NIR-VIS Heterogeneous Face Recognition via Cross-Spectral Joint Dictionary Learning and Reconstruction

Felix Juefei-Xu, Dipan K. Pal, and Marios Savvides
CyLab Biometrics Center, Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA 15213, USA
felixu@cmu.edu, dipanp@andrew.cmu.edu, msavvid@ri.cmu.edu

Abstract

A lot of real-world data is spread across multiple domains. Handling such data has been a challenging task. Heterogeneous face biometrics has begun to receive attention in recent years. In real-world scenarios, many surveillance cameras capture data in the NIR (near infrared) spectrum. However, most datasets accessible to law enforcement have been collected in the VIS (visible light) domain. Thus, there exists a need to match NIR to VIS face images. In this paper, we approach the problem by developing a method to reconstruct VIS images in the NIR domain and vice-versa. This approach is more applicable to real-world scenarios since it does not involve having to project millions of VIS database images into learned common subspace for subsequent matching. We present a cross-spectral joint ℓ_0 minimization based dictionary learning approach to learn a mapping function between the two domains. One can then use the function to reconstruct facial images between the domains. Our method is open set and can reconstruct any face not present in the training data. We present results on the CASIA NIR-VIS v2.0 database and report state-of-the-art results.

1. Introduction

Multi-modal biometric recognition has been a difficult problem to deal with. Vision based biometrics in particular faces this challenge, and has not received much attention from the main-stream computer vision community yet. Nonetheless, many large-scale real-world applications, such as surveillance, actually have to deal with multi-modal data. These applications have to deal with handling images in near infrared (NIR). However, most datasets accessible to law enforcement contain visible light (VIS) images. Cross-spectral heterogeneous face recognition aims at matching face images taken by sensors operating at different wavelengths. Visible light has wavelength $0.38 - 0.7$ μm, near infrared $0.75 - 1.4$ μm, short-wavelength infrared (SWIR) $1.4 - 3$ μm, mid-wavelength infrared (MWIR) $3 - 8$ μm, long-wavelength infrared (LWIR) $8 - 15$ μm, and far infrared (FIR) $15 - 1000$ μm. In this paper, we focus on NIR to VIS matching.

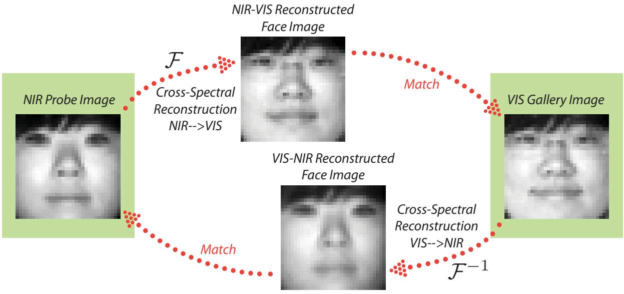
NIR-VIS Reconstructed Face Image
$\mathcal{F}$
NIR Probe Image
Cross-Spectral Reconstruction NIR-->VIS
Match
VIS Gallery Image
VIS-NIR Reconstructed Face Image
Cross-Spectral Reconstruction VIS-->NIR
Match
$\mathcal{F}^{-1}$

Figure 1. Outline of our reconstruction approach to NIR-VIS matching capability. Once learned, the mapping $\mathcal{F}$ can be used to convert a NIR probe image into a VIS image and be matched against a pre-existing VIS gallery. The complimentary process is also possible where one would go from VIS to NIR using $\mathcal{F}^{-1}$. Note that $\mathcal{F}$ and $\mathcal{F}^{-1}$ are not the actual inverse of each other.

In fact, most studies in the vision community have focused on VIS images. These include the development of face recognition systems (FRS). This fact motivates our approach, which is to provide a way for reconstructing a VIS image in the NIR domain and vice-versa. This allows agencies which have already deployed large-scale face matching systems to add NIR-VIS inter-conversion capability as a tool to their arsenal.

In contrast, many studies on handling matching NIR to VIS images either try and build a FRS capable of handling images in both spectra simultaneously, or project the images from both domains onto a common subspace [6]. It, thus, requires a completely separate FRS from the primary FRS that is used by an agency, raising questions on the comparative performance of each system and requiring maintenance of both systems in parallel. Another approach is to project the images onto a common learned subspace. This

978-1-4673-6759-2/15/$31.00 ©2015 IEEE

suffers from being practically expensive to implement. Current agencies have extensive and large databases of VIS images. Mapping all NIR and VIS images onto that subspace is computationally expensive and requires a large amount of pre-processing. Further, both these approaches would not gain from the rapid advancement in exclusively VIS domain matching capabilities. Our approach of reconstructing all NIR to VIS (or the other way if necessary) allows one to use any VIS FRS available since a single image reconstruction is a quick pre-processing step. The outline of our approach is highlighted in Figure 1.

Our algorithm requires NIR and VIS images of the same subject, from which it jointly learns a NIR and a VIS dictionary while constraining the sparse representation of the NIR and VIS images in each dictionary to be the same. This results in two mappings which are optimized to be near (but practically approximate) inverses of one another.

2. Related Work

Unlike traditional single-spectrum face analysis and recognition [19, 18, 28, 12, 13], cross-spectral face recognition requires some additional efforts to bring the two domains to the same platform such that some types of evaluation can make sense. We review some previous work related to our central problem. Zhu *et al.* [31] try to compliment the infeasibility of some classifier learning methods that rely on the corresponding NIR-VIS image pairs of the same target subject. In order to reduce the heterogeneities between VIS and NIR images, the authors propose a transductive subspace model called transductive heterogeneous face matching (THFM) for extracting invariant features for VIS-NIR matching. There are four steps in THFM: (1) domain invariant feature extraction step creates a intra-class scatter-like matrix, (2) target related discriminant model learning step finds a inter-class scatter-like matrix that captures between-class variation in gallery set, (3) cross domain penalization, and (4) locality preserving. The final subspace is obtained by solving a generalized Rayleigh quotient that involves both intra-class and inter-class scatter matrices and their corresponding penalization and locality matrices. An earlier work of theirs can be traced back to [30].

