Received July 31, 2015, accepted September 5, 2015, date of publication September 17, 2015, date of current version September 30, 2015.

Digital Object Identifier 10.1109/ACCESS.2015.2479620

Coupled Auto-Associative Neural Networks for Heterogeneous Face Recognition

BENJAMIN S. RIGGAN[1], (Member, IEEE), CHRISTOPHER REALE[2], (Student Member, IEEE), AND NASSER M. NASRABADI[1], (Fellow, IEEE)

[1]U.S. Army Research Laboratory, Adelphi, MD 20783, USA
[2]University of Maryland, College Park, MD 20742, USA

Corresponding author: B. S. Riggan (bsriggan@ncsu.edu)

This work was supported by the Army Research Laboratory and was accomplished through the Cooperative Agreement under Grant W911NF-12-2-0019.

ABSTRACT Several models have been previously suggested for learning correlated representations between source and target modalities. In this paper, we propose a novel coupled autoassociative neural network for learning a target-to-source image representation for heterogenous face recognition. This coupled network is unique, because a cross-modal transformation is learned by forcing the hidden units (latent features) of two neural networks to be as similar as possible, while simultaneously preserving information from the input. The effectiveness of this model is demonstrated using multiple existing heterogeneous face recognition databases. Moreover, the empirical results show that the learned image representation—common latent features—by the coupled auto-associative produces competitive cross-modal face recognition results. These results are obtained by training a softmax classifier using only the latent features from the source domain and testing using only the latent features from the target domain.

INDEX TERMS Cross-modality, heterogeneous face recognition, common latent features, biometrics, neural networks.

I. INTRODUCTION

Coupled auto-associative neural networks, or coupled autoencoders (CpAEs), are introduced and studied in the context of heterogeneous face recognition, which is face recognition when using different imaging modalities, e.g., visible, near infrared (NIR), shortwave infrared (SWIR), midwave infrared (MWIR), longwave infrared (LWIR), or artistic/forensic sketches [3], [4], [13], [45]. The primary objective of this work is to extract common latent features between two domains (e.g., visible-NIR, visible-SWIR, etc.), so that a classifier trained using one modality may generalize well to the other.

Although the concept of extracting common features for face recognition, itself, is not new [6], its extension to the heterogeneous face recognition problem is relatively novel. Moreover, the method used to extract common features, namely coupling of latent features (hidden units) using neural networks, to our knowledge has never been used for heterogeneous face recognition. Other models (Fig. 1) that extract common latent features have been used in applications using heterogeneous domains, such as audio-to-video speech recognition [24], cross-view object recognition [44], and even video-to-text description/recognition [8].

Coupled neural networks, such as CpAEs, are motivated by the work on multi-modal [37] and cross-modal learning [2], and more recently by the application of bilevel coupled dictionary learning to cross-modal face recognition [31], [32]. The primary difference between multi-modal and cross-modal applications is availability of every modality at test time. Both multi-modal and cross-modal learning use available training data from all modalities (source and target) to train the model. However, when validating and/or testing the model, cross-modal applications do not have data available from all modalities (only target). For example, in thermal-to-visible face recognition application there exists a gallery of visible images of *wanted persons* and a thermal probe image. The goal in this application is to identify the person in the thermal image (given only the probe image). Therefore, a classifier trained in the context of cross-modal applications is usually more difficult to generalize than in the multi-modal setting (i.e., fusion). Therefore, in our work on heterogeneous face recognition, we focus on the more difficult case of cross-modal recognition.

Cross-modal (heterogeneous) face recognition, to some extent, may be considered similar to two recent topics in machine learning literature: domain adaptation [35] and

