \mathbf{\Lambda}_x^{t-1}\right\|_2^2 \\ &\text{s.t. } \mathbf{D}_x^{t^T}\mathbf{D}_x^t = \mathbf{I}, \mathbf{A}_x^T\mathbf{D}_x^t = \mathbf{0}. \end{aligned} \tag{4}$$

The above objective function has a unique solution and is given by $\hat{\mathbf{D}}_x^t = \mathbf{U}\mathbf{V}^T$. The matrices $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices which can be obtained from Singular Value Decomposition given by $(\mathbf{I}_d - \mathcal{P}_A)\mathbf{X}\mathbf{\Lambda}_{xD}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Here, $\mathbf{\Lambda}_{xD}^T$ is the sparse coefficient vector at iteration $(t-1)$ associated with dictionary $\mathbf{D}_x$ and $\mathcal{P}_A$ is an orthogonal projection operator defined by $\mathcal{P}_A\boldsymbol{u}$ = $\mathbf{A}(\mathbf{A}^T\boldsymbol{u})$, $\forall\ \boldsymbol{u} \in \mathcal{R}^d$. Using the above formulation, we separately compute two dictionaries $\mathbf{D}_x$ and $\mathbf{D}_y$ for the data $\mathbf{X}$ and $\mathbf{Y}$ from the two domains.

Since both the dictionaries have m orthogonal atoms computed from data of same classes (in our case, m = d since we set $\mathbf{A}_x$ to a null matrix), they should span the same space. But, since they are learnt separately, there may not be any correspondence between the dictionary atoms of $\mathbf{D}_x$ and $\mathbf{D}_y$, i.e. i^{th} dictionary atom of $\mathbf{D}_x$ may not correspond to the i^{th} atom of $\mathbf{D}_y$. So now the goal is to find the corresponding dictionary atoms, so that they can be aligned. So we transform the columns of both the dictionaries to a common space by using cluster CCA [34] learned from the training data. Cluster CCA is used since it only requires correspondence between clusters of data in the two domains and not one-to-one correspondence between the data points. Then Bipartite Graph Matching [4] is used to learn the correspondence of the dictionary atoms in the transformed space. Given a set of m dictionary atoms from $\mathbf{D}_x$ (denoted as $\mathbf{d}_x^i$, $i = 1, \ldots, m$) and m dictionary atoms from $\mathbf{D}_y$ (denoted as $\mathbf{d}_y^j$, $j = 1, \ldots, m$), the goal is to establish the correspondence between them. Let C_{ij} be the cost of matching the two vectors $\mathbf{d}_x^i$ and $\mathbf{d}_y^j$. The objective function of the bipartite graph matching is as follows

$$\mathbf{H}(\pi) = \sum_i \mathrm{C}(\mathbf{d}_y^i, \mathbf{d}_x^{\pi(i)}) \tag{5}$$

The above objective function is minimized which gives the correspondence between the dictionary atoms of the two domains. $\pi(i)$ is essentially a permutation operation, which is used to permute the atoms of one dictionary such that the dictionary atoms of the two domains have one-to-one correspondence. Note that, the dictionaries are transformed to the common space only to apply bipartite graph matching. Once the correspondence matching $\pi(i)$ is obtained, we apply the permutation on the original dictionary $\mathbf{D}_x$. Let these dictionaries be denoted by $\mathbf{D}_x^c$ and $\mathbf{D}_y^c$, where $\mathbf{D}_x^c$ is obtained by permuting the atoms of $\mathbf{D}_x$ according to $\pi(i)$, and $\mathbf{D}_y^c = \mathbf{D}_y$.

B. Dictionary Alignment, Discriminative coefficients

Once we compute the permuted dictionaries $\mathbf{D}_x^c$ and $\mathbf{D}_y^c$ which have one-to-one correspondence between their columns, we need to address the domain shift of the dictionaries. We reduce this domain shift by aligning one of the dictionary with respect to the other so that we can have a common sparse representation for cross domain data. Inspired by the success of subspace alignment approach [7], we learn a mapping function $\mathbf{T}$ on $\mathbf{D}_x^c$ so that $\mathbf{T} : \mathbf{D}_x^c \rightarrow \mathbf{D}_y^c$. We formulate the objective function to learn the mapping function $\mathbf{T}$ as below

$$\hat{\mathbf{T}} = \arg\min_{\mathbf{T}} \left\| \mathbf{D}_x^c\mathbf{T} - \mathbf{D}_y^c\right\|_2^2 \tag{6}$$

By solving the above objective function, the optimum value of $\mathbf{T}$ can be derived as a closed form solution as $\hat{\mathbf{T}} = \mathbf{D}_x^{c\,T}\mathbf{D}_y^c$. Thus, the x-domain dictionary which is aligned to the y-domain is given by $\mathbf{D}_x^{c,a} = \mathbf{D}_x^c\mathbf{D}_x^{c\,T}\mathbf{D}_y^c$.

Though the dictionaries are aligned to reduce the domain shift, the corresponding sparse coefficient vectors need to be made discriminative so that they can be used for the recognition task. The data from the $\mathbf{X}$ and $\mathbf{Y}$ domains are first used to compute their sparse coefficients as follows

$$\begin{aligned} &\arg\min_{\mathbf{\Lambda_x}} \| \mathbf{X} - \mathbf{D}_x^{c,a}\mathbf{\Lambda}_x\|_2^2 + \alpha \| \mathbf{\Lambda}_x\|_0^2 \\ &\arg\min_{\mathbf{\Lambda_y}} \left\| \mathbf{Y} - \mathbf{D}_y^c\mathbf{\Lambda}_y\right\|_2^2 + \alpha \| \mathbf{\Lambda}_y\|_0^2 \end{aligned} \tag{7}$$

To make the sparse coefficients discriminative, we apply the metric learning approach KISSME [19] on the sparse coefficients of the two domains so that the coefficient vectors of the same class moves closer and of different classes move apart. Please refer to [19] for further details of the algorithm.