Dhamecha *et al.* [4] study how the histogram of oriented gradients (HOG) feature and its variants can help cross-spectral face recognition tasks. In their experiments, three HOG variants, namely, dense scale invariant feature transform (DSIFT), Dalal-Triggs HOG (HOG-DT), and HOG-UoCTTI are compared with the traditional HOG. They have shown that DSIFT feature together with LDA subspace can outperform a commercial matcher as well as other HOG variants by a large margin.

Hou *et al.* [6] capitalize on external face images collected from both the source and target domains for deriving a common subspace for relating and representing cross-domain image data through a novel domain-independent component analysis (DiCA). It is worth mentioning that during the subspace modeling stage, no label information is need for associating different domains which demonstrates practicality for real-world cross-domain classification problems.

Through periocular information, Jillela and Ross [7] has managed to match face images against iris images. This is not only a cross-spectral matching problem since iris images are taken in NIR and face images are taken in VIS, but also a cross-modality matching problem. In their approach, iris images are matched using a commercial matcher, and face images are matched using local binary patterns (LBP), normalized gradient correlation (NGC) and a sparse representation-based matching scheme where they learn a joint dictionary for both iris images and face images by enforcing the same subjects sharing the same sparse coefficients during training, thus, making cross-modality matching possible.

Li *et al.* [23] incorporate various features and a multi-view smooth discriminant analysis to learn a common discriminative feature space for matching NIR-VIS face images. Similarly, Lei and Li [22] model the properties of different types of data separately and then learn two associated projections to project NIR and VIS data respectively into a discriminative common subspace through a learning framework named coupled spectral regression (CBR). Klair and Jain [21] use random subspace projections as well as sparse representation classification for matching NIR-VIS face images. Goswami *et al.* [5] utilize local binary pattern histogram representation in tandem with LDA for cross-spectral matching. Liu *et al.* [25] focus on finding light source invariant features (LSIFs) in order to extract invariant parts between NIR and VIS images. The method is based on a group of differential-based band-pass filters.

More work on matching VIS face images to SWIR [26, 32, 20] and even to MWIR [2, 3] face images can be found accordingly. Our approach is based on dictionary learning which we present in more detail next.

3. Algorithmic Approach

When handling cross-spectral or even cross-modal data, one critical assumption that can be used is the fact that there is some concept common between the sample points. Let $\mathbf{y}_V$ be the VIS image of the subject and $\mathbf{y}_N$ be the corresponding NIR image. Here the "identity" of the sample points is the same. Hence, one overall approach would be to find a space where both points $\mathbf{y}_V$ and $\mathbf{y}_N$ would map very close by. Once an invertible (approximate) map from both domains of images has been found to a point in the common representation space, one can then use the map to reconstruct any NIR image in the VIS domain and vice versa. As mentioned earlier, our overall approach is to reconstruct an image in the given domain/spectrum to a domain/spectrum which a standard FRS can handle. We now present a dictionary learning based method for cross-spectral reconstruction.

3.1. ℓ_0-Dictionary Based Approach for Cross-Spectral Reconstruction

Linear dictionary learning methods have proved themselves to be an useful approach in modeling problems such as patch-based reconstructions. K-SVD is a recent ℓ_0 dictionary learning algorithm that is a natural extension of K-means [1]. The cluster centers are the elements of the learned dictionary and the memberships are defined by the sparse approximations of the signals in that dictionary. Formally, it provides a solution to the problem

$$\underset{\mathbf{D},\mathbf{X}}{\text{minimize}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \text{ subject to } \forall i, \|\mathbf{x}_i\|_0 < K$$

where $\mathbf{Y}$, $\mathbf{D}$ and $\mathbf{X}$ are the data, the learned dictionary and the sparse approximation matrix respectively. Here $\|\cdot\|_0$ is the pseudo-norm measuring sparsity. The sparse approximations of the data elements are allowed to have some maximum sparsity $\|\mathbf{x}_i\|_0 \leq K$. In this paper, we explore the ℓ_0 method since the explicit control over sparsity allows for better model selection.

Let $\mathbf{y}_{Vi}^j$ be the j_{th} image in the VIS domain of the i_{th} subject with $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, n\}$. Let $\mathbf{y}_{Ni}^j$ be the corresponding image in the NIR domain. We also have matrices $\mathbf{Y}_V$ and $\mathbf{Y}_N$ consisting of the concatenated images in the VIS and NIR spectra respectively. One approach to the problem of cross-spectral reconstruction would be to learn two separate dictionaries $\mathbf{D}_V$ and $\mathbf{D}_N$ in the VIS and NIR domains independently using a dictionary learning algorithm such as K-SVD. We could obtain $\mathbf{D}_V$ and $\mathbf{D}_N$ by solving:

$$\mathbf{D}_V = \underset{\mathbf{D},\mathbf{X}}{\arg\min} \|\mathbf{Y}_V - \mathbf{D}\mathbf{X}\|_F^2 \text{ subject to } \forall i, \|\mathbf{x}_i\|_0 < K$$

$$\mathbf{D}_N = \underset{\mathbf{D},\mathbf{X}}{\arg\min} \|\mathbf{Y}_N - \mathbf{D}\mathbf{X}\|_F^2 \text{ subject to } \forall i, \|\mathbf{x}_i\|_0 < K$$

Then, given a NIR image $\mathbf{y}_N$, in order to reconstruct it in the VIS domain, we would then obtain the sparse approximation in $\mathbf{D}_N$, *i.e.* $\mathbf{x}_N = \arg\min_{\mathbf{x}} \|\mathbf{y}_N - \mathbf{D}_N\mathbf{x}\|_F^2$ subject to $\forall i, \|\mathbf{x}_i\|_0 < K$ and then obtain the VIS reconstruction as $\mathbf{y}_V = \mathbf{D}_V\mathbf{x}_N$. To obtain a NIR reconstruction of a VIS image, one would apply a similar approach starting from the VIS dictionary $\mathbf{D}_V$.