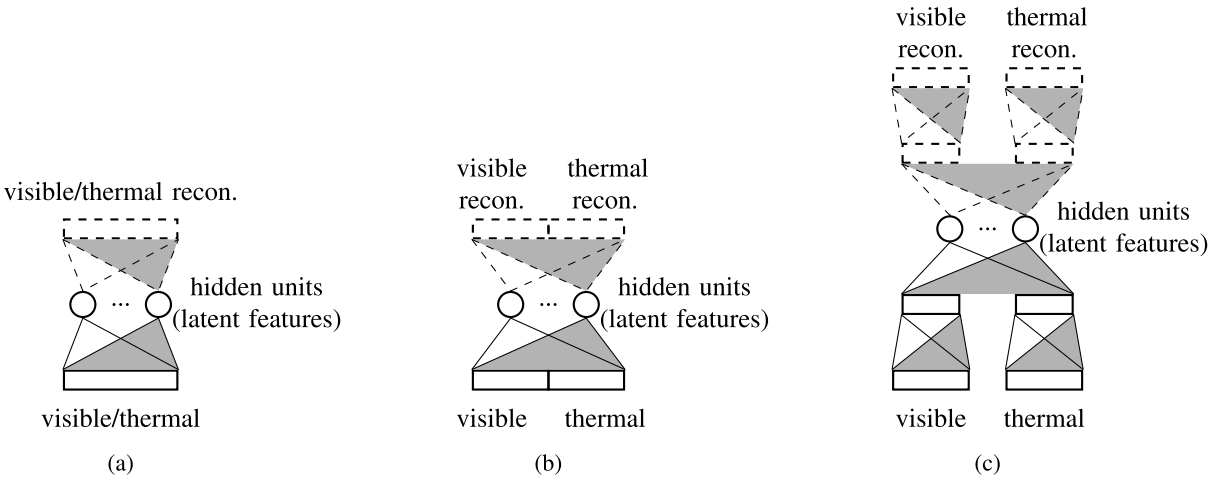

FIGURE 1. **Common cross-modality models: (a) encoder-decoder (autoencoder) with decoder indicated by dotted lines, (b) shared representation, and (c) Deep shared representation.**

self-taught learning [28]. Domain adaptation utilizes labeled training data in one domain (visible) to learn informative features in another domain (IR) [27]. Typically, problems that use domain adaption assume that classes within each domain form compact, but separable, clusters, and between domains corresponding classes are sufficiently "close" by some metric. Self-taught learning uses unlabeled data, such as random images, to learn latent features that are applied to the labeled data for classification. The principal idea concept behind self-taught learning is use of unlabeled data which belongs to dramatically different classes than that of the labeled data.

In contrast to domain adaptation and self-taught learning, clusters from unlabeled data (target domain) and labeled data (source domain) are not assumed to be "close" in the traditional sense due to significant differences in imaging characteristics. Instead, we take advantage of the structural similarities between faces from different imaging modalities and learn a common latent feature representation using CpAEs. The underlying assumption is that there exists a nonlinear common subspace in which corresponding faces belonging to the same individual in each of the domains have almost identical representations. Thus, a classifier trained in one domain should sufficiently classify individuals from a different domain. In some ways this can be considered a type of transfer learning [28], except we do not train separate transfer parameters for coupling.

A. BRIEF HISTORY

Cross-modal face recognition has been studied using feature extraction techniques [4] and partial least squares (PLS) [5], [9]. These studies apply classical pattern recognition tools to the thermal-to-visible face recogntion problem, which yields surprisingly good recognition results. The main idea is to exploit structural similarities between visible and thermal IR facial imagery. The main advantage of thermal imaging sensors is robustness in low lighting conditions where visible light cameras cannot operate effectively. This application is especially useful for personnel from law enforcement, military, and intelligence communities to identify suspects, missing persons, or persons of interest at night.

In recent years, recognition systems have turned away from the classical methods that use hand-crafted features. Instead, modern machine learning models, like autoencoders (AEs) [14] and dictionary learning (DL) [25], [33], are commonly used to learn descriptive features to improve classification rates. Even more recently, there has been a growing interest in "end-to-end" learning machines, where both features and classifiers are learned jointly, such as fully convolutional networks [21]. However, these methods have only been shown to be successful when using "very deep" models in applications that have an overabundance of training data available. In heterogeneous face recognition, the amount of corresponding images from two different domains is lacking due to the time and resources needed to acquire such data.

Face recognition in the visible domain has received a considerable amount of attention, particularly in the area of deep learning. Deep learning has significantly improved performance in face recognition to near human-levels [39]. Deep learning models, such as deep belief nets (DBNs), convolutional neural networks (CNNs), and deep autoencoders (AEs), are typically robust to partial occlusions, pose variations, and facial expressions. Deep learning methods also have been successfully applied to multi-modal and cross-modal recognition applications, such as audio-video speech recognition [24] and cross-view recognition [44].

