Not Afraid of the Dark: NIR-VIS Face Recognition via Cross-spectral Hallucination and Low-rank Embedding

José Lezama[1]*, Qiang Qiu[2]* and Guillermo Sapiro[2]
[1]IIE, Universidad de la República, Uruguay. [2]ECE, Duke University, USA.

Abstract

Surveillance cameras today often capture NIR (near infrared) images in low-light environments. However, most face datasets accessible for training and verification are only collected in the VIS (visible light) spectrum. It remains a challenging problem to match NIR to VIS face images due to the different light spectrum. Recently, breakthroughs have been made for VIS face recognition by applying deep learning on a huge amount of labeled VIS face samples. The same deep learning approach cannot be simply applied to NIR face recognition for two main reasons: First, much limited NIR face images are available for training compared to the VIS spectrum. Second, face galleries to be matched are mostly available only in the VIS spectrum. In this paper, we propose an approach to extend the deep learning breakthrough for VIS face recognition to the NIR spectrum, without retraining the underlying deep models that see only VIS faces. Our approach consists of two core components, cross-spectral hallucination and low-rank embedding, to optimize respectively input and output of a VIS deep model for cross-spectral face recognition. Cross-spectral hallucination produces VIS faces from NIR images through a deep learning approach. Low-rank embedding restores a low-rank structure for faces deep features across both NIR and VIS spectrum. We observe that it is often equally effective to perform hallucination to input NIR images or low-rank embedding to output deep features for a VIS deep model for cross-spectral recognition. When hallucination and low-rank embedding are deployed together, we observe significant further improvement; we obtain state-of-the-art accuracy on the CASIA NIR-VIS v2.0 benchmark, without the need at all to re-train the recognition system.

1. Introduction

In a typical forensic application involving night-time surveillance cameras, a probe image of an individual is captured in the near-infrared spectrum (NIR), and the individual must be recognized out of a gallery of visible spectrum (VIS) images. Whilst VIS face recognition is an extensively studied subject, face recognition in the NIR spectrum remains a relatively unexplored field.

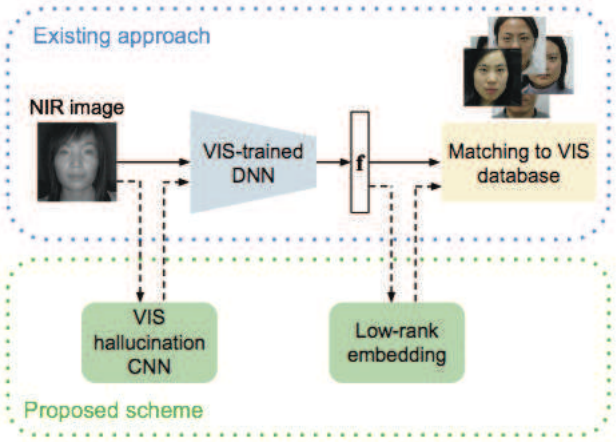


Figure 1. Diagram of the proposed approach. A simple NIR-VIS face recognition system consists in using a Deep Neural Network (DNN) trained only on VIS images to extract a feature vector **f** from a NIR image and use it for matching to a VIS database. We propose two modifications to this basic system. First, we modify the input by hallucinating a VIS image from the NIR sample. Secondly, we apply a low-rank embedding of the DNN features at the output. Each of this modifications produces important improvements in the recognition performance, and an even greater one when applied together.

In standard VIS face recognition, impressive progress has been achieved recently. This is due in part to the excellent performance of deep neural networks [29, 33, 36, 40], which benefit from the availability of very large datasets of face photos, typically mined from the Internet [16, 40]. Such a fruitful strategy for data collection cannot be applied to NIR images, where training data is much scarcer.

Naturally, one would like to leverage the power of state-of-the-art VIS face recognition methods and apply them to NIR. Recent works have made significant progress in this direction [18, 32], but the recognition rates remain much lower than the ones achieved in the VIS spectrum. In this

*Denotes equal contribution.

work, we take a major step towards closing that gap.

We consider using a pre-trained deep neural network (DNN) which has seen only VIS images as a black-box feature extractor, and propose convenient processing of the input and output of the DNN that produce a significant gain in NIR-VIS recognition performance. The proposed approach, summarized in Figure 1, consists of two components, cross-spectral hallucination and low-rank embedding, to optimize respectively the input and output of a pre-trained VIS DNN model for cross-spectral face recognition.

First, we propose to modify the NIR probe images using deep cross-spectral hallucination based on a convolutional neural network (CNN).[1] The CNN learns a NIR-VIS mapping on a patch-to-patch basis. Then, instead of inputting the NIR probe directly to the feature extraction DNN, we input the cross-spectrally hallucinated. Secondly, we propose to embed the output features of the DNN using a convenient low-rank transform [30], making the transformed NIR and VIS features of one subject lie in the same low-dimensional space, while separating them from the other subjects' representations. While here illustrated for the important problem of face recognition, this work provides a potential new direction that combines transfer learning (at the input/output) with joint embedding (at the output).

The two proposed strategies achieve state-of-the-art results when applied separately, and achieve an even more significant improvement when applied in combination. We demonstrate that both techniques work well independently of the DNN used for feature extraction.