IV. Proposed Re-ranking Approach and Testing

During the testing stage, the probe and the gallery features represented by the discriminative sparse coefficient vectors are computed, with which a rank list is obtained. The rank list gives the gallery data with increasing distance from the probe. The proposed re-ranking algorithm uses this rank list along with that given by any other approach to compute an improved

1556-6013 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2868173, IEEE Transactions on Information Forensics and Security

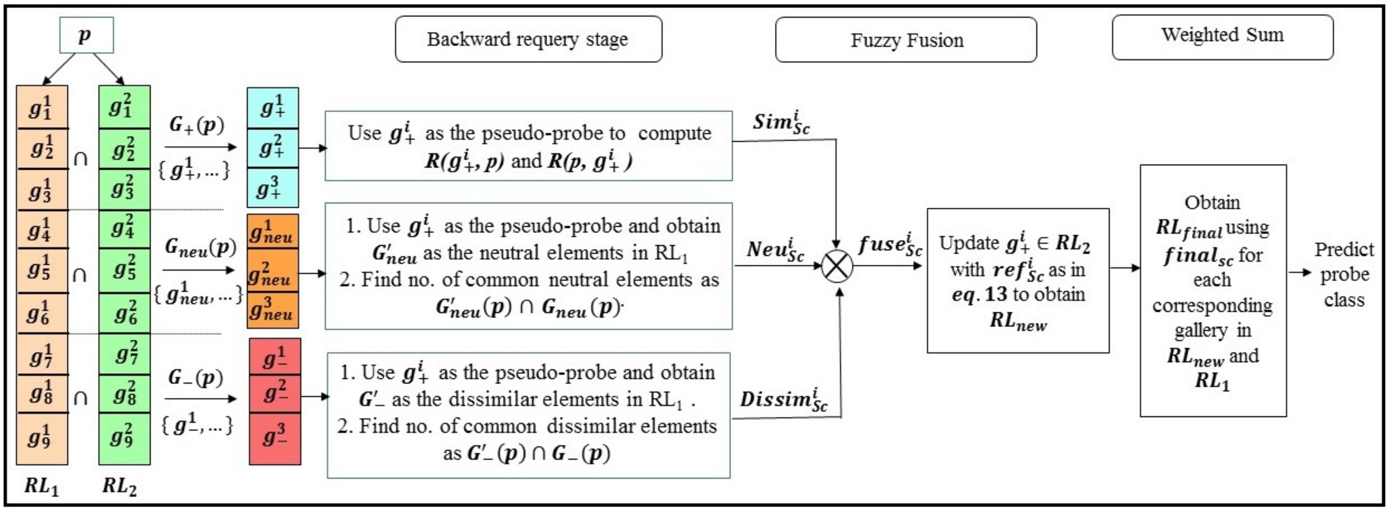
p
g_1^1 g_2^1 g_3^1 g_4^1 g_5^1 g_6^1 g_7^1 g_8^1 g_9^1
g_1^2 g_2^2 g_3^2 g_4^2 g_5^2 g_6^2 g_7^2 g_8^2 g_9^2
$\cap$
RL_1
RL_2
$G_+(p)$
$\{g_+^1, ...\}$
$G_{neu}(p)$
$\{g_{neu}^1, ...\}$
$G_-(p)$
$\{g_-^1, ...\}$
g_+^1 g_+^2 g_+^3
g_{neu}^1 g_{neu}^2 g_{neu}^3
g_-^1 g_-^2 g_-^3
Backward requery stage
Use g_+^i as the pseudo-probe to compute $R(g_+^i, p)$ and $R(p, g_+^i)$
1. Use g_+^i as the pseudo-probe and obtain G'_{neu} as the neutral elements in RL$_1$
2. Find no. of common neutral elements as $G'_{neu}(p) \cap G_{neu}(p)$.
1. Use g_+^i as the pseudo-probe and obtain G'_- as the dissimilar elements in RL$_1$.
2. Find no. of common dissimilar elements as $G'_-(p) \cap G_-(p)$
Sim_{Sc}^i
Neu_{Sc}^i
$Dissim_{Sc}^i$
Fuzzy Fusion
$fuse_{Sc}^i$
Update $g_+^i \in RL_2$ with ref_{Sc}^i as in $eq.$ 13 to obtain RL_{new}
Weighted Sum
Obtain RL_{final} using $final_{sc}$ for each corresponding gallery in RL_{new} and RL_1
Predict probe class

Fig. 2: Illustration of the proposed re-ranking framework.

rank list, which is then used to predict the probe class. For discussion, we will refer to the proposed DAlign as Algo 1 and the other approach as Algo 2. The proposed re-ranking algorithm is general and can be used with any two algorithms.

Consider that the probe is denoted as p and the ranked list of gallery retrieved from Algo 1 and Algo 2 are denoted as RL_1 = $\{g_1^1, g_2^1, g_3^1 .... g_N^1\} \in \mathcal{R}^{d \times N}$ and RL_2 = $\{g_1^2, g_2^2, g_3^2 .... g_N^2\} \in \mathcal{R}^{d \times N}$ respectively. Here, d is the feature dimension and N is the total number of gallery images. The rank list RL_t ($t \in \{1, 2\}$) satisfies the condition $d(p, g_1^t) < d(p, g_2^t) < ..... < d(p, g_N^t)$, where, $d(p, g_i^t)$ ($i \in \{1, 2, 3, ..., N\}$) denotes the Euclidean distance between the probe p and the i^{th} gallery data. It is important to note that in general, g_i^1 may or may not be the same as g_i^2. Using the probe to retrieve the gallery elements is referred to as forward query.

The proposed re-ranking approach has two main steps. First, given a probe, based on the arrangement of the gallery data in the two rank lists, we divide them into three sets, namely strongly similar, strongly neutral and strongly dissimilar with respect to the given probe. Second, using the galleries in the strongly similar set, we refine Algo 2 rank list using backward requery. This refined rank list is fused with that of Algo 1 using weighted sum to compute an improved rank list which is finally used to predict the probe class.

A. Strongly similar, dissimilar and neutral gallery

Now, we discuss the procedure to compute the strongly similar, strongly neutral and strongly dissimilar gallery sets.

Strongly Similar Score: We assume that for both the approaches, even if the correct match is not in rank-1, it will be there in the top k retrieved results in RL_1 and RL_2. This is true for most of the algorithms and is assumed to be true for most re-ranking approaches [21][24]. For example, for DAlign, even though the rank-1 accuracy is 58.38% on the modified CASIA NIR-VIS 2.0 database, it improves to 80.88% for rank-5 and 87.52% for rank-10, which is further discussed in Section VI. Thus if a gallery is present in the top k retrieved results in both the rank lists, i.e. it lies in the intersection of the top k elements of the two rank lists, it is likely to be the correct match for the particular probe and is considered as an element of the strongly similar gallery set. Let $S_{k_+}^1(p)$ and $S_{k_+}^2(p)$ denote the top k gallery elements for the given probe p in RL$_1$ and RL$_2$ respectively, then the strongly similar gallery set is given by

$$G_+(p) = S_{k_+}^1(p) \cap S_{k_+}^2(p) \tag{8}$$

Here, $G_+(p) = \{g_+^1, g_+^2, g_+^3 ....\}$ denotes the strongly similar gallery elements which is likely to contain the correct match for the particular probe and is thus treated as the new gallery set from this point onwards. Now, each element in this new gallery $g_+^i \in G_+(p)$ is used as a pseudo-probe after replacing this gallery element with the original probe which is known as backward requery inspired from [49]. For example, when we use g_+^1 as the pseudo-probe, the gallery consists of $\{g_+^2, g_+^3, ..., p\}$, i.e. all the elements of $G_+(p)$ without g_+^1, in addition to the original probe. This is based on the assumption that if g_+^i is the correct gallery for the probe p, then p should also show up in the top few retrieved elements when g_+^i is used as the pseudo-probe for requery. This assumption has been used in the re-ranking task with great success in [47]. This process is similarly repeated for all other $g_+^i \in G_+(p)$.

