Visible-Light and Near-Infrared Face Recognition at A Distance

Chun-Ting Huang\textsuperscript{a}, Zhengning Wang\textsuperscript{b}, and C.-C. Jay Kuo\textsuperscript{a}

\textsuperscript{a}Ming Hsieh Department of Electrical Engineering, University of Southern California
Los Angeles, CA 90089 USA
\textsuperscript{b}Department of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, China

Abstract

A method to solve the problem of face recognition at a distance (FRAD) under the visible-light (VIS) and the near-infrared (NIR) spectra is presented in this work. For images taken under visible light at day time, we perform the coarse-scale alignment/enhancement to eliminate a set of unlikely candidates at the first stage. Then, the fine-scale alignment/enhancement steps are conducted to refine the candidate list furthermore iteratively at the second stage. To address the additional challenge associated with NIR images captured at night time, we incorporate a restoration mechanism that reconstructs low-quality patches through a locally linear embedding (LLE) process with a local constraint. It is shown by experimental results that our FRAD solution outperforms state-of-the-art methods on both VIS and NIR images.

Keywords: Face recognition, Cross-distance matching, Cross-environment matching, Two-stage filtering, Image restoration, Locally Linear Embedding (LLE)

1. Introduction

With a rapid growth of security demand, authorities in many countries adopt video surveillance systems to monitor possible threats for public security and enforce traffic management. For instance, London has installed about half-million cameras to observe public spaces, and major cities in the United States such as New York and Los Angeles are extending their surveillance networks largely. As the result, the fast increasing volume of large video data has become harder to
manage. Besides, incidents like Boston Marathon bombing require heavy manual labor to examine low-quality face images within the surveillance camera’s footage. The method to improve efficiency and accuracy of automatic face recognition in the context of video surveillance imposes a major challenge in research community.

Even though there have been progressive developments in automatic face recognition, most state-of-the-art methods focus on faces with variant poses yet at a close distance with sufficient quality. The captured face by a surveillance system can be of low resolution and poor quality, which is degraded by an uncontrolled outdoor environment such as long distance at day or even night time. This leaves a noticeable gap between the clear image within datasets and the real footage obtained by surveillance systems. For example, a great majority of face recognition researchers study faces with an arbitrary pose such as those in the Labeled Faces in the Wild (LFW) dataset [1]. In contrast, little research has been conducted to solve the problem of face recognition at a distance (FRAD). Usually, long distance outdoor face images and short distance indoor face images for matching are called probe and gallery sets, respectively, in this context.

Since most surveillance footages taken at a long distance can be easily distorted by illumination and polarization, the recognition performance is severely affected. Normalization is a critical step in handling distortions caused by the long distance. It can be categorized into two types: geometrical normalization and photometrical normalization. They are called face alignment and face enhancement, respectively, in this work.

Under normal lighting circumstances, face alignment includes finding an initial rough face shape reference, and approaching the ground truth via iterative optimization. Facial landmark localization is a well known iterative technique in finding the coordinates of essential components. Cascaded regression is a regression-based method proposed for landmark localization. It was introduced by Dollar et al. [2] for pose estimation in image sequences. Later, it was applied to face alignment. Cao et al. [3] proposed a regression method with two-stage training, where cascaded regression was extended to the context of an affine transform. Xiong et al. [4] applied cascaded regression with the SIFT feature and examined the derived solution from a gradient descent view. It is called the supervised descent method (SDM). Yan et al. [5] adopted a similar framework with the “learn-to-rank” and the “learn-to-combine” modules placed before and after the main alignment module, respectively. Automatic face alignment techniques and systems have been extensively tested. For instance, Wagner et al. [6] used the spare representation in their alignment algorithm for the Multi-PIE dataset. Geng
and Jiang [7] developed an automatic alignment system based on both holistic and local features and conducted experiments on the AR, GT and ORL datasets. Deng et al. [8] proposed a transform-invariant PCA (TIPCA) method to achieve automatic alignment, and tested its performance on the FERET dataset. Recently, deep learning [9] and 2D/3D alignment [10, 11] offer competitive performance in this field.

Most automatic alignment algorithms mentioned above target at unconstrained faces such as different face poses, expressions or even occlusion. Faces in a long distance outdoor environment are distorted due to blurring, illumination and various weather conditions. Ban et al. [12] proposed a face alignment method to address distortions caused by blurring and low resolution. However, since their dataset was built in a short-distance (1 to 3 meters) indoor environment, their alignment problem is different from that occurs in the long-distance outdoor environment.

As compared with geometrical face normalization, photometric face normalization has received much less attention. A system that incorporates wavelet decomposition, deblurring, denoising and linear stretching was proposed in [13] to recover quality loss due to the long distance. The reported recognition performance ranges from 50% to 70% due to low image resolution and quality. Another face enhancement technique rooted in retinex theory was proposed by Land and McCann [14]. It examined relative lightness (instead of absolute lightness) in a local region to mimic the human visual experience. Furthermore, Land [15] presented a lightness computation method. By following their work, Jobson et al. proposed a single-scale retinex (SSR) method in [16] and extended it to a multi-scale retinex (MSR) method in [17]. Rahman et al. [18] proposed MSR with color restoration (MSRCR) to handle illumination variation.

An ideal surveillance system should operate around the clock, including both day and night time. Currently, cameras equipped with flash lights are used for night time to offer acceptable performance. However, they are not appropriate for long distance or covert surveillance. As a result, we need to consider other options for night time face recognition. Methods like near infrared (NIR), short-wave infrared (SWIR), and thermal infrared have been studied in the literatures. NIR has become popular in recent years for several reasons [19]. First, NIR is not visible to human eyes, and it is desired to capture face expressions without interrupting subjects in acquisition. Second, the environmental factor has less impact to NIR when compared with others. Third, the NIR illuminator can penetrate glasses easily, which provides additional information if the test subject wears glasses. Generally speaking, NIR offers a good choice for night time long distance
face recognition.