Some of the more recent work in the application of cross-modal face recognition apply some of the machine learning techniques outlined previously. Our method is influenced by the method in [31], which applies coupled DL to the problem of cross-modal face recognition. In [32], the kernel trick is applied to coupled dictionaries for learning nonlinear latent

features for thermal-to-visible face recognition. Although the coupled DL approach demonstrated some success, the model is shallow. While deeper models, to our knowledge, have not been widely used for cross-modal face recognition, coupled deep Boltzman machines (DBMs) [24] have been used in multi-modal applications like audio-video recognition. Wang et. al [44] use CpAEs that are trained using a max-margin criterion, which explicitly utilizes label information from both views, for cross-view recognition in the visible domain.

The CpAEs that are proposed in this work differ from many of the previous works because two deep networks are trained jointly to learn common latent features that are useful for cross-modal face recognition (and possibly other applications).

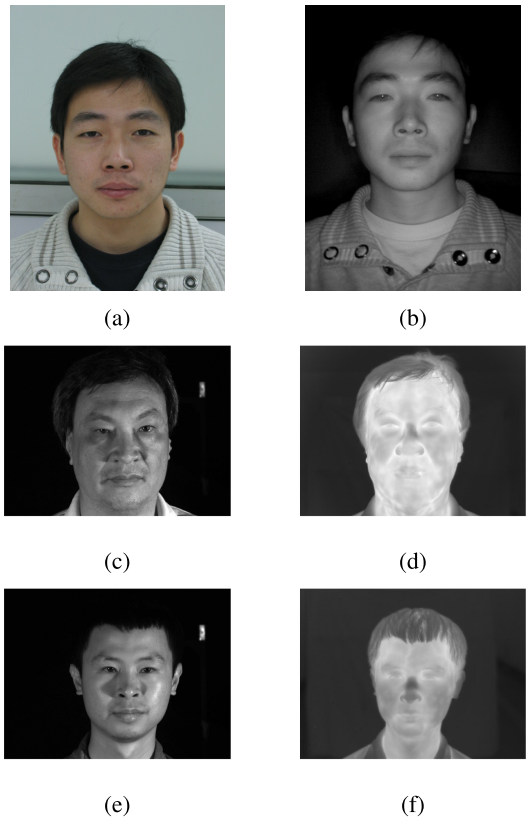

FIGURE 2. Examples of (a) visible and (b) NIR, (c) visible and (d) MWIR, and (e) visible and (f) LWIR imagery.

B. TECHNICAL CHALLENGES

There are several challenges in cross-modal face recognition, including: geometrical, phenomenological, and practical challenges. Similar to visible face recognition, a cross-modal recognition method must be robust, not only to 2D transformations (e.g., translation and image rotation), but also 3D transformations (e.g., pose). Furthermore, the apparent differences between reflective sources and emissive sources are relatively large (even by human standards), as shown in Fig. 2. Other practical challenges encountered include misalignment between domains (for corresponding subjects), varying facial expression and lighting conditions, and local or global varying temperatures (IR only).

C. CONTRIBUTIONS AND OUTLINE

We make three major contributions in this work:

1) We propose a coupled CpAE model for heterogeneous face recognition.
2) We develop a novel coupling objective to learn the model parameters for extracting common latent features.
3) We present a thorough evaluation using multiple datasets: Wright State (WSRI), Notre Dame X1 (UND X1), Night Vision (NVESD), and Casia NIR-VIS 2.0 (Casia).

II. BACKGROUND

Over the past decade, deep learning methods have achieved state-of-the-art recognition results (see [34]). Deep models like deep belief networks (DBNs) [29], CNNs [46], and stacked AEs [14], [41] have been used to solve computer visions problems such as face recognition [38], object detection [30], hand-writing/document recognition [17], and even human action recognition [11] in video segments. The common objective among these applications is to learn a robust set of features for discriminating different classes. Features learned through deep hierarchical networks have been shown to generalize better than most manually selected features. Since deep learning has demonstrated such exemplary performance in object and face recognition, we utilize the generalizability of deep architectures to learn cross-domain mappings.