The modified similarity between g_+^i and the original probe p is computed from the rank (or position) in the two scenarios (forward query and backward requery) as follows

$$Sim_{Sc}^i = \frac{\gamma}{R(g_+^i, p) \times R(p, g_+^i)} \tag{9}$$

Here, R(p, g_+^i) denotes the rank (or position) of g_+^i during forward query, and $R(g_+^i$, $p)$ denotes the rank of the probe p during backward requery. For the correct gallery, both these ranks should be low, which will result in a high similarity score of the gallery for that probe. We experimentally set k to 40 and γ, which is the weighing factor to 0.001.

Strongly Dissimilar Score: The strongly dissimilar gallery score is used to penalize the gallery elements in the new gallery set $G_+(p)$, which we believe are not a suitable match

1556-6013 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

for the probe. The score assigned to the dissimilar gallery elements in this set pushes it away from the probe to output an improved rank list. The strongly dissimilar gallery set is obtained using the intersection between the last *k* gallery elements of RL$_1$ and RL$_2$. If $S^1_{k\_}(p)$ and $S^2_{k\_}(p)$ denote the last *k* gallery elements in RL$_1$ and RL$_2$ respectively, then the strongly dissimilar gallery set is formulated as

$$G_-(p) = S^1_{k\_}(p) \cap S^2_{k\_}(p) \tag{10}$$

Here $G_-(p) = \{g^1_-, g^2_-, g^3_- ....\}$ denotes the strongly dissimilar gallery elements, which are used to penalize the strongly similar gallery elements $g^i_+ \in G_+(p)$ which are not likely to be a correct match for the given probe.

We assume that if a gallery g^i_+ is the correct match for the probe p, then the number of common dissimilar elements when probe p is used for forward query and when g^i_+ is used for backward requery will be high. When probe p is used for forward query, instead of taking the last k retrieved elements as the dissimilar elements, we take the strongly dissimilar set since these are more likely to be the correct dissimilar elements. So, we requery g^i_+ back into the rank list obtained by Algo 1, and let the last *k* gallery elements be denoted as $G'_-(p)$. We then find the number of common elements between $G'_-(p)$ and $G_-(p)$ which is used to assign the dissimilarity score to $g^i_+ \in G_+(p)$ as follows

$$Dissim^i_{Sc} = \alpha_1 * e^{-|G_-(p) \cap G'_-(p)|} \tag{11}$$

where $|.|$ denotes the cardinality and α_1 represents the weighing factor which we experimentally set to 1.0. Higher number of common dissimilar elements indicates that the gallery is likely to be a correct match for the probe and so the dissimilarity score is smaller, and vice versa.

Strongly Neutral Score: The strongly neutral gallery elements are computed so as to lightly penalize the gallery elements in $G_+(p)$, which we believe are neither strongly similar nor dissimilar to the probe. We obtain these elements using the intersection of the intermediate *k* gallery elements of RL$_1$ and RL$_2$. Using the same assumption and the same process as we followed for the strongly dissimilar score, we assign the neutral score for g^i_+ denoted as $Neut^i_{Sc}$.

B. Final rank list using Fuzzy fusion

Using these computed scores, we refine the rank list obtained from Algo 2. First, we use fuzzy aggregation operation to fuse these scores for g^i_+ resulting in a distance score [48]

$$fuse^i_{Sc} = 1 - ({Sim^i_{Sc}}^{\alpha} + {Neut^i_{Sc}}^{\alpha} + {Dissim^i_{Sc}}^{\alpha})^{1/\alpha} \tag{12}$$

Fuzzy aggregation is an averaging function which is an effective tool to fuse multiple scores. With variations in α, this operator can be converted into arithmetic mean or geometric mean. Here, all scores have been given equal weights where we experimentally set $\alpha = 1.9$. Thus, each element in the strongly similar gallery set is assigned this distance score ($fuse_{Sc}$) with which we aggregate its corresponding distance score $d(p, g^2)$ from RL_2 since our goal is to refine rank list of Algo 2. This aggregation gives a modified distance score for these gallery elements in the strongly similar gallery set as shown below

$$ref_{Sc} = ((fuse_{Sc})^{\alpha} + d(p, g^2)^{\alpha})^{1/\alpha} \tag{13}$$

Thus, all $g_+ \in RL_2$ is updated with ref_{Sc}, while scores of the gallery elements which do not belong to the strongly similar gallery set remains unchanged. This results in a refined rank list for Algo 2 which is denoted as RL_{new}. Finally, for each gallery element, the distance scores from RL_{new} and RL_1 are combined using weighted sum as follows:

$$final_{Sc} = \lambda d(p, g^1) + (1 - \lambda) ref_{Sc} \tag{14}$$

Here, λ is the weight factor which we experimentally set as 0.7. This distance score $final_{Sc}$ is used to generate the ranklist RL_{final} which is used to predict the probe class.

V. Database Description

In this section, we describe the different databases available which address the challenging problem of matching NIR with VIS faces along with the details of our collected database.

Available databases for heterogeneous face recognition: Li *et al.* [23] collected the CASIA NIR-VIS 2.0 database containing 725 subjects captured under controlled indoor settings with 1-22 VIS and 5-50 NIR high-resolution face images per subject. Since it contains large number of subjects, it is one of the most widely used datasets for heterogeneous face recognition. But it does not contain surveillance quality images having low-resolution and large variations in pose. Grgic *et al.* [9] collected the Surveillance Cameras face database (SCface) which contains images taken in uncontrolled indoor scenario using three different NIR cameras. There are a total of 4,160 static images from both NIR and VIS spectrum of 130 subjects. Though this dataset is very useful for addressing the surveillance quality scenario, the number of subjects is not too high and also it is difficult to systematically study the effect of low-resolution and pose on the recognition performance. Kang *et al.* [15] propose the Long Distance Heterogeneous Face Database (LDHF-DB) comprising of both VIS and NIR face images. It contains 100 subjects, where for each subject, one frontal image without glasses is captured at each distance during day and night time. In practical scenarios, we encounter non-frontal images with variations in people wearing glasses of different styles which adds to the challenge in matching NIR faces with VIS faces. Maeng *et al.* [28] propose a Near-Infrared Face Recognition at a Distance Database (NFRAD-DB) consisting of NIR images with 50 subjects. In addition to small number of subjects, the images have subtle variations in pose. Singh *et al.* [41] introduce the cross-spectral, cross-resolution video database with 160 subjects consisting of videos captured under uncontrolled settings for both NIR and VIS spectrum. Bourlai *et al.* [42] introduce the SWIR (Short-wave infrared), MWIR (mid-wave infrared) and NIR Mid-Range datasets. These three datasets are captured with 50, 50 and 103 subjects comprising of images with frontal poses. A summary of these databases is given in Table I.