2. Related Work

One common strategy for NIR-VIS face recognition, employed since the early work of Yi et al. [41], is to find a mapping of both the NIR and VIS features to a shared subspace, where matching and retrieval can be performed for both spectrums at the same time. This metric learning strategy was applied in many successive works [11, 15, 22, 28], and more recently using DNN-extracted features [29, 32]. Most metric learning methods learn metrics with triplet [38] or pairwise [7] constraints. The triplet loss is often adopted in deep face models to learn a face embedding [29, 33]. Saxena and Verbeek [32] studied the performance implications of using different feature layers of the DNN in combination with different metric learning approaches, and propose a combination of the two. In this work, we consider the well developed DNN as a black box and use the features produced by the DNN at the second-to-last layer. We adopt a low-rank constraint to learn a face embedding, which proves effective for cross-spectral tasks.

[1]To avoid confusion, in this paper we will refer to the deep neural network used for feature extraction as DNN and to the convolutional neural network used to hallucinate full-resolution VIS images from NIR as CNN.

Another strategy is to convert the NIR probe to a VIS image [18, 23, 31] and apply standard VIS face recognition to the converted version of the probe. One of the first to utilize this strategy for face VIS hallucination, Li et al. [23], learn a patch-based linear mapping from a middle- and long-wavelength infrared (MW-/LWIR) image to a VIS image, and regularize the resulting patches with an MRF. Juefei et al. [18], used a cross-spectral dictionary learning approach to successfully map a NIR face image to the VIS spectrum. On the obtained VIS image, they apply Local Binary Patterns (LBP) [2] and a linear classifier to perform face recognition. Converting from infrared to VIS is a very challenging problem, but has the clear advantage of allowing to use existing traditional face recognition algorithms on the converted images. To the best of our knowledge, this is the first time a deep learning approach is used to hallucinate VIS faces from NIR.

Several works exist for the task of cross-spectral conversion of outdoor scenes [14, 37, 44]. This scenario has the advantage that more multispectral data exists for generic scenes [4]. Building a dataset of cross-spectral face imaging with correct alignment for a significant number of subjects is a much more challenging task. We believe it is in part due to this difficulty that few works exist in this direction.

Given the advantage of thermal images not requiring a light source, a lot of attention has been given to the thermal to VIS face recognition task [3, 5, 31]. Related to our work, Sarfraz et al. [31] used a neural network to learn the reverse mapping, from VIS to MW-/LWIR, so that a thermal face image could be matched to its VIS counterpart. This strategy has the disadvantage of having to apply the mapping to each VIS image in the dataset. We propose to use a convolutional neural network to compute the mapping between the (single) tested NIR and VIS images.

Another important family of work on cross-spectral face recognition focuses on the features used for recognition, and suggested strategies include engineering light source invariant features (LSIF) [25], performing cross-domain feature learning [17, 27, 39, 45], and adapting traditional handcrafted features [8, 21].

Alternative approaches fit existing deep neural networks to a given database, e.g., [26, 13], achieving as expected improvements in the results for that particular dataset. Contrary to those, the generalization power of our proposed framework is born from the technique itself; without any kind of re-training we achieve state-of-the-art results. This is obtained while enjoying the hard work (and huge training) done for existing networks, simply by adding trivial hallucination and linear steps. As underlying networks improve, the proposed framework here introduced will potentially continue to improve without expenses on training or data collection.

In this paper, we build on the ideas of [32] and [18]. We

use a DNN pre-trained on a huge dataset of VIS images as a feature extractor. At the input of this DNN, we propose to preprocess the NIR input using deep cross-modal hallucination. At the output, we propose to embed the feature vector using a low-rank transform. The simplicity of the approach and the use of off-the-shelf well optimized algorithms is part of the virtue of this work. As a side contribution, we derive a secondary dataset from the CASIA NIR-VIS 2.0 dataset [24] consisting of more than 1.2 million pairs of aligned NIR-VIS patches.

3. Cross-spectral Hallucination

Most if not all DNN face models are designed and trained to operate on VIS face images, thanks to the availability of enormous VIS face datasets. It is to expect that such deep models do not achieve their full potential when their input is a NIR face image. In this section we propose to preprocess the NIR image using a convolutional neural network (CNN) that performs a cross-spectral conversion of the NIR image into the VIS spectrum. We will show that using the hallucinated VIS image as input to the feature extraction DNN, instead of the raw NIR, produces a significant gain in the recognition performance. Note that the goal here is not necessarily to produce a visually pleasant image, but a VIS image that is better fit for a VIS-pretrained DNN than the NIR.

The cross-spectral hallucination CNN is trained on pairs of corresponding NIR-VIS patches that are mined from a publicly available dataset, as will be described below. In the VIS domain, we work in a luminance-chroma color space, since it concentrates the important image details in the luminance channel and minimizes the correlation between channels, making the learning more efficient. We find the YCbCr space to give the best results, and we observe no difference between training the three channels with shared layers or independently. For simplification we train three different networks.

The network architecture is inspired in [9]. Because the luminance channel Y contains most of the subject's information, we utilize a bigger network for this channel and smaller networks for the two chromas. Also, because the blue component varies very little in faces, an even smaller network is enough for the blue-difference chroma Cb. The details of the networks architecure is shown in Table 1. The three networks receive a 40x40 input NIR patch and consist of successive convolutional layers with stride 1, no pooling and PReLU activation functions [12] (except in the last layer), and an Euclidean loss function at the end. We pad each layer with zeros to have the same size at the input and output. The three networks have an hour-glass structure where the depth of the middle layers is narrower than the first and last layers [9]. This makes for efficient training while allowing highly non-linear relations to be learned.