The FRAD problem under VIS and NIR spectra is a challenging topic. A two-stage alignment/enhancement filtering (TAEF) system was proposed for VIS FRAD problem in [20]. It contains alignment and enhancement in the coarse-scale stage and facial matching with a refined candidate pool in the fine-scale stage. In this work, we address the FRAD problem under both VIS and NIR spectra as an extended version of [20]. The material on NIR FRAD problem in this work is completely new. To bridge the gap between VIS and NIR, we propose a restoration system that significantly improves the quality of enhanced NIR face images as compared with prior art in [21]. The restoration system is developed using the Locally Linear Embedding (LLE) method [22] that reconstructs image patches learned from two manifolds. It preserves image’s local characteristics by constraining the reconstruction process within a certain region so that the recovered information will be independent of other regions.

The rest of this paper is organized as follows. Cross-distance, cross-environment and cross-spectral datasets are reviewed in Section 2. We present a system to address the cross-distance and cross-environment facial matching problem under the visible light in Section 3. Then, we examine the facial matching problem with the NIR images as the input, and propose a restoration mechanism to relate NIR and VIS images in Section 4. The performance of the proposed solution is evaluated by extensive experimental results including state-of-the-art deep learning approach in Section 5. Error cases are analyzed in Section 6 for additional insights into the proposed solution. Finally, concluding remarks and future works are given in Section 7.

2. FRAD Datasets

There are few publicly available datasets collected for the FRAD problem. The study began with only the cross-distance scenario (VIS-to-VIS) with the UTK-LRHM dataset [13] in 2008. It contains 55 subjects with distances ranging from 10 to 16 meters in an indoor environment and 48 subjects with distances from 50 to 300 meters in an outdoor environment. Another FRAD dataset was built by Rara et al. [23] for sparse-stereo reconstruction in 2009. It has 30 subjects with three distances (namely, 3, 15, and 33 meters). Tome et al. [24] evaluated the effect of distance degradation for several matching methods based on the “Face still dataset” of the NIST multiple evaluation grand challenge (MBGC) [25].

Researchers have started to consider both the cross-distance and the cross-spectral challenges for FRAD datasets since 2011. The near-infrared face recog-
Figure 1: Four exemplary images in the LDHF dataset taken at a distance of 100 and 150 meters. Images in the first row are taken under VIS, whereas the second row represents NIR images from the same subjects.

dition at a distance (NFRAD) dataset was built by Maeng et al. [26]. It has 50 subjects with both VIS and NIR photos in a controlled environment, where the captured distances are 1 meter and 60 meter. This dataset provides some pose change (the frontal view, the slight left and the right face view angles). However, since the NIR illuminator generates a halo-like light pattern around the subject at the 60 meter distance, its usage becomes very limited. The second dataset, collected by Bourlai et al. [27], has 103 subjects captured at 30, 60, 90 and 120 meters in the NIR outdoor environment as the probe set, and 5 feet in the VIS indoor controlled conditions as the gallery set, respectively. This dataset does not cover outdoor VIS images. Finally, the LDHF dataset [28] provides 100 subjects with three standoff distances: 60, 100, 150 meters VIS and NIR outdoor probe images, and 1 meter VIS and NIR gallery images with better quality comparing with other two datasets. The LDHF dataset is the only one that is available in the public domain with both cross-distance and cross-spectral characteristics.

Four exemplary LDHF visible-light images are shown in Figure 1, which are taken at 100 and 150 meters, VIS and NIR, respectively. They have the same image resolution (i.e., 5184 × 3456 pixels) but different face sizes (i.e., 220 × 220 pixels for the 100-meter image and 120 × 120 pixels for the 150-meter image.
The illumination in the LDHF dataset can be roughly categorized into three types: normal, foggy, and back-lighted. The two images in the first row of Figure 1 demonstrate how image quality can be affected by the foggy and the back-lighted environments. Apparently, both face alignment and enhancement techniques are needed before facial matching.

Experimental results have been conducted and reported for the LDHF dataset recently. Maeng et al. [26] applied Gaussian smoothing and histogram equalization as the preprocessing step and extracted the dense scale invariant feature transform (Dense-SIFT) [29] from 32 \times 32 overlapping patches. Then, each patch was divided into 4 \times 4 grids, where an 8-bin gradient orientation histogram was calculated to form a 128-D feature vector. The matching distance between feature vectors of two VIS images was the Euclidean distance while the linear discriminant analysis (LDA) was adopted for the matching between NIR and VIS images. The LDA projection matrices were learned from another dataset called CASIA HFB [30] so that feature vectors of LDHF’s NIR images can be projected through learned matrices. The method in [26] demands accurate alignment to achieve high performance since dense-SIFT can be affected by slight deviation. No alignment details were provided in [26]. Furthermore, the matching score drops severely for the cross-spectral (i.e., NIR-to-VIS) matching case. This demonstrates the shortcoming of the dense-SIFT feature used in recognition without a proper pre-processing step.

In addition, Omri et al. [31] proposed a multispectral fusion method for the LDHF dataset. They decomposed VIS and NIR images with the discrete wavelet transform (DWT), selected the most salient features through SVD or PCA and, then, fused the results together. They attempted to solve the cross-distance matching problem with the fused cross-spectral (VIS and NIR) information.