1556-6013 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2868173, IEEE Transactions on Information Forensics and Security

TABLE I: Summary of the available databases along with the proposed HPR database for matching NIR and VIS face images.

Name of database	No. of subjects	Videos	Pose variation	Environment	Image Resolution	Distance variation (Range in meters)
CASIA NIR-VIS 2.0 [23]	725	No	Low	Indoor	High	N/A
HPR database	**200**	**400 (NIR)**	**Extreme**	**Indoor**	**Low**	**1 - 5**
Cross-spectral cross-resolution video database [41]	160	98 (VIS) & 95 (NIR)	Extreme	Indoor & Outdoor	Low	0 - 10
SCface database [9]	130	No	Low	Indoor	Low	1 - 4.2
LDHF-DB [15]	100	No	Frontal Only	Indoor & Outdoor	Low	1, 60, 100, 150
NFRAD-DB [28]	50	No	Moderate	Indoor	Low	1, 60
SWIR [42]	50	No	Frontal Only	Indoor	Low	50, 106
MWIR [42]	50	No		Indoor	Low	2
NIR Mid-Range [42]	103	515 (NIR)		Outdoor	Low	30, 60, 90, 120

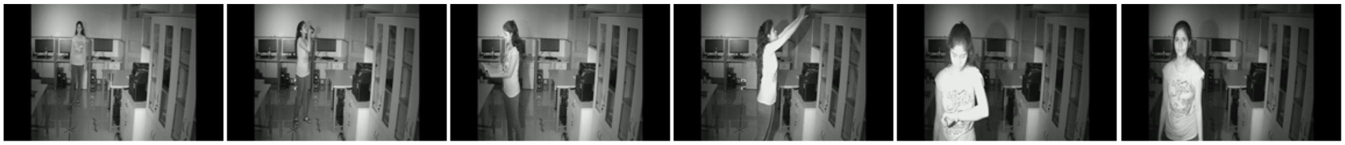

Fig. 3: Sample video frames from one video from our HPR database captured by NIR camera.

Proposed HPR Database: Taking into account the advantages and limitations of the available databases, we introduce a new database termed as HPR keeping two goals in mind, (1) It should be able to reflect the practical surveillance scenario and (2) It should help in systematic study of the effect of resolution and pose in the recognition performance. The database has images from 200 subjects captured with two surveillance quality cameras which can work under low light conditions in NIR mode and one HR VIS camera which captures mugshot photos.

Data Collection Details: For data collection, we use two NIR cameras, namely, CP Plus PTZ Camera with 2 MP resolution (NIR Cam01) and HIK Vision Bullet Camera with 3 MP resolution (NIR Cam02). A 54 array NIR LED illuminator is mounted on the same platform in between the two NIR cameras for providing active lighting source in the dark. Canon PowerShot Pro Series S5 IS 8.0MP digital camera is used for capturing HR VIS images. The VIS images are captured under normal lighting conditions but the NIR images are captured by switching off all the light sources, with the NIR illuminator on. We take the HR VIS mugshot gallery images of subjects at 4ft. distance from the camera at seven poses (Fig. 1), which accounts to a total of 1400 HR VIS images. The NIR still images are captured when the subjects stand at three different distances (8ft., 12ft. and 16ft.) from the two NIR cameras at seven different poses to systematically study the effect of distance and pose on the performance, thus accounting to a total of 4200 images from each camera. Fig. 1 shows images taken from the first camera for different poses and distances. The resolution of the face images decreases as the subject moves further from the camera allowing us to study the effect of varying resolution on the performance. For the NIR videos, the subjects perform different tasks like drinking water, reading a book, reaching out for something from a cupboard and looking at the watch. They are not directed to look at specific locations and they are free to do the task as they want, which results in significant variations in the videos, thus reflecting the practical scenarios (Fig. 3). Most of the volunteers for our database are university students and the data capture protocol was approved by the Institute Human Ethics Committee.

Once the images are captured, Viola Jones face detection technique [43] is used to detect and crop the face region in the image, which are then manually verified and corrected. The summary of the proposed database is presented in Table I. The database has the following interesting points:

- It has real LR NIR images and HR VIS images of 200 subjects to address the problem of low-resolution heterogeneous face recognition with extreme poses which are very difficult to recognize.
- The database is useful for systematic evaluation and understanding of how variations in pose and resolution affect the performance of heterogeneous face recognition.
- It also has videos of 200 subjects captured while they are performing various tasks.
- Experimental protocols are defined for our database and are benchmarked with available methods in literature.

VI. Experimental Evaluation

Extensive experiments are conducted to evaluate the usefulness of the proposed algorithms for the task of low-resolution heterogeneous face recognition task along with analysis on the collected dataset. We first evaluate the performance of matching synthesized LR NIR images with HR VIS faces using the modified CASIA NIR-VIS 2.0 database [23]. Then, we evaluate the performance of the algorithms on real LR images of the SCface database [9] and our database. The proposed re-ranking approach is evaluated by combining DAlign with the other algorithms to justify its usefulness. In our experiments, we concatenate *pool5*, *fc6* and *fc7* features from VGG face

1556-6013 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2868173, IEEE Transactions on Information Forensics and Security

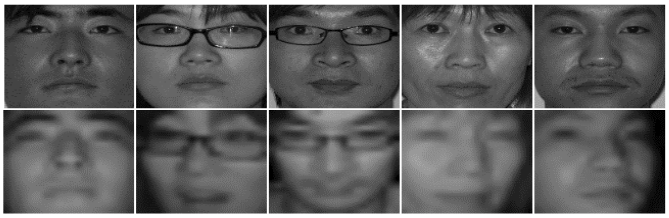

Fig. 4: Sample images from the modified CASIA NIR-VIS 2.0 database. Row 1: HR VIS gallery images; Row 2: Corresponding LR NIR images.

network [33] and apply PCA to reduce the dimension to 2500 which is used as the feature vector [29].