3.2. Cross-Spectral Joint Dictionary Learning

In the previous subsection, we presented a method to perform cross-spectral reconstruction. However, the method suffers from a fundamental oversight. The reconstruction step, $\mathbf{y}_V = \mathbf{D}_V\mathbf{x}_N$, assumes that the images $\mathbf{y}_V$ and $\mathbf{y}_N$ have the same sparse representation in the dictionaries $\mathbf{D}_V$ and $\mathbf{D}_N$ respectively. There is no reason that the sparse representation of the two images (in the different domains) is shared between the two separate dictionaries, since they were trained independent of each other. This problem can be handled by implementing a joint framework for learning the dictionaries. During training, we would like to constrain the sparse representation for each pair of NIR and VIS images to be the same. Thus, the joint optimization problem becomes

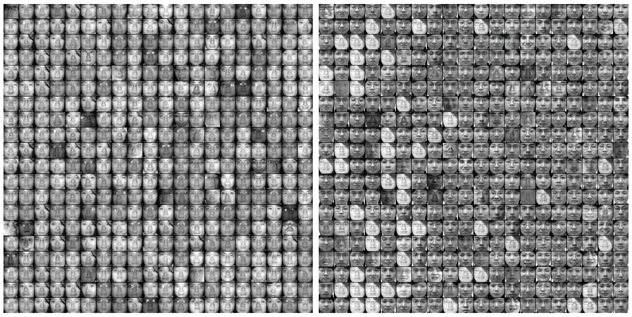

Figure 2. Jointly learned dictionaries from NIR (left) and VIS (right) training samples in fold 1 of the View 2 partition of [24].

$$\underset{\mathbf{D},\mathbf{X}}{\text{minimize}} \|\mathbf{Y}_V - \mathbf{D}_V\mathbf{X}\|_F^2 + \|\mathbf{Y}_N - \mathbf{D}_N\mathbf{X}\|_F^2 \quad (1)$$
$$\text{subject to } \forall i, \|\mathbf{x}_i\|_0 < K$$

Notice that the sparse representation matrix $\mathbf{X}$ is shared between the two terms. Upon some rearrangement we arrive at the cross-spectral joint dictionary learning method.

$$\underset{\mathbf{D}_N,\mathbf{D}_V,\mathbf{X}}{\arg\min} \left\| \begin{pmatrix} \mathbf{Y}_V \\ \mathbf{Y}_N \end{pmatrix} - \begin{pmatrix} \mathbf{D}_V \\ \mathbf{D}_N \end{pmatrix} \mathbf{X} \right\|_F^2 \quad (2)$$
$$\text{subject to } \forall i, \|\mathbf{x}_i\|_0 \leq K$$

This translates to the standard K-SVD problem where we $\text{minimize}_{\mathbf{D}',\mathbf{X}'} \|\mathbf{Y}' - \mathbf{D}'\mathbf{X}\|_2$ under $\|\mathbf{x}_i\|_0 \leq K$. with $\mathbf{Y}' = (\mathbf{Y}_V^T, \mathbf{Y}_N^T)^T$ and $\mathbf{D}' = (\mathbf{D}_V^T, \mathbf{D}_N^T)^T$. During reconstruction, for instance from NIR to VIS, we obtain the sparse approximation in $\mathbf{D}_N$, *i.e.* $\mathbf{x} = \arg\min_{\mathbf{x}} \|\mathbf{y}_N - \mathbf{D}_N\mathbf{x}\|_F^2$ such that $\forall i, \|\mathbf{x}_i\|_0 < K$ and then obtain the VIS reconstruction as $\mathbf{y}_V = \mathbf{D}_V\mathbf{x}$. Recall that due to the joint constrained learning of the dictionaries, the sparse representation $\mathbf{x}$ is shared between the two domains. For reconstruction from VIS to NIR, one would follow the opposite procedure of representing the image in $\mathbf{D}_V$ first before reconstructing it in $\mathbf{D}_N$. This method is open set thereby allowing the reconstruction of any face that is not present in the training set.

As a final detail, we define K_1 to be the sparsity constraint going from NIR to VIS, *i.e.* $\mathbf{x} = \arg\min_{\mathbf{x}} \|\mathbf{y}_N - \mathbf{D}_N\mathbf{x}\|_F^2$ such that $\forall i, \|\mathbf{x}_i\|_0 < K_1$. Analogously, we define K_2 to be the sparsity constraint in reconstructing from NIR to VIS.

3.3. Choice of Sparsity

Once we have learned the joint dictionary, we can split it into two parts, one corresponds to NIR face images $\mathbf{D}_N$, and the other for VIS face images $\mathbf{D}_V$ as shown in Figure 2. As previously discussed, we can therefore reconstruct face cross spectrum using the coupled dictionary. It is worth noticing that the choice of sparsity level is crucial in sparse coding during the reconstruction. Here we follow a simple greedy search approach to determine the best sparsity level. The fidelity of reconstruction is measured by the peak signal-to-noise ratio (PSNR).

Due to the fact that NIR and VIS images in the CASIA NIR-VIS 2.0 database [24] were captured at different sessions (see section 4), with a lot of other variations such as slight pose, expression *etc*., there is not a single pair of NIR-VIS image that is perfectly aligned, with only spectral variations. However, what we care about is identity preservation after the NIR-VIS reconstruction. We can therefore, as a very rough estimate, determine the sparsity level by cross-validating the PSNR between the original NIR image $\mathbf{y}_N$ and the reconstructed VIS image $\mathcal{F}(\mathbf{y}_N)$ since PSNR is based off the Euclidean distance between the two images. PSNR is used as a rough similarity measure. Here, $\mathcal{F}$ is the NIR-VIS mapping. We repeat this process for all the NIR images in the development set and compute average PSNR accordingly. Similarly, we can also evaluate VIS-NIR reconstruction using the same approach.

The cross-domain mapping $\mathcal{F}$ is non-linear. It is worth noted that the dictionary learning process itself is non-linear due to the OMP step in the sparse coding stage, even though the image can be linearly represented by the dictionary atoms. After dictionaries for both domains are jointly learned, mapping from one domain to the other is done through sparse coding which is again non-linear. A distinction should be made between the linearity in the representation of a dictionary, and the non-linearity in dictionary learning process and the cross-domain mapping in this case.

There are over $40,000$ NIR-VIS image pairs available in the development training set, where each pair is identity consistent[1]. The optimal sparsity levels are determined by using the trained dictionary from the development training set to reconstruct the development testing set (probe set vs. gallery set).