Next, we briefly review various coupled learning models, particularly those that have been applied to cross-modal face recognition.

A. COUPLED LEARNING

Recently, coupled learning has been used to generalize features between multiple modalities. Yang et al. [47] proposed a coupled dictionary learning approach for image super-resolution, where the sparse codes are forced to be equal. In [43], the constraint that the codes must be equal was relaxed, so that the sparse codes between domains are allowed to be related by a linear mapping. These methods are primarily applied to the problem of image super-resolution. In [31] bilevel coupled dictionary learning was applied to the problem of face recognition. In [32], the kernel trick is applied to bilevel coupled dictionaries for learning nonlinear latent features for thermal-to-visible face recognition. Although this approach demonstrated some success, the model is shallow and relies on hand-extracted histogram of oriented gradient (HOG) [7] features. A deeper model may be capable of learning better features.

Learning useful features for classification from unlabeled data is a concept dating back to [26], which demonstrated the ability to learn edge filters from unlabeled images.

Recently, this concept has been extended to deeper models that are able learn more complex features. For example, in [18] stacked RBMs are used to learn more complex representations from unlabeled data, and [19] used convolutional DBNs for generating object parts (including faces). Le et al. [16] applied this concept using a deep AE model with local receptive fields and pooling in order to extract meaningful deep features for face detection.

In our work, we utilize a similar concept for cross-modal face recognition. We demonstrate that CpAEs can learn deeply coupled features for face recognition. The primary difference that separates our work from others is the way that common latent features are learned.

III. COUPLED AUTOENCODERS

The proposed CpAE approach for heterogeneous face recognition has 3 main components:

1) preprocessing
2) coupling (or domain adaptation)
3) classifiction.

The data preprocessing phase is a procedure that all images (training and testing) undergo in order to reduce significant variations between different modalities. The principle concept behind feature coupling is depicted in Fig. 3, where the goal of this step is to map corresponding source and target domain images to common latent subspace in a way that enables cross-modal classification. In this work, a CpAE-based model is trained to extract common latent features between source and target modalities. The final step, classification, involves training an *off-the-shelf* multi-class classifier, such as k-nearest neighbors, one-vs-all support vector machine (SVM), or softmax classifier.

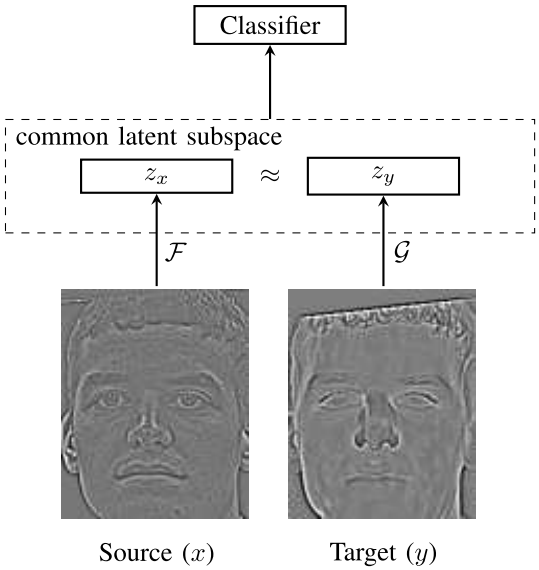

FIGURE 3. A corresponding source/target image pair are mapped to a common features space so that face recognition may be performed using images from the either source or target domain. The cross-modal face recognition scenario—trained with source domain and tested with target domain—is the most realistic application of this framework.

A. PREPROCESSING

The preprocessing phase in face recognition applications commonly include alignment, cropping, and filtering methods, and sometimes hand-crafted feature extraction methods, e.g., SIFT [22] or HOG [7] (as in [9], [31], and [32]). The diagram in Fig. 4 depicts a typical preprocessing subroutine that is applied prior to any optimization/recognition algorithms. However, the actual preprocessing methods used is very dependent on the imaging modalities (visible, NIR, thermal, or sketch imagery) and the variations (pose, lighting, temperature, etc.) within the dataset.