A. Experiments on modified CASIA NIR-VIS 2.0 Database

We choose CASIA NIR-VIS 2.0 database [23] since it is the largest publicly available database with 725 subjects. The HR VIS images are of size 128×128 while the NIR images are downsampled to 20×20 and then upsampled back to the original size using bi-cubic interpolation to simulate the LR images. Sample images of the database used in our experiments are shown in Fig. 4. We follow the standard protocol of the database and report the mean recognition rate and standard deviation over 10 folds for the proposed dictionary alignment (DAlign) approach. The comparisons with other approaches are reported in Table II. We observe that the performance of DAlign is 58.38%, which is higher than all the other compared approaches. But this is considerably lower than the performance of 87.45%, when the probe images were HR as reported in [29], which illustrates the difficulty of the problem addressed in this work.

We also evaluate the proposed re-ranking approach using DAlign along with other compared approaches taken as Algo 2 whose results are reported in Table II. We observe that the proposed re-ranking framework is able to consistently outperform both DAlign and the other baseline. For this experiment, the best rank-1 performance of 68.30% is obtained by applying re-ranking on DAlign and CBFD. We compare our approach with recent state-of-the-art re-ranking algorithms [30], [38] and [2] and the results are reported in Table II.

During testing, for a given probe feature, it takes 0.1022 and 0.5331 seconds to identify the probe ID using DAlign and CBFD respectively from 358 gallery images. Given their scores, the re-ranking takes 1.5294 seconds to output the probe class. This is the CPU computational time on a machine with 32GB DDR4 RAM, 2.4GHz clock frequency and 64bit OS using an unoptimized MATLAB code.

B. Analysis of the Proposed Approach

Here, we analyze the proposed algorithms on fold 3 of the modified CASIA NIR-VIS 2.0 database and take the best performing methods from Table II. The rank-1 recognition on fold 3 for DAlign, randKCCA, CBFD and CA-LBFL are 57.56%, 45.60%, 42.80% and 46.29% respectively.

Comparison with Score Fusion Techniques: Here, we illustrate the effectiveness of the proposed re-ranking approach as compared to some of the standard score based fusion techniques. We use four fusion techniques, namely Weighted Sum, Max. Fusion [17], Min. Fusion [17] and Product Fusion [50] to fuse the similarity scores for comparison. We observe from Table III that the proposed re-ranking approach combines the scores effectively to give better performance than the standard fusion techniques.

TABLE II: Evaluation (%) on the modified CASIA NIR-VIS 2.0 database [23]. Notation: (A+B) denotes the performance of the re-ranking approach using the approaches A and B.

Algorithm	Rank-1±std	Rank-5±std	Rank-10±std
randKCCA [25]	45.42±1.51	71.63±1.07	80.45±0.80
CA-LBFL [6]	44.41±1.12	63.74±1.22	71.81±1.26
ClusterCCA [34]	42.64±1.61	69.00±1.22	78.41±0.82
CBFD [27]	42.47±0.97	61.92±1.08	70.26±0.99
meanCCA [34]	41.82±1.87	67.72± 1.22	77.10±1.00
GMA [40]	41.56±1.72	67.71±1.50	77.57±0.98
DAlign [29]	**58.38±1.19**	**80.88±1.06**	**87.52±0.72**
DAlign + randKCCA	64.29 ±1.66	84.57±0.79	89.44±0.75
DAlign + ClusterCCA	62.32±1.28	82.79±0.78	87.94±0.50
DAlign + meanCCA	61.90±1.33	82.21±0.87	87.22±0.46
DAlign + CCA	62.18±1.24	82.51±0.95	87.78±0.61
DAlign + GMA	62.39±1.27	82.74±1.10	88.04±0.87
DAlign + CA-LBFL	68.21±0.87	83.49±0.76	89.70±0.43
DAlign + CBFD	**68.30±1.17**	**82.74±0.93**	**90.14±0.62**
Re-ranking on LR [30]	60.21±1.26	82.64±1.02	88.29±0.63
ECN re-ranking [38]	66.84±2.39	82.25±1.31	84.71±1.22
RDP re-ranking [2]	61.13±1.19	82.71±1.05	88.69±0.72

Effect of different terms: Here we analyze the effect of each of the following terms, namely, Sim_{Sc} (S_1), $Dissim_{Sc}$ (S_2) and $Neut_{Sc}$ (S_3) on the performance of the proposed re-ranking framework using DALign with CBFD as the two baselines. We conduct an experiment on fold 3 of the modified CASIA NIR-VIS 2.0 database by adding one term at a time. We obtain a rank-1 performance of 65.99% using S_1 alone, adding S_2 improves it to 66.17%, and finally combining S_3 along with S_1 and S_2 gives the highest performance of 67.80%. We observe that, each term helps in improving the recognition performance, though the most significant improvement is given by the first term.

The proposed re-ranking approach is also able to effectively utilize multiple scores from more than one algorithm to further improve the result. We perform an experiment by taking DAlign as Algo1 and three other algorithms, randKCCA, CBFD and CA-LBFL as secondary baseline algorithms. We observe that the rank-1 recognition rate improves to 70.30%, while rank-5 recognition rate increases to 88.19% and rank-10 recognition becomes 92.94%. But all the results in the paper are with one secondary algorithm.

1556-6013 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2868173, IEEE Transactions on Information Forensics and Security

TABLE III: Comparison of the proposed re-ranking algorithm (Rank-1%) with standard fusion techniques on fold-3 of the modified CASIA NIR-VIS 2.0 database [1].

Algorithms	Weighted Sum	Max. Fusion	Min. Fusion	Product Fusion	**Re-ranking**
DAlign + GMA	56.26	56.60	46.63	55.93	**62.39**
DAlign + CCA	59.96	47.61	57.97	59.67	**62.18**
DAlign + MeanCCA	58.17	51.95	54.55	57.53	**61.90**
DAlign + ClusterCCA	58.77	53.74	54.62	58.77	**62.32**
DAlign + randKCCA	58.50	58.79	45.62	55.50	**64.29**
DAlign + CBFD	60.88	61.14	59.25	56.87	**68.30**
DAlign + CA-LBFL	60.59	52.96	58.85	58.71	**68.21**

Effect of Super-resolution: One standard way of matching LR with HR images is to first apply super-resolution (SR) on LR image to synthesize the corresponding HR image and then perform matching. For this task, we apply the state-of-the-art deep SR technique proposed by Lai *et al.* [20] on the LR images to synthesize the corresponding HR images. In this work, a deep Laplacian pyramid network based framework is proposed [20] to progressively reconstruct sub-band residuals of the HR images. Since the trained models for scaling factors of 2 and 4 are provided, we use LR NIR input images of resolution 20×20, 30×30 and 60×60. For 30×30 and 60×60, we use the scaling factor of 2 and 4 directly to upsample them back close to its original resolution. For image resolution of 20×20, we apply the SR technique with scaling factor of 4 to obtain a 80×80 SR image, which is then resized back to the original resolution of 128×128.