The reconstructed NIR image of the original VIS image $\mathbf{y}_V$ can be represented by $\mathcal{F}^{-1}(\mathbf{y}_V)$, where $\mathcal{F}^{-1}$ is the VIS-NIR mapping. Also, the following relationship should ideally hold: $\|\mathbf{y}_N - \mathcal{F}^{-1} \circ \mathcal{F}(\mathbf{y}_N)\|_F^2 \leq \epsilon,\ \forall \mathbf{y}_N, \exists \epsilon$, as well as $\|\mathbf{y}_V - \mathcal{F} \circ \mathcal{F}^{-1}(\mathbf{y}_V)\|_F^2 \leq \epsilon,\ \forall \mathbf{y}_V, \exists \epsilon$.

The red plot in Figure 3 shows the average PSNR as a function of sparsity level K for NIR-VIS reconstruction evaluation, and the blue plot shows the same for VIS-NIR reconstruction evaluation. Note that we do not expect PSNR to be high, since that could also mean that the reconstruction is very similar to the original domain image. All we need, is to make sure that the PSNR is not too high and to pick the sparsity corresponding to the highest PSNR. Figure 3 shows that indeed the PSNR for all sparsities peak at a reasonable PSNR. In our experiments, the optimal sparsity level for NIR-VIS reconstruction is $K_1 = 53$, and for VIS-NIR reconstruction, the optimal sparsity level is $K_2 = 73$.

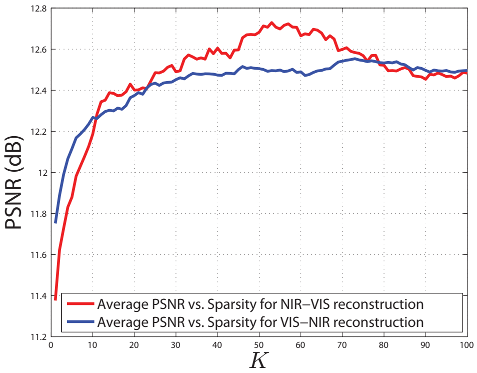

Figure 3. Average PSNR as a function of K for NIR-VIS and VIS-NIR reconstruction evaluations on the development set. Optimal choices of sparsity for NIR-VIS reconstruction is $K_1 = 53$, and for VIS-NIR reconstruction is $K_2 = 73$.

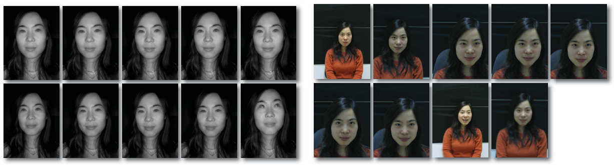

Figure 4. Within-subject variations in the same session for NIR image (left) and VIS image (right) captures.

4. Database and Protocol

The database used in this paper is the CASIA NIR-VIS 2.0 Face Database [24]. This is so far the largest face database across NIR and VIS spectrum, in terms of the number of subjects (725), and the number of face images (17,580). This database also exhibits within-class variations such as pose, expression, eyeglasses, and capture distance.

Figure 4 shows some sample images from this database, which illustrates within-subject variations in the same session. Variations across all four sessions is expected to be more. The database also provides cropped images of resolution 128×128. In our experiments, we down-sample them to 32×32. By doing so, we don't lose performance in face verification, while making overcomplete dictionary training more feasible. The protocol defines two views or subsets of the database. View 1 is meant for algorithm development,

[1] Namely, if a subject has n_1 VIS images and n_2 NIR images available, we obtain a total of $n_1 \times n_2$ VIS-NIR pairs.

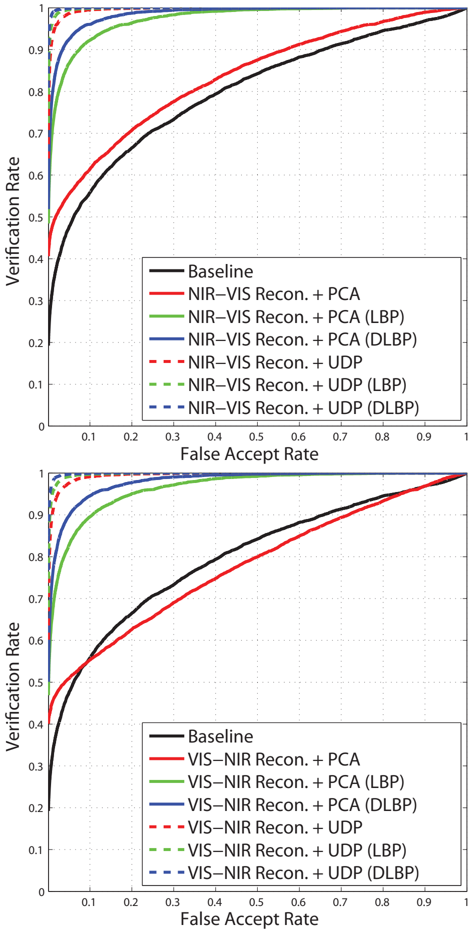

Figure 5. ROC curves for the face verification experiments. (Top) Experiments that convert all the NIR images into VIS ones for both training and testing. (Bottom) Experiments that convert all the VIS images into NIR ones for both training and testing.

using which parameters are to be tuned. View 2 is to be used for performance evaluation which is further divided into 10 folds. For both views, the number of subjects in training and testing are the same. Further, the subjects in the training and the corresponding testing set are non-overlapping. The receiver operating characteristic (ROC) curves, which are generated using all similarity scores across all ten folds, as well as the Rank-1 identification rates are used to evaluate the performance. For the Rank-1 identification rate, the mean accuracy and standard deviation of ten folds should be reported.

5. Experiments and Results

In this section, we first demonstrate the cross-spectral face reconstruction fidelity results using the proposed joint dictionary learning approach. Then we conduct face verification experiments to evaluate whether the proposed method can help improve the face recognition performance.