Figure 2. Sample images of two subjects from the CASIA NIR-VIS 2.0 Face Dataset. Top: NIR. Bottom: VIS.

3.1. Mining for NIR-VIS patches

We use the CASIA NIR-VIS 2.0 dataset [24] to obtain pairs of NIR-VIS patches. This dataset contains 17,580 NIR-VIS pairs of images with an average of 24 images of each modality for each subject. This dataset cannot be used for training directly because the NIR-VIS image pairs are not aligned and the subjects pose and facial expression vary a lot (Figure 2). In [18], this problem was partially avoided by subsampling 128x128 crops of the original images to 32x32 images. Yet in their reported results, some smoothing and visual artifacts due to the training set misalignment can be observed ([18], Figure 8). In this work we perform no subsampling. Instead, we mine the CASIA NIR-VIS 2.0 dataset for consistent pairs of NIR-VIS patches at the best possible resolution, Figure 3. Through this process we are able to derive a secondary dataset with more than one million pairs of 40x40 NIR-VIS image patches.

We use the luminance channel and the NIR image to find correspondences. The first step is to pre-process the images by aligning the facial landmarks of the two modalities (centers of the two eye pupils and center of the mouth), using [19], and cropping the images to 224x224 pixels. Secondly

Ch.	layers	first and last	intermediate	skip-connections
Y	11	148x11x11 str. 1, pad 5 PReLU	36x11x11 str. 1, pad 5 PReLU	input to last layer
Cb	7	66x3x3 str. 1, pad 1 PReLU	32x3x3 str. 1, pad 1 PReLU	none
Cr	8	148x5x5 str. 1, pad 2 PReLU	48x5x5 str. 1, pad 2 PReLU	none

Table 1. Architecture of the CNN used for cross-spectral hallucination. The first and last layers have deeper filters than the layers in-between, mimicking an encoding-decoding scheme.

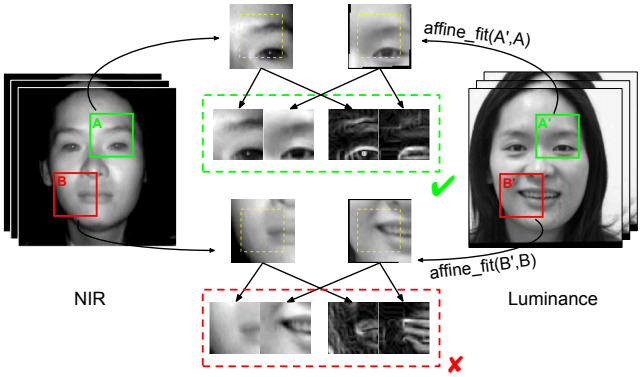


Figure 3. Mining the CASIA NIR-VIS 2.0 dataset for valid patch correspondences. We compare every NIR image to the luminance channel of every VIS image for each subject. Note that the NIR and VIS images are captured under different pose and facial expression. We use 224x224 crops of the original images with facial landmarks aligned. A sliding 60x60 patch is extracted at the same location from both images. The VIS patch is affine-registered to the NIR patch. We then crop a 40x40 region inside the registered patches. We keep the pair if the correlation of the two patches and of their gradients are above a threshold. In this example, patches A and A' form a qualifying pair, whilst B and B' are discarded.

we normalize the mean and standard deviation of the NIR and color channels of all the images in the dataset with respect to a fixed reference chosen from the training set. The facial landmark alignment is insufficient, as discrepancies between the NIR and VIS images still occur. Training a CNN with even slightly inconsistent pairs produces strong artifacts at the output. In order to obtain a clean training dataset, we run a sliding window of 60x60 pixels, with stride 12, through both images, and extract patches at the same locations. Note that the patches are roughly aligned based on the facial landmarks, but this alignment is typically not fully accurate. We then fit an affine transform between the 60x60 luminance patch and NIR patch. Next, we crop the center 40x40 regions and compute a similarity score to evaluate if the patches of both modalities coincide. The similarity score consists of the correlation between the patches plus the correlation between their gradient magnitudes. If the sum of both values is above 1 and neither of them is below 0.4 we consider the pair a valid match. Note that a cross-spectral NIR-VIS patch similarity metric using a CNN was proposed in [1]. In this work we opt for using plain correlation for the sake of efficiency.

This patch mining strategy allowed us to collect more than 700,000 pairs of NIR-VIS patches. We then pruned this dataset to ensure that the patches were approximately uniformly distributed around the face. After this pruning we kept a total of 600,000 patches. We horizontally flip them to form a final dataset of 1.2 million aligned NIR-VIS patches that we use for training and validation of the cross-spectral hallucination CNN. Figure 4 shows example input and output patches, including the result of the hallucination CNN for patches not seen during training. The convolutional filters learned by the three networks can be applied to any NIR image to produce a VIS image equivalent. Note that we retained the subject identification for each patch so that the dataset can be split without subject overlap.