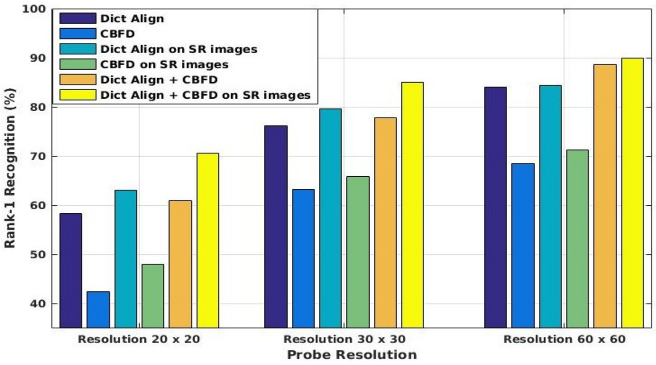

Fig. 5: Effect of applying SR [20] on the LR NIR faces before performing the recognition.

For this experiment, we chose DAlign and CBFD [27] as the baselines for the proposed re-ranking approach. We observe from Fig. 5 that the SR technique is able to improve the rank-1 performance of the individual baseline algorithms. We also observe that the SR images helps to further boost the performance of the proposed re-ranking approach. For comparison, when HR NIR images are used as probe and HR VIS images are used as gallery, we obtain rank-1 performance of 84.90%, 72.68% and 92.01% with DAlign, CBFD and DAlign with CBFD respectively.

Effect of Face hallucination: Here, we compare the proposed approach with the state-of-the-art heterogeneous face recognition approach which uses cross-spectral hallucination and low-rank embedding [22] referred to as CSH. For this experiment, as in [22], we use the three channel RGB and NIR images of the database [23] which are cropped to 224×224 and aligned using the fiducial locations as discussed in [16]. Here, we want to analyze two things, (1) How does our approach compare with the state-of-the-art deep neural network based technique [22] and (2) Can the proposed re-ranking approach use this technique along with DAlign to give a performance which is better than both the approaches.

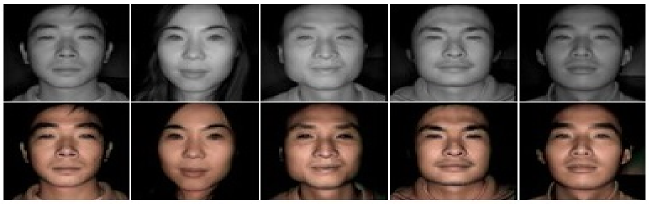

Fig. 6: Few examples of hallucinated faces on CASIA NIR-VIS 2.0 database. Top Row: Output of SR [20] on LR images of size 112×112, Second row: Corresponding hallucinated images after applying CSH [22].

TABLE IV: Rank-1 recognition rates (%) obtained for two different features using CSH [22] and DAlign.

Algorithm	CSH [22]	DAlign	DAlign + CSH [22]
Feat 1	76.29	85.70	**89.88**
Feat 2	93.78	96.55	**97.96**

For the experiment, we downsample the original NIR images to create the LR NIR probe images of resolution 112×112. (We also tried with image resolution of 20×20, but the hallucinated images obtained when the super-resolution images using a scaling factor of 8 were given as input were distorted and the performance was quite low.) We then apply the state-of-the-art SR technique [20] with a scale factor of 2 to create the HR images and then apply the cross-spectral hallucination method [22]. Sample LR NIR images and their corresponding hallucinated images are shown in Fig. 6. We observe that for the LR image resolution of 112×112, the hallucinated images look quite natural. Then the VGG Face network is used to extract the features of 1024d (Feat 1) as discussed in [22] after which low-rank embedding is applied to obtain the recognition performance. We directly use the code provided by the authors [20] [22]. We observe from

1556-6013 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

TABLE V: Rank-1 recognition (%) on SCface database [9].

Algorithm	Dist 1	Dist 2	Dist 3
DAlign (Feat 2)	13.12	32.50	46.87
DAlign (Feat 3)	33.75	32.50	33.75
DAlign (Feat 2) + DAlign (Feat 3)	**34.37**	**39.38**	**49.37**

Table IV (Row 1) that the proposed DAlign approach gives a better rank-1 recognition performance compared to the state-of-the-art approach for LR heterogeneous face recognition. We also observe that the proposed re-ranking approach is able to combine the rank lists given by DAlign and CSH to further improve the recognition performance.

We repeat the same experiments with our feature (by concatenating *pool5*, *fc6* and *fc7* followed by PCA) denoted by Feat 2, and the results are reported in Table IV (Row 2). We observe similar pattern in the results, except that these results are significantly higher as compared to those using Feat 1.

C. Experiments on SCface database:

In this section, we present the experimental evaluation on the SCface database [9] to illustrate the effectiveness of the proposed approaches on surveillance quality heterogeneous face images. We randomly take 50 subjects for training and the remaining 80 subjects for testing. Both the VIS and NIR images are resized to 224×224 to compute the feature (Feat 2). There are images from two IR cameras captured at three distances. We conduct the experiment on images from both cameras together under each distance separately to illustrate the effect of distance variations on the recognition performance. The frontal HR VIS images are taken as gallery and the NIR images from both the cameras are taken as probe. Here, the probe images are themselves of surveillance quality.

We observe that both CBFD [27] and CA-LBFL [6] perform poorly on these images. This may be because these approaches work on pixel difference vectors and the images are of poor quality. We also examine the performance of the other approaches reported in Table II, but all of them perform poorly (around 5%). So for the re-ranking algorithm, instead of taking two approaches, we explore whether different features with the same algorithm can be used as inputs to the re-ranking approach. We conduct experiments with the proposed DAlign using two different image features, Feat 2 and rootSIFT [26] (Feat 3). We consider DAlign with Feat 2 and Feat 3 features as Algo 1 and Algo 2 respectively for the re-ranking approach. We report the rank-1 recognition accuracies (%) for the three distances in Table V. Interestingly, the performance of Algo 1 is less than Algo 2 for Dist 1, same as Algo 2 for Dist 2 and better than Algo 2 for Dist 3. But the proposed re-ranking approach is able to successfully improve the recognition performance as compared to both Algo 1 and Algo 2 for matching LR NIR images with good quality VIS images.