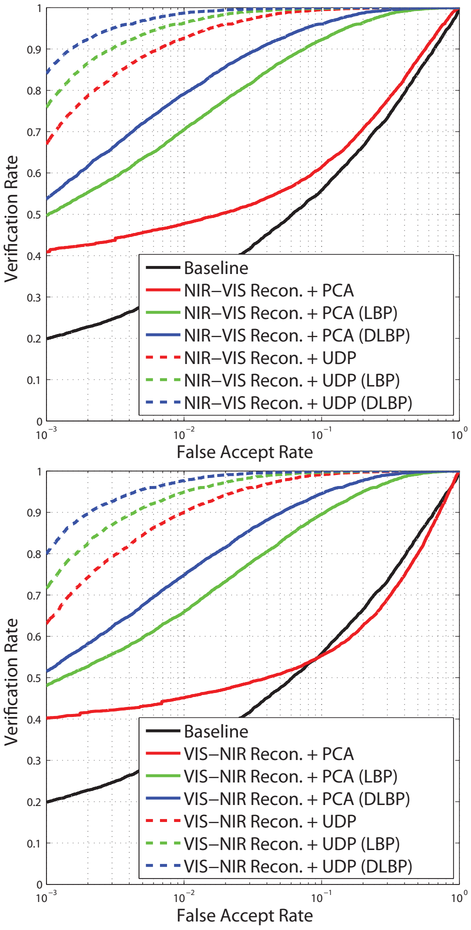

Figure 6. ROC curves for the face verification experiments. (Top) Experiments that convert all the NIR images into VIS ones for both training and testing. (Bottom) Experiments that convert all the VIS images into NIR ones for both training and testing. These two ROC plots are the same as Figure 5, but shown in semi-log scale to emphasize the performance difference at very low FAR.

One implementation detail worth mentioning is that due to the fact that there is no exact cross-spectral mapping for every image in the database, *i.e.* NIR and VIS images for the same subject are not taken at the same time, we need to manually specify all the pairings between NIR and VIS images. For example, subject i has p images in the NIR set, and q images in the VIS set, the way we create the pairing set is to pair each of the NIR image to each of the VIS image of the same subject, resulting in $p \times q$ pairs. Then, we do this for all the subjects on all the images available to us. By doing this, we have significantly augmented the number of training pairs for learning the joint dictionary while making sure that each pair is of the same subject. Therefore the dictionary learns the NIR-VIS mapping $\mathcal{F}$ and the VIS-NIR inverse mapping $\mathcal{F}^{-1}$ while being agnostic about the subjects' identities which is essential for generalizing to unseen subjects.

5.1. Cross-Spectral Face Reconstruction Fidelity

After obtaining the joint cross-spectral dictionary, we split it into the NIR $(\mathbf{D}_N)$ and the VIS $(\mathbf{D}_V)$ part. For any input NIR images, we can first apply sparse coding on the NIR dictionary to obtain the sparse coefficient vector, which will be used to pick out atoms from the VIS dictionary for NIR-VIS reconstruction. Similarly, for VIS-NIR reconstruction.

Quantitative results for face reconstruction fidelity is reported in Table 1 showing mean PSNR across all ten folds for both NIR-VIS and VIS-NIR reconstructions. The mean PSNR for NIR-VIS reconstruction is 12.723 dB and the mean PSNR for VIS-NIR reconstruction is 12.586 dB. Both PSNR readings are considered high because we are comparing original images to its reconstruction into the counterpart domain. Here, PSNR is served as a soft clue for preserving identity.

Figure 8 and Figure 9 show the ten best and worst NIR-VIS reconstruction results from the first NIR probe set. The ranking is according to the PSNR between the reconstructed image and the original image. Similarly, Figure 10 and Figure 11 show the ten best and worst VIS-NIR reconstruction results from the first VIS gallery set. It can be observed that for good reconstruction results, corresponding spectrum feature are clearly reconstructed and the subject identity is well preserved.

By taking a closer look at the worse reconstruction results for both cases as shown in Figure 9 and Figure 11, we come to understand more about the challenges posed by this database. This database is fairly unconstrained in the sense that subjects exhibit facial expression, and pose variations. Also, because the database is a collection of multiple capture sessions, the images within the same subject can be quite different across different sessions. There are also camera-related factors such as blurred images, glare on eye-glasses, and different zooming factors. If an input image exhibits quite unique artificial facial features (for example, one female subject may has her hair blocking the left eye region, or a male subject may wear IR-reflective eye-glasses during the acquisition of NIR images where severe glare is shown on the eye-glasses), or quite extreme facial expression and pose, such unique information will either spike on one or just a few dictionary atoms, or it can not be well represented by atoms at all. Either way, this would lead to poor reconstruction.

5.2. Face Verification Experiments

According to the face verification protocol set forth by [24], there are 10-fold experiments. For the testing part in each fold, the gallery set is always of size 358, which contains one VIS image per subject, a total of 358 subjects. The probe set has over 6,000 NIR images from the same 358 subjects. All the probe NIR images are to be matched against all the VIS gallery images, resulting a similarity matrix of size 358 by around 6,000. The ground truth mask is provided. So the ideal algorithm would result in a block-diagonal structure in the similarity matrix.

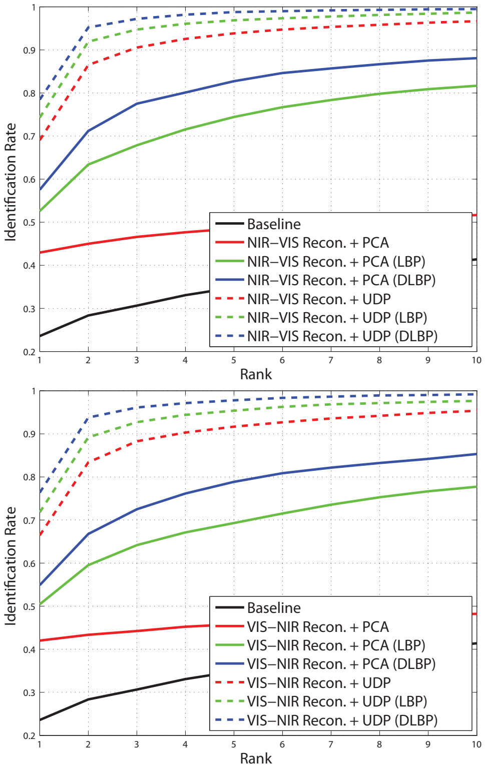

Figure 7. CMC curves for the face verification experiments showing Rank-1 through Rank-10 identification rates. (Top) Experiments that convert all the NIR images into VIS ones for both training and testing. (Bottom) Experiments that convert all the VIS images into NIR ones for both training and testing.