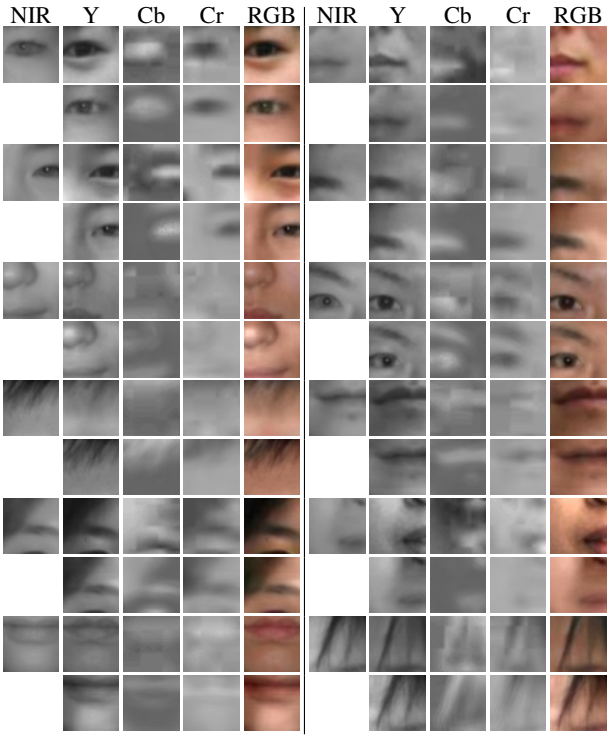


Figure 4. Example patches extracted from CASIA-NIS-VIR 2.0 database using the proposed patch mining method. For each patch, the top row shows the NIR input and ground truth Y, Cb, Cr and RGB signals, and the bottom row shows the output of the cross-spectral hallucination CNN. The Cb and Cr values have been scaled for better visualization. In total, we were able to mine 1.2 million NIR-VIS pairs, equally distributed along the face. All the patches in this figure belong to the validation set and have not been seen during training. (Best viewed in electronic format.)

3.2. Post-processing

Ideally, one would not like to lose all the rich information contained in the original NIR image. Despite our methodology for mining aligned patches, it is not at all impossible that the CNN introduces small artifacts in unseen patches. To safeguard the valuable details of the original NIR, we propose to blend the CNN output with the original NIR image. A successful blending smooths the result of the cross-spectral hallucination and maintains valid information from the pure NIR image. We will later analyze this fact in the experimental section. We perform the blending only on the

luminance channel, by computing the image

$$Y = \hat{Y} - \alpha \cdot G_\sigma^2 * (N_{ir} - \hat{Y}), \quad (1)$$

where Y is the final luminance channel estimation, $\hat{Y}$ is the output of the cross-spectral CNN, N_{ir} is the NIR image, G_σ is a Gaussian filter with $\sigma = 1$, and $*$ denotes convolution. The parameter α balances the amount of information retained from the NIR images and the information obtained with the CNN and allows to remove some of the artifacts introduced by the CNN ($\alpha = 0.6$ in our experiments).

Figure 5 shows example results of the cross-modal hallucination for subjects not seen during training. Note how the blending helps correcting some of the remaining artifacts in the CNN output but maintaining a more natural-looking face than the NIR alone.

4. Low-rank Embedding

In this section, we propose a simple way to extend a DNN model pre-trained on VIS face images to the NIR spectrum, using low-rank embedding at the output layer. The mathematical framework behind low-rank embedding is introduced in [30], where a geometrically motivated transformation is learned to restore a within-class low-rank structure, and meanwhile introduce a maximally separated inter-class structure. With a low-rank embedding layer appended at the end, a DNN model that sees only VIS images produces deep features for VIS and NIR images (or the previously described hallucinated images) in a common space.

4.1. Low-rank Transform

Many high-dimensional data often reside near single or multiple subspaces (or after some non-linear transforms). Consider a matrix $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N \subseteq \mathbb{R}^d$, where each column $\mathbf{y}_i$ is a data point in one of the C classes. Let $\mathbf{Y}_c$ denote the submatrix formed by columns of $\mathbf{Y}$ that lie in the c-th class. A $d \times d$ low-rank transform $\mathbf{T}$ is learned to minimize

$$\sum_{c=1}^{C} ||\mathbf{T}\mathbf{Y}_c||_* - ||\mathbf{T}\mathbf{Y}||_*, \quad (2)$$

where $||\cdot||_*$ denotes the matrix nuclear norm, i.e., the sum of the singular values of a matrix. The nuclear norm is the convex envelope of the rank function over the unit ball of matrices [10]. An additional condition $||\mathbf{T}||_2 = 1$ is originally adopted to prevent the trivial solution $\mathbf{T} = 0$. We drop this normalization condition in this paper, as we never empirically observe such trivial solution, with $\mathbf{T}$ being initialized as the identity matrix. The objective function (2) is a difference of convex functions program, and the minimization is guaranteed to converge to a local minimum (or a stationary point) using the concave-convex procedure [34, 42].