D. Experiments on the proposed HPR database

In this section, we provide details of the different evaluation protocols and the results obtained by the different state-of-the-art algorithms on our database. We provide three challenging protocols by which we can systematically analyze the effect of variations in head pose and the distance between the camera and subject (or resolution) on the recognition performance. In all our experiments, we consider the HR VIS faces taken under frontal pose as the gallery. Thus, both during training and testing, there is only a single image per subject for the VIS domain. Out of the 200 subjects, 50 subjects are taken for training and the remaining 150 subjects are considered for testing. Though we have landmark points for each facial image in our database, we did not use this information for our experiments. We use rank-1 recognition rate (%) as the performance metric for all the experimental protocols on our database. For this dataset also, the best algorithms in Table II, namely CBFD [27] and CA-LBFL [6] performed poorly. So we use randKCCA as Algo 2 to demonstrate the performance of our re-ranking framework. In this work, all the experiments are on the still images.

Protocol 1. Effect of Varying Resolution: This protocol is designed to analyze how the distance between camera and the subject affects the recognition performance. For this protocol, all the poses captured at a particular distance by both the NIR cameras are considered as the probe images. The results of the proposed dictionary alignment approach and the proposed re-ranking approach (taking randKCCA as the other baseline) is shown in Table VI for different distances. We observe that the performance drops as the distance between camera and subject increases. For example, for the proposed re-ranking approach, the performance drops from 60.62% to 56.91% when the distance increases from 8 ft. to 12 ft., and further drops to 50.52% when the distance increases to 16 ft. This is because, as the subject moves far from the camera, the size of face becomes smaller and hence the available discriminative information in the face is less, thereby resulting in lower recognition performance.

Protocol 2. Effect of Pose variations: This protocol is designed to analyze the variations of head pose on the recognition performance. For a particular distance (say 8 ft.), we take one pose (say Pose00) captured from both the NIR cameras as probe images. The experiment is repeated for each distance and each pose to observe the effect of pose variations. We also consider an even more challenging setting, where we fix the pose and take all the images captured at all the three distances together as probe. We repeat this setting over all the poses. Each of the first three rows in each block in Table VI shows the performance of the corresponding algorithm for the seven different head poses for a particular distance. The fourth row shows the performance for all the three distances combined for the different poses. As expected, the performance is best if the probe has frontal pose (Pose00) and captured from nearby distance (8 ft.).

Protocol 3. Effect of all the poses and distances together: In real surveillance scenarios, there is usually no restriction on the pose and distance of the subjects from the camera. So the most realistic and also the most challenging protocol

1556-6013 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2868173, IEEE Transactions on Information Forensics and Security

TABLE VI: Rank-1 recognition (%) for the three protocols on the HPR database.

Method	Distances	Pose00	Pose01	Pose02	Pose03	Pose04	Pose05	Pose06	All Poses
DAlign	Dist 1 (8 ft.)	96.33	40.33	84.67	91.33	48.15	42.33	47.33	54.81
	Dist 2 (12 ft.)	96.00	35.00	82.67	87.00	46.33	34.00	46.00	51.90
	Dist 3 (16 ft.)	93.00	31.00	76.00	76.33	31.33	22.33	43.00	44.49
	All Distances	94.33	36.55	79.11	84.33	40.55	34.11	44.67	50.16
randKCCA	Dist 1 (8 ft.)	71.67	17.00	44.33	45.00	16.33	14.00	16.00	39.05
	Dist 2 (12 ft.)	70.33	13.33	42.67	42.00	15.67	13.33	15.00	37.71
	Dist 3 (16 ft.)	63.67	12.00	34.00	29.67	11.00	9.67	15.00	31.95
	All Distances	74.67	18.44	46.00	43.78	17.56	13.00	19.44	39.43
DAlign + randKCCA	Dist 1 (8 ft.)	96.33	41.00	85.67	92.00	49.33	45.67	48.67	60.62
	Dist 2 (12 ft.)	96.33	35.33	83.33	88.00	47.67	35.33	46.33	56.91
	Dist 3 (16 ft.)	93.67	32.67	77.33	77.56	31.67	24.67	43.33	50.52
	All Distances	94.89	37.44	80.67	85.33	41.56	35.11	46.78	56.46

is when the probe consists of all the seven poses and all the three distances together. The results are reported in Table VI at "All Distance" vs. "All Poses". We see that the performance of the dictionary alignment approach for this protocol is 50.16%. The proposed re-ranking approach is able to improve the performance to 56.46%. So we see that a lot of work is still required to achieve satisfactory performance for low-resolution heterogeneous face recognition task.

Effect of SR on real LR faces: We have observed that SR techniques help to improve the performance for the LR NIR images. Inspired by this, we conduct another experiment to study the effect of SR techniques on real LR images. For this experiment, we take the NIR images at 16 ft. distance as LR images and apply SR technique [20] to reconstruct the HR images before performing the matching. We observe that there is almost no improvement in performance, which indicates that improving the matching performance of real LR images using SR techniques is much more challenging compared to that of synthetic LR images.

Effect of real vs. synthesized LR images: We perform another experiment in which we synthesize LR images by downsampling the images at 8 ft. to the size of images at 16 ft. and computing the recognition accuracy with our approach. We then compare its recognition performance with that of the real LR images at 16 ft. The rank-1 performance on the synthesized (real LR) images using DAlign, randKCCA and the re-ranking framework are 51.71% (44.49%), 35.19% (31.95%) and 56.46% (50.52%) respectively. Clearly, the synthesized images have significantly better recognition accuracy compared to the real LR images, which further illustrates the importance of working with real LR images as compared to synthetically generated LR images.

VII. Acknowledgement

This work is partly supported through a research grant from MeitY, India. We would like to thank Prof. K. R. Ramakrishnan (Retd. Professor, IISc) and Mr. Suresh Kirthi K. (IISc) for helpful discussions.

VIII. Conclusion

In this work, we address the challenging problem of matching uncontrolled LR NIR images with HR VIS images in the database, which will be very useful for low-light or night-time surveillance scenarios. We propose a dictionary alignment approach for this task. We also propose a re-ranking approach, which can effectively combine the dictionary alignment approach with another approach to further improve the performance. We further propose our own HPR database to systematically study the effect of head pose and image resolution on the recognition performance. We observe from the experimental results that most of the existing algorithms including the proposed one are significantly affected with non-frontal head pose of the probe image. One way to address this is to utilize approaches which frontalize faces along with the proposed approach for improved performance on non-frontal faces, and this will be part of our future work.