For the training in each fold, there are around 2,500 VIS images and around 6,100 NIR images from around 360 subjects, which are mutually exclusive from the 358 subjects in testing. The training set in each fold is used for learning the joint dictionary as well as for modeling the linear subspace to be discussed. Note that the parameters K_1 and K_2 were optimized using View 1, and were set to the same values in this experiment.

With our proposed cross-spectral reconstruction capability, we can essentially reconstruct the VIS counterparts for all the NIR images in both training and testing set for each fold. By doing this, the classifier learning and face matching will be done entirely in the VIS domain. Similarly, if we reconstruct the NIR counterparts for all the VIS images, we end up training and matching entirely in the NIR domain. The cross-spectral reconstruction can dramatically eliminate the cross-spectral variations, making face matching a more feasible task.

The baseline algorithm is shown in [24], where they have

Table 1. Mean PSNR for both NIR-VIS and VIS-NIR face reconstruction across all ten folds.

	Mean PSNR (dB)	Standard Deviation
NIR-VIS reconstruction across 10 NIR probe sets	12.723	2.26 %
VIS-NIR reconstruction across 10 VIS gallery sets	12.586	2.18 %

Table 2. Experimental results for the 10-fold face verification tasks including mean accuracy (rank-1 identification rate), with its standard deviation, verification rates (VR) at 0.1% false accept rate (FAR), equal error rates (EER), and the area under the ROC curve (AUC).

	Mean Accuracy	Std. Dev.	VR at 0.1% FAR	EER	AUC
Baseline [24] (our own implementation)	0.2358	1.91 %	0.202	0.278	0.799
NIR-VIS Reconstruction + PCA	0.4296	2.25 %	0.415	0.254	0.836
NIR-VIS Reconstruction + PCA (LBP)	0.5259	2.17 %	0.504	0.086	0.973
NIR-VIS Reconstruction + PCA (DLBP)	0.5754	2.24 %	0.548	0.061	0.985
NIR-VIS Reconstruction + UDP	0.6906	1.89 %	0.691	0.028	0.996
NIR-VIS Reconstruction + UDP (LBP)	0.7428	2.15 %	0.778	0.018	0.998
NIR-VIS Reconstruction + UDP (DLBP)	**0.7846**	1.67 %	0.858	0.011	0.999
VIS-NIR Reconstruction + PCA	0.4201	2.20 %	0.403	0.306	0.779
VIS-NIR Reconstruction + PCA (LBP)	0.5043	1.53 %	0.487	0.102	0.965
VIS-NIR Reconstruction + PCA (DLBP)	0.5488	1.77 %	0.523	0.074	0.980
VIS-NIR Reconstruction + UDP	0.6645	1.54 %	0.646	0.035	0.995
VIS-NIR Reconstruction + UDP (LBP)	0.7179	1.59 %	0.734	0.023	0.997
VIS-NIR Reconstruction + UDP (DLBP)	0.7637	2.32 %	0.816	0.014	0.999

achieved 23.70% mean accuracy by using a variant of PCA called Hetero-Component Analysis (HCA) together with augmented samples by face symmetry. We have attempted to re-implement their method and have been able to achieve a mean accuracy of 23.58%, fairly comparable to what is reported in [24].

We divide our experiments into two major parts. The first part is to carry out NIR-VIS reconstruction using proposed method to convert all the images into the VIS domain, both in training and testing stages. The second part is to carry out VIS-NIR reconstruction to convert all the images into the NIR domain, both in training and testing stages. Three features are explored namely raw pixel, local binary patterns (LBP), and DCT encoded local binary patterns (DLBP) [17, 10, 11, 8, 16, 15, 14, 27, 9]. Two linear subspace methods are adopted, namely principal component analysis (PCA) and unsupervised discriminant projections (UDP) [29]. It is worth noticing that both methods are unsupervised and no label information is ever capitalized. Normalized cosine distance (NCD) is used for measuring the similarities between data/feature samples.

The experimental results are consolidated in Table 2 where the mean accuracy (rank-1 identification rate) with its standard deviation cross all ten folds, verification rates (VR) at 0.1% false accept rate (FAR), equal error rates (EER), and the area under the ROC curve (AUC) are shown for each algorithm. Figure 5 shows the ROC curves for both NIR-VIS and VIS-NIR experiments. Figure 6 shows the same ROC curves as Figure 5 but in semi-log scale to emphasize the performance at very low FAR. Figure 7 shows the cumulative match characteristic (CMC) curves with rank-1 through rank-10 identification rates for both parts of the experiments.

5.3. Discussions

From the results of the face verification experiments, we find that the proposed cross-spectral joint dictionary reconstruction can significantly improve the face recognition accuracy by reconstructing the probe and gallery images into the same spectrum domain. Once reconstructed in a common spectrum, face recognition tasks are therefore made easier and less sophisticated classifiers can perform well.

We are able to significantly outperform the baseline [24] as well as some good results reported in [4] (73.28%) by obtaining a high 78.46% mean accuracy which to the best of our knowledge is currently state-of-the-art. The best performing algorithm is to reconstruct all the NIR probe images into VIS ones and then apply the DLBP feature, followed by the UDP subspace method. We also show that competitive results are achieved by NIR-VIS reconstruction, when compared to that of VIS-NIR reconstruction, which showcases the mutuality of the proposed method, reconstructing from one domain to the other.

6. Conclusion and Future Work

In this paper, we present a cross-spectral joint dictionary learning technique to reconstruct images between the NIR and VIS domain. Our method is open set and can reconstruct faces not present in the training set. Further, once an image is reconstructed in either domain, any FRS can be used to match. We experiment with a few feature-classifier pairings and find that they perform very well after all images were reconstructed in either domain.