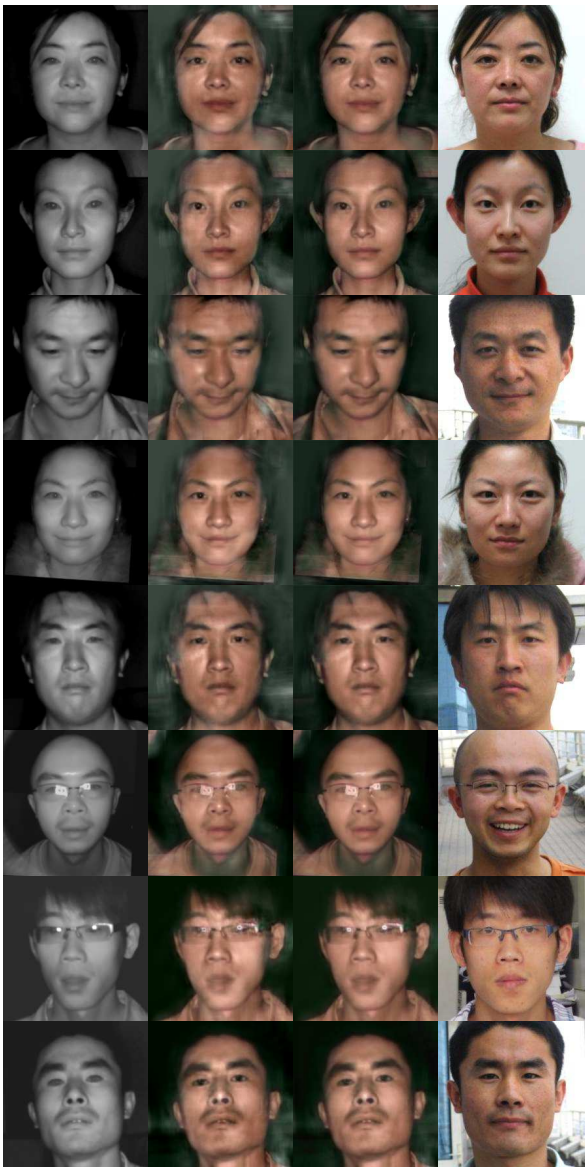
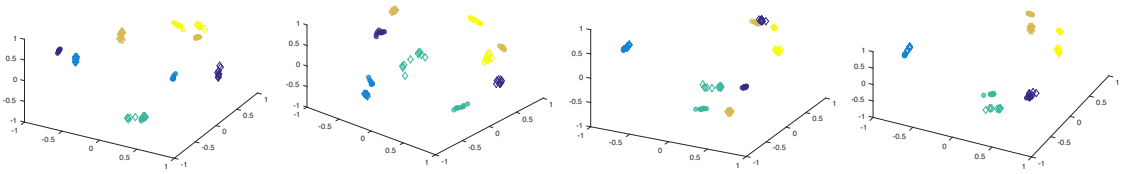


Figure 5. Results of the deep cross-modal hallucination for subjects in the validation set. From left to right: Input NIR image; Raw output of the hallucination CNN; output of the CNN after post-processing; one RGB sample for each subject. The post-processing helps removing some of the artifacts in the CNN output. See for example the faces with glasses, which cause the CNN to create notorious artifacts. Note that the CNN was trained only on face patches so the color of the clothes cannot be hallucinated. (Best viewed in electronic format.)

Theorem 1. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be matrices of the same row dimensions, and* $[\mathbf{A}, \mathbf{B}]$ *denote their column-wise concatenation. Then,* $||[\mathbf{A}, \mathbf{B}]||_* \leq ||\mathbf{A}||_* + ||\mathbf{B}||_*$, *and equality holds if the column spaces of* $\mathbf{A}$ *and* $\mathbf{B}$ *are orthogonal.*

Proof. This results from properties of the matrix nuclear norm.



(a) No embedding (b) Pairwise embedding (c) Triplet embedding (d) Low-rank embedding

Figure 6. Deep features generated using the VGG-face model [29] for VIS (filled circle) and NIR (unfilled diamond) face images from five subjects, one color per subject. Data are visualized in two dimensions using PCA. In (a), without embedding, VIS and NIR faces from the same subject often form two different clusters respectively. In (d), low-rank embedding successfully restores a low-rank structure for multi-spectrum faces from the same subject. In (b) and (c), popular pair-wise and triplet embeddings still show significant intra-class variations across spectrums (best viewed zooming on screen).

Based on Theorem 1, the objective function (2) is non-negative, and it achieves the value zero if, after applying the learned transformation $\mathbf{T}$, the column spaces corresponding to different classes become orthogonal (that is, the smallest principal angle between two different subspaces is $\frac{\pi}{2}$). Note that minimizing each of the nuclear norms appearing in (2) serves to reduce the variation within a class. Thus, the low-rank transform simultaneously minimize intra-class variation and maximize inter-class separations.

4.2. Cross-spectral Embedding

In traditional single-spectrum VIS face recognition, DNN models adopt the classification objective such as softmax at the final layer. Deep features produced at the second-to-last layer are often l_2-normalized, and then compared using the cosine similarity to perform face recognition [29, 35]. Thus, successful face DNN models expect deep features generated for VIS faces from the same subject to reside in a low-dimensional subspace.

In Figure 6a we illustrate the following, which motivates the proposed embedding: Face images from five subjects in VIS and NIR are input to VGG-face [29], one of the best publicly available DNN face models. The generated deep features are visualized in two dimensions using PCA, with filled circle for VIS, unfilled diamond for NIR, and one color per subject. We observe that VIS and NIR faces from the same subject often form two different clusters respectively. Such observation indicates that a successful DNN face model pre-trained on VIS faces is able to generate discriminative features for NIR faces; however, when a subject is imaged under a different spectrum, the underlying low-rank structure assumption is often violated.

Our finding is that the low-rank transform $\mathbf{T}$ in (2) can still effectively restore for the same subject a low-rank structure, even when $\mathbf{Y}_c$ contains mixed NIR and VIS training data from the c-th subject. As no DNN retraining is required, a very important advantage of our approach in practice, the learned low-rank transform can be simply appended as a linear embedding layer after the DNN output layer to allow a VIS model accepting both VIS and NIR images. As shown in Figure 6d, low-rank embedding effectively unifies cross-spectral deep features in Figure 6a.