Kang et al. [21] presented a general framework to achieve cross-distance and cross-spectral matching. They collected random sample patches from the whole face region and constructed a dictionary by relating high- and low-quality patches extracted from the same location using the LLE method. Based on the dictionary, an input low-quality patch can be reconstructed using nearby high-quality neighbor patches with a weighted sum. Finally, results from the Multiscale Local Binary Pattern (MLBP) [32] and the SIFT preprocessing filters were incorporated for score-level fusion. This scheme improves image quality from loss caused by long distance and NIR acquisition. On the other hand, its recognition rate of long distance VIS images is worse than that of [21].

In this work, we adopt different pre-processing techniques to handle VIS and NIR probe images separately. For VIS probe images, the main problem lies in
mis-alignment between the probe and galley images so that we need to pay more attention to geometrical normalization. For the NIR probe images, the main problem lies in degraded image quality. We propose a restoration solution to improve the quality of NIR face images. The details are provided in Section 3.

3. Cross-Distance Facial Matching Under Visible Light

In this section, we introduce a solution to the problem of cross-distance facial matching under the visible light. It is called the two stage alignment/enhancement filtering (TAEF) system whose diagram is shown in Figure 2. At the first stage, approximate facial landmarks are obtained using cascaded regression in the alignment step and MSRCR in the enhancement step. The objectives of this stage are two folds: filtering out unlikely candidates to reduce the size of the candidate pool and providing a better initialization for further processing. At the second stage, fine-scale alignment, enhancement and matching operations are performed iteratively to reduce the least probable candidate one by one until the last one is reached.

3.1. Coarse-scale Processing

3.1.1. Coarse-scale Alignment (C-Alignment)

Typically, the initial location of the face region is determined by a face detection algorithm. It can be affected by the appearance variation such as poses, expressions and occlusion. In this work, we focus on frontal faces with limited
Figure 3: Two exemplary enhanced facial images (from left to right): the original image, the enhanced images by MSRCR, histogram equalization, dark channel prior [34], Laplacian sharpening and wavelet decomposition. MSRCR can strike a balance between dehazing and contrast enhancement.

facial expression as shown in the LDHF dataset. With this simplified condition, we adopt the Viola-Jones [33] algorithm as the face detector. The detector can capture most long distance faces during day-time since they are frontal. Even for the distance of 150 meters, the detector has only 8 false positive cases among 100 faces. These cases can be easily removed by simple background examination. We will provide more details on face detection for NIR images in Section 4.

In the cross-distance matching problem, the test image contains a long distance face, the ground truth is a short-distance face of higher resolution and better quality. Apparently, there is a mismatch between the training and testing data. To address the mismatch problem, we design a distortion filter that mimics the long distance effect so as to generate the facial landmarks of synthetic long distance images. The cascaded regression scheme has to be retrained using the distance-adjusted data. Besides, we need an initial location for each landmark in the cascaded regression. To achieve this goal, we conduct Procrustes analysis on all landmarks of images in the distance-adjusted training set to yield a face model formed by the averaged locations of landmarks. After the face region bounding box on the test image is generated and a face model is constructed based on the distance-adjusted training set, we map these landmarks to their corresponding locations in the test face region to generate initial landmark locations. Then, given an input face image, we apply the cascaded regression to reduce the distance between the estimated landmark position and that of the training data. This process can be written mathematically as follows.

To conduct the cascaded regression, we need initial facial shape and training
set of multiple subjects. The initial facial shape is represented by the coordinates of \( N \) initial facial landmarks in form of \( S^0 = [x_1, y_1, \ldots, x_N, y_N] \). The training set is denoted by \( \{(I_k, S_k)\}_{k=1}^{K} \), where \( I_k \) is the \( k \)th subject, \( S_k \) is the corresponding landmark-based facial representation, and \( K \) is the total number of training subjects. With these two inputs, cascaded regression generates a sequence of approximations \( S^1, \ldots, S^t, \ldots, S^T \), where \( S^T \) is the converged output. The \( t \)-th facial shape is updated based on

\[
S^t = S^{t-1} + R^t(I, S^{t-1}),
\]

where

\[
R^t = \arg \min_R \sum_{k=1}^{K} ||S^t_k - [S^{t-1}_k + R(I_k, S^{t-1}_k)]||
\]

is learned. \( R \) represents a regressor and \( S^{t-1}_k \) is the estimated shape produced in the previous stage. In this work, \( R \) is chosen to be a linear regressor since it can handle the desired task efficiently. To explain the above concept in words, we design a sequence of regressors, where each regressor is trained based on the difference between the estimated result from the previous stage, \( S^{t-1}_k, k = 1, \ldots, K \), and the ground truth, \( \overline{S}_k \). This iteration process stops when the training error converges. In our work, we adopt multi-scale HOG features [35] as the input descriptor for regressor’s training. To be more specific, the large-scale HOG feature is extracted in the beginning stage while only a small-scale area around each estimated landmark is considered in the later stage.

Furthermore, we regulate the solution at the end of each iteration with two constraints. First, we adopt the sketch token feature from [36] as a reference for the face contour since it offers a reliable edge map using the mid-level feature. It is observed that the trained sketch token model offers an exceptional result on long distance faces in the designated region. Second, the estimated landmarks are constrained based on the face shape with the closest distance, so that no landmarks will deviate from the ground truth too much because of image quality degradation.

To sum up, the predicted facial landmarks for the \( k \)-th subject at the end of the \( t \)-th iteration are obtained as the fusion of results from: 1) the predicted result from the \( t \)-th regressor, 2) the output obtained by imposing the sketch-token-based face contour constraint, and 3) the closest landmark model selected from the training set.
Figure 4: Comparison of the 150-to-1 meter face matching result with different enhancement methods: ROC (left) and CMC (right). This comparison is generated based on the intermediate output from coarse-scale stage without fine-scale stage’s processing. MSRCR’s result stands out from other methods, and it also helps to aggregate candidates in the next stage.