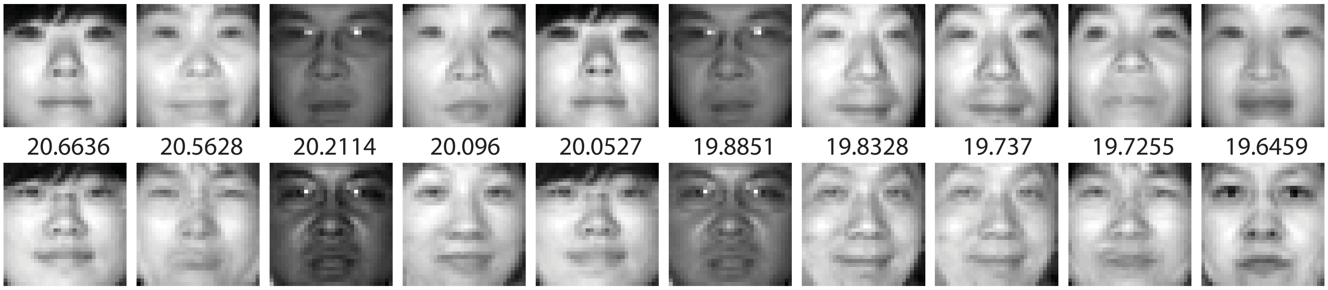

Figure 8. Ten *best* NIR-VIS reconstruction results in terms of the PSNR from the first NIR probe set. The first row shows original NIR images, the second row shows NIR-VIS reconstructions along with the PSNR.

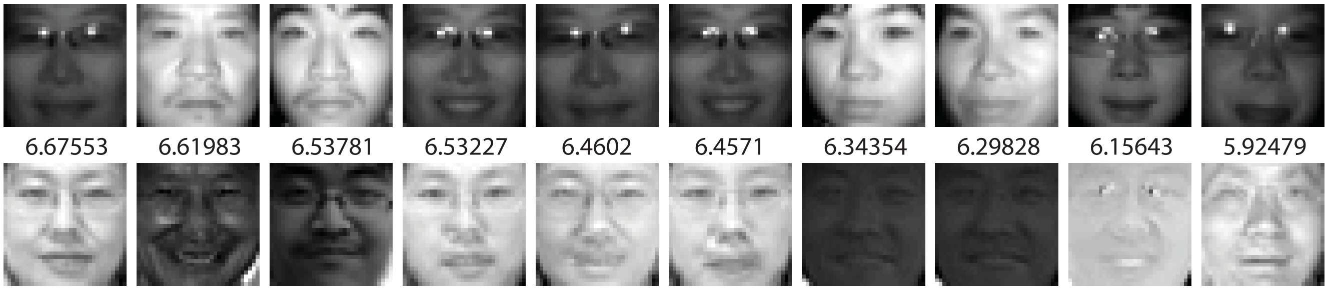

Figure 9. Ten *worst* NIR-VIS reconstruction results in terms of the PSNR from the first NIR probe set. The first row shows original NIR images, the second row shows NIR-VIS reconstructions along with the PSNR.

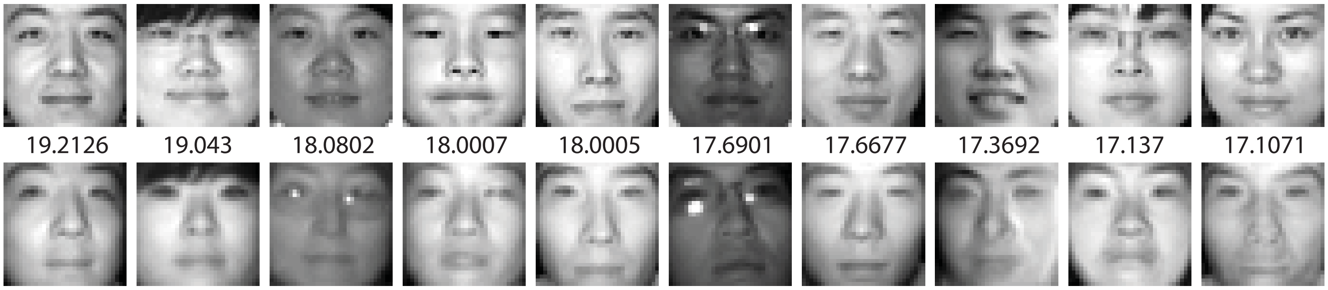

Figure 10. Ten *best* VIS-NIR reconstruction results in terms of the PSNR from the first VIS gallery set. The first row shows original VIS images, the second row shows VIS-NIR reconstructions along with the PSNR.

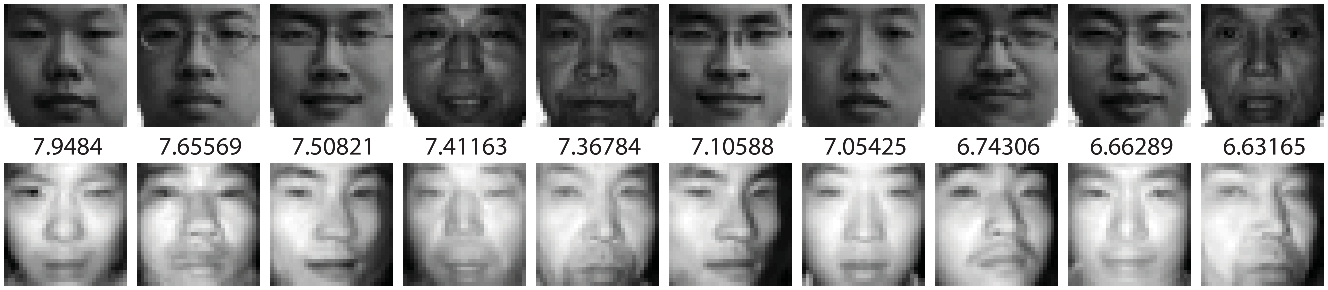

Figure 11. Ten *worst* VIS-NIR reconstruction results in terms of the PSNR from the first VIS gallety set. The first row shows original VIS images, the second row shows VIS-NIR reconstructions along with the PSNR.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, Nov 2006. 3

[2] T. Bourlai and B. Cukic. Multi-spectral face recognition: Identification of people in difficult environments. In *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on*, pages 196–201, June 2012. 2

[3] Z. Cao and N. Schmid. Matching heterogeneous periocular regions: Short and long standoff distances. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 4967–4971, Oct 2014. 2

[4] T. Dhamecha, P. Sharma, R. Singh, and M. Vatsa. On Effectiveness of Histogram of Oriented Gradient Features for Visible to Near Infrared Face Matching. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1788–1793, Aug 2014. 2, 7

[5] D. Goswami, C. H. Chan, D. Windridge, and J. Kittler. Evaluation of face recognition system in heterogeneous environments (visible vs NIR). In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2160–2167, Nov 2011. 2