References

[1] S. Bai and X. Bai. Sparse contextual activation for efficient visual re-ranking. *IEEE Transactions on Image Processing*, 25(3):1056–1069, 2016.

[2] S. Bai, X. Bai, Q. Tian, and L. J. Latecki. Regularized diffusion process on bidirectional context for object retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[3] C. Bao, J. F. Cai, and H. Ji. Fast sparsity-based orthogonal dictionary learning for image restoration. *IEEE International Conference on Computer Vision*, pages 3384–3391, 2013.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[5] Y. Chu, T. Anand, G. Bebis, and L. Zhao. Low-resolution face recognition with single sample per person. *Signal Processing*, pages 144–157, 2017.

[6] 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*, 40(5):1139–1153, 2018.

[7] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. *IEEE International Conference on Computer Vision*, pages 2960–2967, 2013.

[8] J. Garcia, N. Martinel, A. Gardel, I. Bravo, G. Foresti, and C. Micheloni. Discriminant context information analysis for post-ranking person re-identification. *IEEE Transactions on Image Processing*, 26(4):1650–1665, 2017.

[9] M. Grgic, K. Delac, and S. Grgic. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011.

[10] H. Han, S. S. A. K. Jain, and, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2017.

1556-6013 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2868173, IEEE Transactions on Information Forensics and Security

[11] R. He, X. Wu, Z. Sun, and T. Tan. Learning invariant deep representation for nir-vis face recognition. *Association for the Advancement of Artificial Intelligence*, 4:2000–2006, 2017.

[12] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[13] G. Hu, X. Peng, Y. Yang, T. Hospedales, and J. Verbeek. Frankenstein: Learning deep face representations using small data. *IEEE Transactions on Image Processing*, 27(1):293–303, 2018.

[14] Y. Jin, J. Lu, and Q. Ruan. Coupled discriminative feature learning for heterogeneous face recognition. *IEEE Transactions on Information, Forensics and Security*, 10(3):640–652, 2015.

[15] D. Kang, H. Han, A. K. Jain, and S. W. Lee. Nighttime face recognition at large standoff: Cross-distance and cross-spectral matching. *Pattern Recognition*, 47(12):3750–3766, 2014.

[16] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–12, 2014.

[17] J. Kittler, R. D. M. Hatef, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[18] B. F. Klare and A. K. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1410 –1422, 2013.

[19] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, 2012.

[20] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2017.

[21] Q. Leng, R. Hu, C. Liang, Y. Wang, and J. Chen. Person re-identification with content and context re-ranking. *Multimedia Tools and Applications*, 74(17):6989–7014, 2015.

[22] J. Lezama, Q. Qiu, and G. Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6807–6816, 2017.

[23] S. Li, D. Yi, Z. Lei, and S. Liao. The casia nir-vis 2.0 face database. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 348–353, 2013.

[24] W. Li, Y. Wu, M. Mukunoki, and M. Minoh. Common near-neighbor analysis for person re-identification. *IEEE Conference on Image Processing*, pages 621–1624, 2012.

[25] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schlkopf. Randomized nonlinear component analysis. *International Conference on Machine Learning*, pages 1359–1367, 2014.

[26] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[27] 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*, 37(10):2041–2056, 2015.

[28] H. Maeng, H. C. Choi, U. Park, S. W. Lee, and A. K. Jain. Nfrad: Near-infrared face recognition at a distance. *IEEE International Joint Conference on Biometrics*, pages 1–7, 2011.

[29] S. P. Mudunuri and S. Biswas. Dictionary alignment for low-resolution and heterogeneous face recognition. *IEEE Winter Conference on Applications of Computer Vision*, pages 1115–1123, 2017.

[30] S. P. Mudunuri, S. Venkataramanan, and S. Biswas. Improved low resolution heterogeneous face recognition using re-ranking. *6th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, pages 446 – 456, 2018.

[31] B. S. Oh, K. Oh, A. B. J. Teoh, Z. Lin, and K. A. Toh. A gabor-based network for heterogeneous face recognition. *Neurocomputing*, pages 253–265, 2017.

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

[33] O. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *British Machine Vision Conference*, pages 1–12, 2015.

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

[35] C. Reale, H. Lee, and H. Kwon. Deep heterogeneous face recognition networks based on cross-modal distillation and an equitable distance metric. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 32–38, 2017.

[36] C. Reale, N. M. Nasrabadi, H. Kwon, and R. Chellappa. Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 54–62, 2016.

[37] H. Roy and D. Bhattacharjee. Local-gravity-face (lg-face) for illumination-invariant and heterogeneous face recognition. *IEEE Transactions on Information, Forensics and Security*, 11(7):1412–1424, 2016.

[38] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. *arXiv preprint arXiv:1711.10378.*, 2017.

[39] S. Saxena and J. Verbeek. Heterogeneous face recognition with cnns. *European Conference of Computer Vision Workshops*, pages 1 – 8, 2016.

[40] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2167, 2012.

[41] M. Singh, S. Nagpal, N. Gupta, S. Ghosh, R. Singh, and M. Vatsa. Cross-spectral cross-resolution video database for face recognition. *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–7, 2016.

[42] a. B. C. T. Bourlai. Multi-spectral face recognition: identification of people in difficult environments. *IEEE International Joint Conference on Intelligence and Security Informatics*, pages 196–201, 2012.

[43] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2001.

[44] S. Wang, D. Zhang, Y. Liang, and Q. Pan. Semi coupled dictionary learning with applications to image super resolution and photo sketch synthesis. *IEEE International Conference on Computer Vision*, pages 2216–2223, 2012.

[45] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *arXiv preprint arXiv:1511.02683*, pages 1–13, 2015.

[46] X. Wu, L. Song, R. he, and T. Tan. Coupled deep learning for heterogeneous face recognition. *arXiv preprint arXiv:1704.02450*, pages 1–9, 2017.

[47] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. leng, X. Xiao, J. Chen, and R. He. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566, 2016.

[48] R. Yu, Z. Zhou, S. Bai, and X. Bai. Divide and fuse: A re-ranking approach for person re-identification. *arXiv preprint arXiv:1708.04169*, 2017.

[49] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas. Query specific fusion for image retrieval. *European Conference of Computer Vision Workshops*, pages 660–673, 2012.

[50] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian. Query-adaptive late fusion for image search and person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1741–1750, 2015.

[51] L. Zheng, Y. Yang, and A. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv: 1610.02984*, 2016.

[52] J. Zhu, W. Zheng, J. Lai, and S. Li. Matching nir face to vis face using transduction. *IEEE Transactions on Information Forensics and Security*, 9(3):501 – 514, 2014.

1556-6013 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.