3.1.2. Coarse-scale Enhancement (C-Enhancement)

After getting landmarks from the C-alignment procedure, we attempt to re-store the distorted facial color so as to allow robust cross-environment facial matching. We test several enhancement algorithms, including histogram equalization, dark channel prior [34], Laplacian sharpening, wavelet decomposition and MSRCR, and conclude that MSRCR provides a superior performance on foggy and back-lighted images. Two examples are given in Figure 3 for subjective visual comparison. The top and bottom original images in Figure 3 are distorted by the foggy and back-lighted conditions, respectively. The goal of enhancement is to remove these environmental factors to allow cross-environment matching. We see that MSRCR does provide better results against the original ones.

For objective performance evaluation, we compare the receiver operating characteristic (ROC) curve and the cumulative match characteristic (CMC) curve of the matching result under different enhancement algorithms in Figure 4, where the matching result in this figure is generated using only the coarse-scale alignment/enhancement and will be detailed in the next subsection. We see from this figure that only MSRCR can improve the matching performance. It is worthwhile to point out that most image enhancement algorithms have been developed for white noise removal. The white noise model is however not suitable in characterizing outdoor distortions. The matching performance is actually worsened by these enhancement algorithms designed for other purposes. In contrast, MSRCR compensates the environment effect with a more suitable design and, as a result, it can offer better performance over the original one. Furthermore, the possible can-
didates also aggregates forward after MSRCR’s enhancement, which implies the necessity of C-enhancement.

3.1.3. Facial Matching (C-Matching and F-Matching)

The facial matching component appears in both the coarse-scale and the fine-scale processing modules, where the same matching method is adopted and described below.

After proper alignment and enhancement, the first step is to extract feature descriptors (e.g. HOG and SIFT) and geometric features (i.e. facial landmarks) in polar coordinates. Polar coordinates are adopted because it can represent relative locations of aligned landmarks conveniently. Feature descriptors are extracted from two cropped face regions called the interior face region and the bounded face region, respectively. However, due to the constraint set by NIR environment, the third cropping scale remains the same with Kang’s setting [21]. Two examples are illustrated in Figure 5. The interior face region includes major facial components up to eyebrows and down to part of chin without ears and hair. The bounded face region has the whole face including the face contour and partial hair such as bangs. Since the bounded face region is sensitive to background and hairstyle change, it is not used in traditional face recognition systems. However, for the cross-distance and cross-environment face recognition problem, the information contained in the interior face region could be too little. The additional information contained in the bounded face region can be helpful, and it is not proper to discard any relevant information due to aggressive cropping. The discussion related to cropped NIR face region will be explained in Section 4. We will focus on having VIS images as the input for the TAEF system, since it is easier to show the argument.

Experiments (Table 2 in Section 5) show that the performance of using the information from both interior and bounded face regions is better than that from only one of them. Thus, in our implementation, both HOG and SIFT features from the two regions are used separately as individual classifiers for further processing. Moreover, both regions share the same interpupillary distance (IPD) as 40 pixels. The number 40 is chosen because it is the average IPD for 150 meter images, and the cutoff range is decided based on the total average face boundary. The reason to fix the IPD in both regions is to maintain the resolution alignment since we may generate distortions during resizing by allowing images of different scales. After collecting all features from both interior and bounded regions of aligned and enhanced face images, we can measure the Euclidean distance of feature vectors and generate rank-order lists from all classifiers. Then, a weighted voting will be used to pile all classification results into one single rank-ordered list. By gradu
ally eliminating less probable candidates in various stages, the TAEF system will provide the final ranked result.

3.2. Fine-scale Processing

An iterative alignment/enhancement filtering process is adopted in this stage. It means that, after eliminating the least possible candidate from the selection pool through alignment, enhancement and matching, features are extracted from re-normalized images based on remaining images in the pool at the next iteration. This process is described in detail below.

Sets of probe and gallery images are denoted by \( P = \{P_1, \cdots, P_k, \cdots, P_{N_p}\} \) and \( G = \{G_1, \cdots, G_k, \cdots, G_{N_g}\} \), where \( N_p \) and \( N_g \) are their sizes. Furthermore, \( O_1, \cdots, O_k, \cdots, O_{N_p} \) are candidate pools for probe images \( P_1, \cdots, P_k, \cdots, P_{N_p} \), respectively. The iterative filtering process consists of two steps at each iteration. First, probe image \( P_i \) is geometrically and photometrically normalized with subjects left in its candidate pool \( O_i \) so that the normalized probe image can be written as

\[
P_i' = \Gamma(G_j, \Lambda(G_j, P_i)), \quad G_j \in O_i,
\]

where \( \Lambda \) and \( \Gamma \) are the fine-scale alignment (F-alignment) and enhancement (F-enhancement) operations, respectively. Afterwards, the new candidate pool is expressed as

\[
O_i' = \{O_i \mid V_j \geq N_v\},
\]

where

\[
V_j = \sum_{l=1}^{N_c} w_l \cdot C_l(\Psi_l(P_i'), j)
\]

is the vote collected from classifiers \( C_l, l = 1, \cdots, N_c \) using feature transform \( \Psi_l \) and weighting factor \( w_l \), and \( N_v \) is a threshold of vote count for the pool.