[6] C.-A. Hou, M.-C. Yang, and Y.-C. Wang. Domain Adaptive Self-Taught Learning for Heterogeneous Face Recognition. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 3068–3073, Aug 2014. 1, 2

[7] R. Jillela and A. Ross. Matching face against iris images using periocular information. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 4997–5001, Oct 2014. 2

[8] F. Juefei-Xu, M. Cha, J. L. Heyman, S. Venugopalan, R. Abiantun, and M. Savvides. Robust Local Binary Pattern Feature Sets for Periocular Biometric Identification. In *Biometrics: Theory Applications and Systems (BTAS), 4th IEEE Int'l Conf. on*, pages 1–8, sep 2010. 7

[9] F. Juefei-Xu, M. Cha, M. Savvides, S. Bedros, and J. Trojanova. Robust Periocular Biometric Recognition Using Multi-level Fusion of Various Local Feature Extraction Techniques. In *IEEE 17th International Conference on Digital Signal Processing (DSP)*, 2011. 7

[10] F. Juefei-Xu, K. Luu, M. Savvides, T. Bui, and C. Suen. Investigating Age Invariant Face Recognition Based on Periocular Biometrics. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–7, Oct 2011. 7

[11] F. Juefei-Xu, D. K. Pal, and M. Savvides. Hallucinating the Full Face from the Periocular Region via Dimensionally Weighted K-SVD. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, June 2014. 7

[12] F. Juefei-Xu, D. K. Pal, K. Singh, and M. Savvides. A Preliminary Investigation on the Sensitivity of COTS Face Recognition Systems to Forensic Analyst-style Face Processing for Occlusions. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on*, June 2015. 2

[13] F. Juefei-Xu and M. Savvides. Can Your Eyebrows Tell Me Who You Are? In *Signal Processing and Communication Systems (ICSPCS), 2011 5th International Conference on*, pages 1–8, Dec 2011. 2

[14] F. Juefei-Xu and M. Savvides. Unconstrained Periocular Biometric Acquisition and Recognition Using COTS PTZ Camera for Uncooperative and Non-cooperative Subjects. In *Applications of Computer Vision (WACV), 2012 IEEE Workshop on*, pages 201–208, Jan 2012. 7

[15] F. Juefei-Xu and M. Savvides. An Augmented Linear Discriminant Analysis Approach for Identifying Identical Twins with the Aid of Facial Asymmetry Features. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 56–63, June 2013. 7

[16] F. Juefei-Xu and M. Savvides. An Image Statistics Approach towards Efficient and Robust Refinement for Landmarks on Facial Boundary. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–8, Sept 2013. 7

[17] F. Juefei-Xu and M. Savvides. Subspace Based Discrete Transform Encoded Local Binary Patterns Representations for Robust Periocular Matching on NIST's Face Recognition Grand Challenge. *IEEE Trans. on Image Processing*, 23(8):3490–3505, aug 2014. 7

[18] F. Juefei-Xu and M. Savvides. Facial Ethnic Appearance Synthesis. In *Computer Vision - ECCV 2014 Workshops*, volume 8926 of *Lecture Notes in Computer Science*, pages 825–840. Springer International Publishing, 2015. 2

[19] F. Juefei-Xu and M. Savvides. Weight-Optimal Local Binary Patterns. In *Computer Vision - ECCV 2014 Workshops*, volume 8926 of *Lecture Notes in Computer Science*, pages 148–159. Springer International Publishing, 2015. 2

[20] N. Kalka, T. Bourlai, B. Cukic, and L. Hornak. Cross-spectral face recognition in heterogeneous environments: A case study on matching visible to short-wave infrared imagery. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–8, Oct 2011. 2

[21] B. Klare and A. Jain. Heterogeneous Face Recognition: Matching NIR to Visible Light Images. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1513–1516, Aug 2010. 2

[22] Z. Lei and S. Li. Coupled Spectral Regression for matching heterogeneous faces. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1123–1128, June 2009. 2

[23] J. Li, Y. Jin, and Q. Ruan. Matching NIR face to VIS face using multi-feature based MSDA. In *Signal Processing (ICSP), 2014 12th International Conference on*, pages 1443–1447, Oct 2014. 2

[24] S. Li, D. Yi, Z. Lei, and S. Liao. The CASIA NIR-VIS 2.0 Face Database. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 348–353, June 2013. 3, 4, 6, 7

[25] S. Liu, D. Yi, Z. Lei, and S. Li. Heterogeneous face image matching using multi-scale features. In *Biometrics (ICB), IAPR Int'l Conf. on*, pages 79–84, March 2012. 2

[26] F. Nicolo and N. Schmid. Long Range Cross-Spectral Face Recognition: Matching SWIR Against Visible Light Images.

Information Forensics and Security, IEEE Transactions on, 7(6):1717–1726, Dec 2012. 2

[27] M. Savvides and F. Juefei-Xu. Image Matching Using Subspace-Based Discrete Transform Encoded Local Binary Patterns, Sept. 2013. US Patent US 2014/0212044 A1. 7

[28] K. Seshadri, F. Juefei-Xu, D. K. Pal, and M. Savvides. Driver Cell Phone Usage Detection on Strategic Highway Research Program (SHRP2) Face View Videos. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on*, June 2015. 2

[29] J. Yang, D. Zhang, J.-Y. Yang, and B. Niu. Globally Maximizing, Locally Minimizing: Unsupervised Discriminant Projection with Applications to Face and Palm Biometrics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(4):650–664, April 2007. 7

[30] J.-Y. Zhu, W.-S. Zheng, and J. Lai. Transductive VIS-NIR face matching. In *Image Processing (ICIP), IEEE International Conference on*, pages 1437–1440, Sept 2012. 2

[31] J.-Y. Zhu, W.-S. Zheng, J.-H. Lai, and S. Li. Matching NIR Face to VIS Face Using Transduction. *Information Forensics and Security, IEEE Transactions on*, 9(3):501–514, March 2014. 2

[32] J. Zuo, F. Nicolo, N. Schmid, and S. Boothapati. Encoding, matching and score normalization for cross spectral face recognition: Matching SWIR versus visible data. In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, pages 203–208, Sept 2012. 2