In DNN-based face recognition, deep feature embeddings with pairwise or triplet constraints are commonly used [29, 32, 33]. Two popular DNN embedding schemes, pair-wise (ITML [7]) and triplet (LMNN [38]) embeddings, are shown in Figure 6b and Figure 6c respectively, and contrary to our approach, significant intra-class variations, i.e., the distance between same color clusters, are still observed across spectrums.

5. Experimental Evaluation

We consider pre-trained VIS DNNs as black-boxes, thereby enjoying the advances in VIS recognition, and perform cross-spectral hallucination to input NIR images, and/or low-rank embedding to output features, for cross-spectral face recognition. To demonstrate that our approach is generally applicable for single-spectrum DNNs without any re-training, we experiment with three pre-trained VIS DNN models from different categories:

- The **VGG-S** model is our trained-from-scratch DNN face model using the VGG-S architecture in [6].
- The **VGG-face** model is a publicly available DNN face model,[2] which reports the best result on the LFW face benchmark among publicly available face models.
- The **COTS** model is a commercial off the shelf (COTS) DNN face model to which we have access.

5.1. Dataset

The CASIA NIR-VIS 2.0 Face Dataset [24] is used to evaluate NIR-VIS face recognition performance. This is the largest available NIR-VIS face recognition dataset and contains 17,580 NIR and VIS face images of 725 subjects. This dataset presents variations in pose, illumination, expressions, etc. Sample images are shown in Figure 2. The

[2]Downloaded at http://www.robots.ox.ac.uk/~vgg/software/vgg_face.

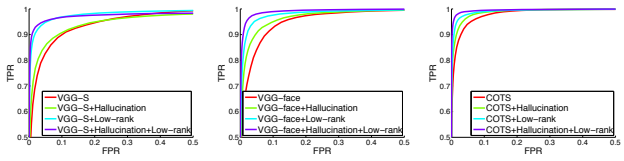


Figure 7. ROC curves for Table 2

CASIA-Webface dataset [40] is used to train our trained-from-scratch VGG-S model. CASIA-Webface is one of the largest public face datasets containing 494,414 VIS face images from 10,575 subjects collected from IMDB.

5.2. Hallucination Networks Protocol

We first train the three CNNs used to hallucinate VIS faces from input NIR images described in Table 1. We use our mined dataset of NIR-VIS image patches. Given that not all the images in the CASIA NIR-VIS 2.0 dataset provide the same amount of aligned patches, the standard protocol for this dataset, which splits the dataset in two, does not provide enough training data for the cross-spectral hallucination CNN. For that reason, to properly evaluate the hallucination contribution we split the dataset into 6 folds. We use five folds (1,030,758 pairs of patches) for training and one fold (206,151 pairs) for testing.[3] The folds are not arbitrary, but follow the natural order of the numbering scheme of the original dataset. We make sure that there is no subject overlap between the training and testing dataset.

We implement the luminance and chroma hallucination networks in the Caffe learning framework. We train the three networks using ADAM optimization [20], with initial learning rate 10^{-5}, and the standard parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$. We observe it is enough to train the networks for 10 epochs.

5.3. Cross-spectral Face Recognition Protocol

Our goal is to match a NIR probe face image with VIS face images in the gallery. We consider three pre-trained single-spectrum DNNs, VGG-S, VGG-face and COTS, as black-boxes, and only modify their inputs (cross-spectral hallucination) and/or outputs (low-rank embedding) for cross-spectral face recognition. All three models expect RGB inputs. When no VIS hallucination is used, we replicate the single-channel NIR images from the CASIA NIR-VIS 2.0 dataset into three channels to ensure compatibility. When hallucination is used, we first apply the hallucination CNN to the NIR images and then feed the hallucinated VIS images to the single-spectrum DNN.

We generate deep features from the respective DNN models, which are reduced to 1024 dimensions using PCA.

[3]Our data partition protocol is included in supplementary material, and will be available to the public for reproducing our experimental results.

	Accuracy (%)
VGG-S	75.04
VGG-S + Colorization [43]	76.14
VGG-S + Hallucination	80.65
VGG-S + Low-rank	89.88
VGG-S + Hallucination + Low-rank	95.72
VGG-face	72.54
VGG-face + Colorization [43]	82.45
VGG-face + Hallucination	83.10
VGG-face + Low-rank	82.26
VGG-face + Hallucination + Low-rank	91.01
COTS	83.84
COTS + Colorization [43]	90.18
COTS + Hallucination	93.02
COTS + Low-rank	91.83
COTS + Hallucination + Low-rank	**96.41**

Table 2. Cross-spectral rank-1 identification rate on CASIA NIR-VIS 2.0 (see text for protocol). We evaluate three pre-trained single-spectrum (VIS) DNN models: VGG-S, VGG-face, and COTS. This experiment shows the effectiveness of cross-spectrally hallucinating an NIR image input, or low-rank embedding the output (universally for all the tested DNNs). When both schemes are used together, we observe significant further improvements, e.g., 75.04% to 95.72% for the VGG-S model. The proposed framework gives state-of-the-art (96.41%) without touching at all the VIS recognition system.

We then learn a 1024-by-1024 low-rank transform matrix to align the two spectrums. Note that, for efficiency, the PCA and low-rank transform matrices can be merged. We use cosine similarity to perform the matching.