The same cascaded regression in C-alignment is applied to the probe image in the F-alignment but with one major difference. That is, it is aligned with each individual gallery image \( G_j \) in the candidate pool \( O_i \) one by one, where the Procrustes analysis is conducted to derive the transform array. Translation, orthogonal rotation, reflection, and scale component are all calculated in this process. Furthermore, only reliable landmarks are selected to reduce the influence from inaccurate landmarks. For example, tips of eyes and the mouth often have higher steadiness and the nose rim location is difficult to determine in long distance. With these improvements, the F-alignment, denoted by \( \Lambda \) in Eq. (3) can reduce the estimation error based on the improvement in the last iteration.
The F-enhancement process, denoted by $\Gamma$ in Eq. (3), is needed for photometric normalization, and it is achieved by region-based histogram matching. In contrast with the traditional histogram matching method that calculates the histogram of the whole face, we match histograms of the probe image and each individual gallery image in face sub-regions segmented by localized landmarks in the F-alignment step so as to differentiate images in a small candidate pool.

4. Cross Spectra (NIR/VIS) Facial Matching

For NIR-to-VIS facial image matching, facial alignment and enhancement as detailed in Section 3 are still needed. Furthermore, we need several special techniques to process NIR images.

4.1. Image Pre-Processing

Due to the low signal-to-noise ratio (SNR) of NIR images, we adopt a tighter cropped facial region for NIR images as shown in Figure 5, where the IPD is set to 92 pixels. This choice not only rules out most background noise but also preserves needed facial texture. In this pre-processing step, all NIR face images are cropped into the same size ($192 \times 240$ pixels), where eye locations are automatically detected and aligned using the technique described in Section 3.

Because of very poor NIR image quality at 100 and 150 meters, the Viola-Jones detector is not as effective as being applied to VIS images. It gives 35 false positive and 5 false negative cases among one hundred 150-meter NIR images.
We explore pupil’s reflection from the NIR illuminator as an auxiliary tool to mark the eye location. With this extra procedure, we are able to detect all faces on NIR 150 meter images without any error. For image enhancement, we would like to maintain fidelity without removing too many details. The $3 \times 3$ median filter is applied to suppress high frequency noise, and a simple image contrast adjustment is adopted to enhance the image. The contrast adjustment is decided based on whole image’s gray-scale histogram distribution. That is, since the NIR image does not receive sufficient light, we adjust the intensity values to meet the condition that 1% of data of the whole image is saturated at low and high intensities.

To partially recover the lost information caused by the long distance and the spectral difference partially, we need to bridge the gap between 1 meter and 150 meter NIR images by building their correspondence. Being inspired by the work in [37], we adopt the LLE method to achieve this goal, which is explained in detail in the next subsection.

4.2. LLE-based Image Restoration

The proposed restoration scheme is based on the framework in [21] but with additional features to boost up the performance. It consists of two stages: 1) the correspondence building stage and 2) the correspondence finding stage. In the first stage, we build the correspondence between high quality and low quality patches. In the second stage, we use the correspondence to reconstruct patches to restore the quality of the probe image.

The operation in the first stage is illustrated in Figure 6. We first partition 150-meter (low quality) and 1-meter (high quality) NIR images in the gallery set into multiple subregions, and extract a large number of patches from each subregion. The pixels of low quality and high quality patches are cascaded into a single vector. Then, we use the tree-structured vector quantization (TSVQ) to cluster vectors separately based on their source subregion. TSVQ clusters vectors into two groups at each level, and repeat the same procedure at each group recursively until the desired cluster number is reached. The centroid of each cluster is called a codeword, and the set of all codewords generated by the TSVQ is called a codebook (or a dictionary). Thus, we can associate a codebook with each subregion.

Mathematically, we use $\{(G^H_k, G^L_k)\}_{k=1}^{K}$ to denote the 1 meter and 150 meter NIR images from the same subject in the gallery set $G$, where $K$ is the total number of subjects and superscripts $L$ and $H$ indicate low and high quality images, respectively. We divide each image into smaller subregions denoted by $\{D_t, t = 1, 2, ..., T\}$, and extract pairs of corresponding patches from the same
Figure 6: The construction of two corresponding codebooks using 150-meter and 1-meter NIR patches, where low quality (150-meter NIR) and high quality (1-meter NIR) patches are extracted from the same subregion of the same subject and cascaded into a vector. TSVQ is used to generate a codebook for each subregion.

Figure 7: The process of restoring low quality patches using the LLE-based method. Patches are extracted from the designated subregion. Then, the system inquires the corresponding codebook, locates its neighbor patches, and reconstructs the target via LLE. Finally, the restored face image is resembled by all reconstructed patches.

From manifold learning, the correspondence between two manifolds can be learned when they possess similar local geometries. Here, we use LLE to learn the relationship between high quality (1 meter) and low quality (150 meter) patches in each cluster. Based on the learned relationship, we can reconstruct the 150 meter probe patches in each subregion and, then, restore the whole image accordingly.

Figure 7 demonstrates the restoration procedure. Once a probe image’s patch
is extracted, we can use it to locate the closest cluster, \( \{C_o^L, o = 1, 2, ..., O\} \), in each subregion through the minimum Euclidean distance, where \( O \) indicates the total number of clusters in each subregion. For a low quality patch denoted by \( \pi_j^L \), we select its \( S \) nearest neighbor patches of the same cluster and use them to calculate weights \( \{w_s, s = 1, 2, ..., S\} \) that minimize the following error

\[
\varepsilon_j = ||\pi_j^L - \sum_{\pi_s^L \in N_j^L} w_s \pi_s^L||,
\]

where \( N_j^L \) denotes the nearest neighbor low quality patches from \( C_o^L \), and \( \|\cdot\| \) is the Euclidean norm.