5.4. Results

We evaluate the performance gain introduced by cross-spectral hallucination and the low-rank transform, and by both techniques combined. As explained, our cross-spectral hallucination CNN requires more training data than the one available in the standard CASIA NIR-VIS 2.0 benchmark training set, so we define a new splitting of that dataset into 6 folds. We use the same protocol for VIS hallucination as for face recognition, i.e. five folds for training and one fold for testing. There is no subject overlap between our training and testing partitions. The evaluation is performed on the testing partition, using VIS images as the gallery and NIR images as the probe.

In Table 2 we report the rank-1 performance score of the single-spectrum DNN, with and without hallucination and with and without the low-rank embedding for the three face models. The corresponding ROC curves are shown in Figure 7. The results show that it is often equally effective to hallucinate a VIS image from the NIR input, or to low-rank embed the output. Both schemes independently introduce a significant gain in performance with respect to using single-spectrum DNNs. When hallucination and low-rank

	Accuracy (%)
Jin et al. [17]	75.70 ± 2.50
Juefei-Xu et al. [18]	78.46 ± 1.67
Lu et al. [27]	81.80 ± 2.30
Saxena et al. [32]	85.90 ± 0.90
Yi et al. [39]	86.16 ± 0.98
Liu et al. [26]	**95.74 ± 0.52**
VGG-S	57.53 ± 2.31
VGG-face	66.97 ± 1.62
COTS	79.29 ± 1.54
VGG-S + Triplet	67.13 ± 3.01
VGG-face + Triplet	75.96 ± 2.90
COTS + Triplet	84.91 ± 3.32
VGG-S + Low-rank	82.07 ± 1.27
VGG-face + Low-rank	80.69 ± 1.02
COTS + Low-rank	*89.59 ± 0.89*

Table 3. Cross-spectral rank-1 identification rate on the 10-fold CASIA NIR-VIS 2.0 benchmark. The evaluation protocol defined in [24] is adopted. We evaluate three single-spectrum DNN models: VGG-S, VGG-face, and COTS. Single-spectrum DNNs perform significantly better for cross-spectral recognition by applying the proposed low-rank embedding at the output (universally for all the tested DNNs), which is a simple linear transform on the features. The popular triplet embedding [38] shows inferior to low-rank embedding for such a cross-spectrum task. Excluding [26], we report the best result on this largest VIS-NIR face benchmark. As discussed before, [26] tunes/adapts the network to the particular dataset, achieving results slightly below to those we report in Table 2 for our full system; results we obtain without any need for re-training and thereby showing the generalization power of the approach, enjoying the advantages of both existing and potentially new-coming VIS face recognition systems.

embedding are used together, we observe significant further improvements, e.g., from 75.04% to 95.72% for VGG-S. Using the COTS and the combination of hallucination and low-rank embedding, the proposed framework yields a state-of-the-art 96.41% rank-1 accuracy, without touching at all the VIS pre-trained DNN and with virtually no additional computational cost. We include the result of applying a state-of-the-art colorization method [43] instead of the proposed hallucination CNN. Note that [43] was trained on perfectly aligned images (whereas ours is not, Section 3.2); and it was trained for the grayscale to RGB visualization task and not the NIR to RGB for recognition task.

For completeness, we also present our results using the low-rank transform for the standard CASIA NIR-VIS 2.0 evaluation protocol in Table 3 (recall that the standard protocol is not possible for the added hallucination step). These results demonstrate that the low-rank embedding dramatically improves single-spectrum DNNs VIS-NIR rank-1 identification rate, 57.53% to 82.07% for VGG-S, 66.97% to 80.69% for VGG-face, and 79.29% to 89.59% for COTS. One of the most effective triplet embedding methods, LMNN [38], shows inferior performance to the proposed low-rank embedding for this cross-spectrum task.

The results obtained by our full system (Table 2) and our partial system (Table 3), indicate that the combination of hallucination and low-rank embedding produce state-of-the-art results on cross-spectral face recognition, without having to adapt or fine-tune an existing deep VIS model.

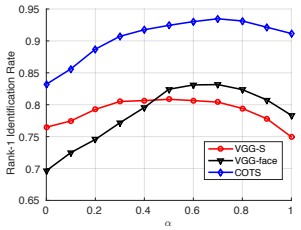


Figure 8. The impact of the blending parameter α in (1) on face recognition. We evaluate three single-spectrum DNN models: VGG-S, VGG-face, and COTS.

As discussed in Section 3.2, to preserve the details of the original NIR, we blend the hallucinated luminance of the CNN output with the original NIR image to remove possible artifacts introduced by the CNN. Figure 8 shows the impact of the blending parameter α in (1) on face recognition. The parameter $\alpha \in [0, 1]$ balances the amount of information retained from the NIR images and the information obtained with the hallucination CNN. We usually observe the peak recognition performance when α is around 0.6-0.7. With $\alpha = 0.6$ we also obtain a more natural-looking face; this is the value used in Table 2 and Figure 5.

6. Conclusion

We proposed an approach to adapt a pre-trained state-of-the-art DNN, which has seen only VIS face images, to generate discriminative features for both VIS and NIR face images, without retraining the DNN. Our approach consists of two core components, cross-spectral hallucination and low-rank embedding, to adapt DNN inputs and outputs respectively for cross-spectral recognition. Cross-spectral hallucination preprocesses the NIR image using a CNN that performs a cross-spectral conversion of the NIR image into the VIS spectrum. Low-rank embedding restores a low-rank structure for cross-spectral features from the same subject, while enforcing a maximally separated structure for different subjects. We observe significant improvement in cross-spectral face recognition with the proposed approach. This new approach can be considered a new direction in the intersection of transfer learning and joint embedding.

Acknowledgments

Work partially supported by ONR, NGA, ARO, NSF, ANII Grant PD_NAC_2015_1_108550. José Lezama performed part of this work while at Duke University.

References

[1] C. A. Aguilera, F. J. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo. Learning cross-spectral similarity measures with deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2016. 4

[2] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *European Conference on Computer Vision (ECCV)*, pages 469–481. Springer, 2004. 2

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

[4] M. Brown and S. Süsstrunk. Multispectral SIFT for scene category recognition. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR11)*, pages 177–184, Colorado Springs, June 2011. 2

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

[6] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference (BMVC)*, 2014. 6

[7] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, Corvalis, Oregon, USA, June 2007. 2, 6

[8] T. I. Dhamecha, P. Sharma, R. Singh, and M. Vatsa. On effectiveness of histogram of oriented gradient features for visible to near infrared face matching. In *ICPR*, pages 1788–1793, 2014. 2

[9] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision*, pages 391–407. Springer, 2016. 3

[10] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002. 5

[11] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2960–2967, 2013. 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR15)*, pages 1026–1034, 2015. 3

[13] J. Hoffman, S. Gupta, and T. Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 826–834, 2016. 2

[14] M. A. Hogervorst and A. Toet. Fast natural color mapping for night-time imagery. *Information Fusion*, 11(2):69–77, 2010. 2

[15] C.-A. Hou, M.-C. Yang, and Y.-C. F. Wang. Domain adaptive self-taught learning for heterogeneous face recognition. In *ICPR*, pages 3068–3073, 2014. 2

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 1

[17] Y. Jin, J. Lu, and Q. Ruan. Large margin coupled feature learning for cross-modal face recognition. In *2015 International Conference on Biometrics (ICB)*, pages 286–292, 2015. 2, 8

[18] F. Juefei-Xu, D. K. Pal, and M. Savvides. NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 141–150, 2015. 1, 2, 3, 8

[19] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR14)*, 2014. 3

[20] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[21] B. Klare and A. K. Jain. Heterogeneous face recognition: Matching nir to visible light images. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1513–1516. IEEE, 2010. 2

[22] Z. Lei and S. Z. Li. Coupled spectral regression for matching heterogeneous faces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR09).*, pages 1123–1128. IEEE, 2009. 2

[23] J. Li, P. Hao, C. Zhang, and M. Dou. Hallucinating faces from thermal infrared images. In *2008 15th IEEE International Conference on Image Processing (ICIP)*, pages 465–468. IEEE, 2008. 2

[24] S. Z. Li, D. Yi, Z. Lei, and S. Liao. The CASIA NIR-VIS 2.0 face database. In *9th IEEE Workshop on Perception Beyond the Visible Spectrum (PBVS, in conjunction with CVPR 2013)*, June 2013. 3, 6, 8

[25] S. Liu, D. Yi, Z. Lei, and S. Z. Li. Heterogeneous face image matching using multi-scale features. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 79–84. IEEE, 2012. 2

[26] X. Liu, L. Song, X. Wu, and T. Tan. Transferring deep representation for nir-vis heterogeneous face recognition. In *2016 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2016. 2, 8

[27] J. Lu, V. E. Liong, X. Zhou, and J. Zhou. Learning compact binary face descriptor for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):2041–2056, 2015. 2, 8

[28] A. Mignon and F. Jurie. CMML: a new metric learning approach for cross modal matching. In *Asian Conference on Computer Vision*, pages 14–pages, 2012. 2

[29] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference (BMVC)*, 2015. 1, 2, 6

[30] Q. Qiu and G. Sapiro. Learning transformations for clustering and classification. *Journal of Machine Learning Research*, 16:187–225, 2015. 2, 5

[31] M. S. Sarfraz and R. Stiefelhagen. Deep perceptual mapping for thermal to visible face recognition. *arXiv preprint arXiv:1507.02879*, 2015. 2

[32] S. Saxena and J. Verbeek. Heterogeneous face recognition with CNNs. In *ECCV TASK-CV 2016 Workshops*, 2016. 1, 2, 6, 8

[33] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR15)*, pages 815–823, 2015. 1, 2, 6

[34] B. K. Sriperumbudur and G. R. G. Lanckriet. A proof of convergence of the concave-convex procedure using zangwill's theory. *Neural Computation*, 24(6):1391–1407, 2012. 5

[35] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014. 6

[36] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR14)*, June 2014. 1

[37] A. Toet and M. A. Hogervorst. Progress in color night vision. *Optical Engineering*, 51(1):010901–010901, 2012. 2

[38] K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009. 2, 6, 8

[39] D. Yi, Z. Lei, and S. Z. Li. Shared representation learning for heterogeneous face recognition. In *International Conference on Automatic Face and Gesture Recognition (2015)*, 2015. 2, 8

[40] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 1, 7

[41] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li. Face matching between near infrared and visible light images. In *International Conference on Biometrics (ICB)*, pages 523–530. Springer, 2007. 2

[42] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 4:915–936, 2003. 5

[43] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016. 7, 8

[44] Y. Zheng and E. A. Essock. A local-coloring method for night-vision colorization utilizing image analysis and fusion. *Information Fusion*, 9(2):186–199, 2008. 2

[45] J.-Y. Zhu, W.-S. Zheng, J.-H. Lai, and S. Z. Li. Matching NIR face to VIS face using transduction. *IEEE Transactions on Information Forensics and Security*, 9(3):501–514, 2014. 2