With weights \( w_s \) obtained from the above equation, we can use them to reconstruct the corresponding high quality patch

\[
\pi_j^H = \sum_{\pi_s^H \in N_j^H} w_s \pi_s^H,
\]

where \( N_j^H \) is the corresponding nearest neighbor high quality patches from cluster \( C_o^H \). We should emphasize that \( C_o^H \) contains high quality patches in the same geometric location of \( C_o^L \), except that they are extracted from 1 meter NIR image of the same subject. Furthermore, we take the regional background into consideration by averaging the restored image with the cluster mean image \( C_o^H \) in the subregion (e.g., one half from the restored image and the other half from the cluster mean image). Therefore, the restored result will be less sensitive if the nearest neighbor patches fail to represent the input probe patch. Finally, we reassemble patches back to the subregion to restore the whole facial image. We use overlapping subregions to reduce the blocking effect.

The major difference between VIS and NIR images is image quality. Furthermore, since there is no large-scale NIR face dataset with labeled landmark annotation, the performance of NIR facial landmark localization is limited. Here, we replace the C-alignment and the C-enhancement steps with “integration from face detection and eye location marking” and “LLE-based image restoration”, respectively. The TAEF system will output the final decision for NIR images right after C-matching. There is no processing needed in the fine-scale stage.

5. Experimental Results

5.1. Implementation Details

The LDHF dataset has a limited number of gallery images. Since training landmark localization requires a relatively large dataset with annotation, we in-
clude the MUCT dataset [38]. It consists of 3755 images with 76 landmarks collected from 276 subjects. We only select 25 out of 76 landmarks in the C-alignment training process by focusing on visible landmarks such as the center and the edge tips of eyes and the mouth in the long distance. In addition, we include LDHF’s gallery images along with those in MUCT in the training set for C-alignment. For LDHF’s gallery images, each face is manually labeled with 25 landmarks since the ground truth is not provided by the dataset. We simulate three long distance scenarios for each gallery image so that the number of training images is tripled. The three scenarios are: 1) contrast change due to the long distance, 2) fog and 3) back-lighted. This allows the regressor to learn the cross-environment setting.

In the C-enhancement step, we apply MSRCR to the whole probe image, where the model parameters are determined by the histogram of each probe image focusing on the facial area. If the histogram of an image contains a concentrated peak in the dark area without high intensity values, it belongs to the back-lighted scenario. If the histogram spans over a broad area with intermediate intensity values, it is under the foggy condition. Otherwise, it is the normal situation. To compare enhancement methods, we set parameters $\alpha = 0.7$ and $c = 0.5$ in the Laplacian sharpening method and parameters $\text{threshold} = 50$ and $C = 2$ in the wavelet decomposition method. Moreover, it is evaluated using the coarse-scale output only. That is, we list the output from C-matching in Figure 4, where the verification rates under FAR=0.1% are 12% for the dark channel prior, 18% for the histogram equalization, 22% for the wavelet decomposition, 32% for the Laplacian sharpening, 44% for the original and 58% for MSRCR.

In the fine-scale stage, TAEF collects votes from all classifiers to build up a candidate pool. We observe that HOG and SIFT features have the ability to select candidates of high similarity but with low first-rank accuracy. They can be used as the main features for both interior and bounded face regions, yet they need to be assisted with geometric features offered by landmarks. As a result, we have six classifiers based on the following feature sets: HOG and SIFT from interior and bounded face regions, landmark’s angle and radius distributions (represented in polar coordinates). The voting mechanism collects votes from all six classifiers, and the top $N$ candidates that receive most votes are placed in the initial candidate pool ($N = 5$ in the experiment). Then, one candidate is removed at each iteration until the final one is reached.

For the restoration mechanism, we select $48 \times 64$ as the subregion size and $8 \times 8$ as the patch size. They are determined by considering the trade off between restoration performance and system efficiency. A larger subregion offers higher
efficiency but poor restoration capability, and vice versa. To generate more samples, we allow overlapping subregions and patches with a quarter of their boundary size. This also allows smoother transition across the boundaries of subregions and patches. As a result, there are 153 subregions per image and 609 patches per subregion. Since we adopt 10-fold cross-validation in the dictionary building stage, there are 54,810 patches per grid. We apply the 8-level TSVQ with 256 clusters per subregion. There are 214 patches per cluster on the average. In the patch reconstruction phase, we identify the associated cluster for a long-distance probe patch, choose its five nearest neighbors from the same cluster, and calculate their weights to approximate the probe one. Then, we use these weights and their corresponding 1 meter patches to reconstruct the 1 meter patch.

5.2. Performance for Visible Light Images

We compare the ROC and the CMC curves in Figure 8 to demonstrate the performance of the TAEF system. Note that we need a distance table among all candidates to draw ROC and CMC, where the distance table is built based on the received number of votes. A higher vote number means a closer distance and vice versa, and the distance is weighed by the iteration number. We can see the performance improvement in the ROC curve as the iteration number increases. For example, when $FAR = 0.001$, TAEF gives a verification rate of 12%, 20%, 45%, 48% and 97% at the 1st, 2nd, 3rd, 4th and 5th. The superior performance of the TAEF method is also demonstrated by the CMC plot. At the first iteration, the first rank recognition rate is 51% and it rapidly climbs up to 99% in rank 5.

Figure 8: Performance of the 150-to-1 meter VIS-to-VIS face matching with different iterations: ROC (left) and CMC (right). The performance improves as the iteration number of the fine-scale stage increases. This demonstrates the effectiveness of the image-pair normalization.
follows the same aggregation pattern for later iterations. Its first rank recognition rates are 68%, 81%, 81% and 97% for iteration numbers 2, 3, 4 and 5, respectively.

For performance benchmarking, we compare the results of the TAEF system with those of Maeng et al. [26] and Kang et al. [21]. Note that the former did not provide sufficient details on their alignment process while the latter relied on a commercial software called FaceVACS and manually provided eye locations when the software failed to detect. For these reasons, we can only take the reported data from their papers for the comparison purpose. We list the ROC verification rates of three methods (TAEF and theirs) for 150-meter visible-light images in LDHF in Table 1. TAEF has the best performance among the three. Moreover, we test the TAEF method using 60 meter and 100 meter visible-light images in LDHF, and it gives 100% first rank recognition rate. Thus, TAEF offers the state-of-the-art performance for the FRAD problem in an outdoor setting.

We also compare the first rank recognition rate using features from only the interior or bounded face region or both under the same settings. The results for the 150-meter visible-light images at the first rank are shown in Table 2. It is interesting to see that the performance of the bounded face region alone is better than that of the interior face region. Since the interior face region is blurred due to the long distance effect, its extracted features have limited discriminant power. The additional information from the bounded face region such as the face contour and hairstyle can play an important role although it is less robust. The TAEF system takes both into account and achieves the best performance.

Table 1: Comparison of ROC verification rates for 150-to-1 meter VIS-to-VIS face recognition.

<table>
<thead>
<tr>
<th>Methods</th>
<th>0.1% FAR</th>
<th>1% FAR</th>
<th>10% FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maeng [26]</td>
<td>93%</td>
<td>99%</td>
<td>100%</td>
</tr>
<tr>
<td>Kang [21]</td>
<td>75%</td>
<td>87%</td>
<td>99%</td>
</tr>
<tr>
<td>TAEF</td>
<td>97%</td>
<td>99%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 2: Comparison of the first rank recognition rates for different face regions (VIS-to-VIS).

<table>
<thead>
<tr>
<th>Face region</th>
<th>Interior</th>
<th>Bounded</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st iteration</td>
<td>35%</td>
<td>46%</td>
<td>51%</td>
</tr>
<tr>
<td>2nd iteration</td>
<td>49%</td>
<td>54%</td>
<td>68%</td>
</tr>
<tr>
<td>3rd iteration</td>
<td>60%</td>
<td>69%</td>
<td>81%</td>
</tr>
<tr>
<td>4th iteration</td>
<td>60%</td>
<td>72%</td>
<td>81%</td>
</tr>
<tr>
<td>5th iteration</td>
<td>66%</td>
<td>86%</td>
<td>97%</td>
</tr>
</tbody>
</table>
Figure 9: Performance of the 150-to-1 meter VIS-to-VIS face matching result with different iterations: ROC (left) and CMC (right). The gallery set is enlarged to contain 191 images by including face images from the MUCT dataset.

Figure 10: The restoration result of 150 meter NIR images.

To further verify the effectiveness of the TAEF system, we select 91 faces from the MUCT dataset and combine them the original LDHF 1 meter indoor gallery set. As a result, we have 191 images for matching which is about twice of the original size. We are able to maintain high ROC and CMC accuracy as shown in Figure 9. The first rank accuracy rate reaches 94%, which demonstrates the strength of the two-stage strategy.

5.3. Performance for Near Infrared Images

Five pairs of pre-processed (or intermediate) and restored (or final) face images are shown in Figure 10, where intermediate results after the pre-processing step in Sec. 4.1 are shown in the first row and the final output images using
the LLE-based restoration method are presented in the second row. Clearly, the restored ones give better visual quality than the pre-processed ones. They have less noise and bear higher similarity with the corresponding 1 meter NIR images. Moreover, we compare the visual appearances of the subregion and global-region restoration schemes in Figure 11. Note that there is no subregion decomposition in the global-region restoration scheme. All collected patches are restored by LLE in one single system. Its result is vulnerable to local variants. In contrast, the sub-region restoration scheme gives significantly better results because of its ability to preserve local characteristics within each subregion.

Our restoration results are compared with those obtained by Kang et al. [21] in Figure 12. We see from the figure that the local facial textures are better preserved by our method. For instance, the top subject in Figure 12 has sharper and richer eyebrow shape as compared to the benchmark one. The bottom subject possesses distinct mouth characteristics such as the lip contour and corner.

With restoration, we are able to boost up the first rank accuracy rate on the 150 meter NIR images from 8% to 52% with the HOG feature, and 62% to 76% with the SIFT feature as illustrated in Figure 13. This performance gain demonstrates the ability of recovering some lost information by restoration. We further apply the C-matching step of TAEF to obtain weighted votes from HOG and SIFT extracted from the cropped NIR face region and the interior face region. The proposed TAEF with restoration can achieve 45% verification rate at 0.1% FAR for
5.4. Deep Learning

Deep learning is a popular tool in computer vision applications nowadays. However, due to the limited size of the LDHF dataset, it is difficult to train an effective deep network in our current context. We should emphasize that there is only one image for each subject under the same distance in this dataset, and the correspondence between images of each subject at various distances is fixed. The same data augmentation technique (e.g., rotation, mirroring or random sampling) has to be applied to both the input and output image pairs simultaneously. As a result, the technique does not offer more discriminant power among subjects.

On the other hand, several image processing (denoising, inpainting, and super-resolution) problems have been solved by the deep learning technology. In this subsection, we examine the application of deep learning to NIR image restoration.

The super-resolution convolutional neural network (SRCNN) was proposed in [39] to learn the mapping between low- and high-resolution images. Here, we use it as another benchmark for image restoration. We adopt the same network architecture in [39] except changing the first layer’s filter size from $9 \times 9$ to $7 \times 7$ to fit the input image size of our problem better. The restoration relation between

![150m Input | 1m Reference | Benchmark (Kang’s) | Restored Output](image)

Figure 12: Comparison of restoration results obtained by our method (the last column) and by the method proposed by Kang et al. [21] (the 3rd column), where the first column shows the input 150 meter images and the second column shows the 1 meter reference images.
low- and high-quality images is learned from training samples. Besides local restored LLE and SRCNN, we conduct experiments by cascading LLE and SRCNN (namely, applying SRCNN to the restored LLE’s output). We compare the restored images obtained by the three methods in the left side of Figure 14. We see that images restored by the SRCNN are smoother and sometimes blurred with sufficient details. We also compare the CMC curves of the three methods in Figure 14. As shown in the figure, the CMC performance of the SRCNN is worst among the three. The local restored LLE method and the cascaded method have comparable performance. The cascaded method provides slightly better performance because the SRCNN method can improve the image quality based on restored images.

Finally, we compare the ROC verification rates of several methods in Table 3. Although the SRCNN method outperforms the local restored LLE method by 4% at the 0.1% FAR, it does not perform well on 1% FAR and 10% FAR. Overall, the cascaded method provides the best performance among all benchmarking methods in Figure 14 and Table 3.

6. Error Analysis

6.1. Visible Light Images

We examine the error cases for VIS images in this subsection. Among the 100 probe images located at the 150-meter distance, there are three failure cases for TAEF as shown in Figure 15. We provide two intermediate results of probe images
Figure 14: Left: Exemplary restored images obtained by the local restored LLE method, the SRCNN method, and the local restored LLE followed by the SRCNN. Right: The CMC curves of the three methods.

in the first two columns: the output after C-alignment in the first column and the final normalized result in the second column. Furthermore, their ground truth of the 1-meter gallery image is shown in the fourth column while their predicted match by TAEF is shown in the third column. The ground truth images of these three subjects from top to bottom rank as No. 2, No. 2, and No. 4, respectively. In other words, TAEF can still offer the correct prediction if we allow the top two or four ranked candidates in these cases.

One reason for the error is attributed to the difference of the hair style between the gallery and the probe images. For example, for the case in the bottom row, the hair style of the 1-meter gallery ground truth is completely changed in her corresponding 150-meter probe image. Generally speaking, the hair style and the chin shape visible in the bounded face region do contribute positively to the recognition performance. This case happens to work against this policy.

Another reason is due to environmental factors such as blurring, which is not yet considered in the TAEF system. For the first two rows, the interior face regions of the final output images from the TAEF system are still blurred. The loss of details in pupils and eye’s shape can mislead HOG-based and SIFT-based classifiers. With these two blurred probe images, the TAEF system fails to choose the correct one in the last round.
Table 3: Comparison of ROC verification rates for 150-to-1 meter NIR-to-VIS face recognition.

<table>
<thead>
<tr>
<th>Methods</th>
<th>0.1% FAR</th>
<th>1% FAR</th>
<th>10% FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maeng [26]</td>
<td>6%</td>
<td>20%</td>
<td>56%</td>
</tr>
<tr>
<td>DWT-SVD [31]</td>
<td>29%</td>
<td>29%</td>
<td>64%</td>
</tr>
<tr>
<td>DWT-PCA [31]</td>
<td>36%</td>
<td>39%</td>
<td>68%</td>
</tr>
<tr>
<td>Kang [21]</td>
<td>37%</td>
<td>62%</td>
<td>92%</td>
</tr>
<tr>
<td>Local Restored</td>
<td>45%</td>
<td>72%</td>
<td>95%</td>
</tr>
<tr>
<td>SRCNN [39]</td>
<td>49%</td>
<td>65%</td>
<td>93%</td>
</tr>
<tr>
<td>Local Restored + SRCNN</td>
<td>49%</td>
<td>75%</td>
<td>97%</td>
</tr>
</tbody>
</table>

6.2. Near Infrared Images

We examine the error cases for NIR images in this subsection. We show several of them in Figure 16. The pre-processed image, the restored output, the predicted match and the ground truth are displayed in order along each row. The hair style difference between the gallery and the probe images plays an important factor. For instance, the subject in the first row has a smaller fringe in the probe image which is similar to the gallery one. Furthermore, blurred NIR images may lead to different gradients/contours that may confuse the classifier. As compared with the VIS case, NIR image quality degradation is much more significant and serious.

7. Conclusion and Future Work

We first examined and verified the performance of TAEF on LDHF dataset with cross-distance and cross-spectral conditions. It consists of two stages: coarse-scale processing and fine-scale processing. In the coarse-scale stage, a quick scan-through procedure of face alignment, enhancement and matching is executed on all gallery images. After that, a candidate pool is established. In the fine-scale stage, the procedure is further conducted into more detailed alignment and enhancement by pair-wise reference and unlikely candidates are removed one by one. The efficiency of the system is preserved by the two-stage processing procedure.

Next, we presented a restoration system for cross-spectral face matching problem. The proposed restoration system can restore the NIR cross-distance probe image via learning the modality gap between VIS and NIR face images. In addition, given 1 meter and 150 meter corresponding NIR images, each subregion has its own LLE model to recover high-quality patches, which later become parts of
Figure 15: Error cases for 150 meter visible-light probe images (from left to right): after C-alignment, final result after fine-scale processing, the final matching output, and the ground truth.

Figure 16: Error cases of 150 meter NIR probe images (from left to right): the original input after preprocessing, the restored output, the final matching result, and the ground truth.
the restored image. The effectiveness of the solution is demonstrated by higher accuracy rates.

The main issue in the FRAD problem is the lacking of a large long distance face dataset. It is critical to build such a dataset for further research advancement along this direction. Also, the convolutional neural network (CNN) has been tested in short-distance facial recognition problems and reported to have an impressive performance gain. It will be interesting to try the CNN solution if a large labeled long distance facial image dataset is available.

**Acknowledgment**

Computation for the work described in this paper was supported by the University of Southern California’s Center for High-Performance Computing (hpc.usc.edu).


[38] S. Milborrow, J. Morkel, F. Nicolls, The muct landmarked face database, Pattern Recognition Association of South Africa.