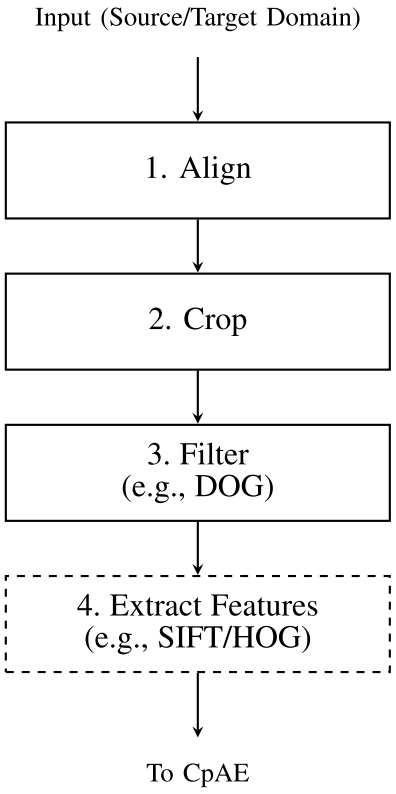

FIGURE 4. Preprocessing both source and target domain typically includes: alignment of corresponding source and domain inputs, cropping images around the face, and filtering. Some methods will reduce the gap between two domains further using hand-crafted feature extraction. In our work, we choose leave this step out, but we show it in the diagram above for completeness.

Face alignment is usually applied by using fiducial points (e.g., eyes, nose, mouth, etc), which are either manually located or automatically detected, to determine an affine transform that converts corresponding images to canonical coordinates. This produces a resulting set of images from different modalities which are geometrically aligned.

After alignment, all the images are cropped around each face.

Next, filtering methods such as Difference of Gaussians (DOG) filtering is applied. DOG filtering in known to reduce different variations within each modality. For example, in the visible domain local illumination variations are reduces [40], and in the thermal domain temperature variations are reduced. The purpose for DOG filtering is

to reduce the apparent difference between each imaging modality.

Notice that the images in Fig. 3 have already been preprocessed.

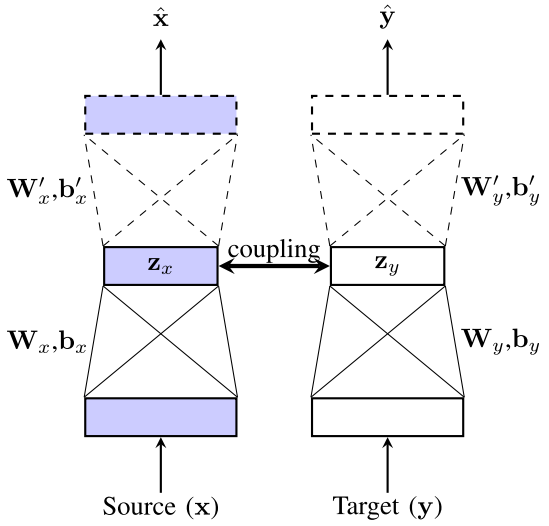

FIGURE 5. A CpAE is a pair of AEs where the hidden units (latent features) are coupled. The latent features, $\mathbf{z}_x$ and $\mathbf{z}_y$, are computed from the source and domain inputs, $\mathbf{x}$ and $\mathbf{y}$, and the encoder parameters: $\mathbf{W}_x$, $\mathbf{b}_x$ and $\mathbf{W}_y$, $\mathbf{b}_y$. Additionally, source and domain reconstructions, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, are computed using the latent features and decoder parameters: $\mathbf{W}'_x$, $\mathbf{b}'_x$ and $\mathbf{W}'_y$, $\mathbf{b}'_y$.

B. DEFINITION

Consider a CpAE to be defined as a pair of encoders and decoders (Fig. 5). Let the source and target input images be denoted by $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^d$ respectively. Given $\mathbf{x}$ and $\mathbf{y}$, the corresponding latent features are given as the output of the following encoders:

$$\mathbf{z}_x = f(\mathbf{b}_x + \mathbf{W}_x\mathbf{x}), \tag{1}$$

and

$$\mathbf{z}_y = f(\mathbf{b}_y + \mathbf{W}_y\mathbf{y}), \tag{2}$$

where $\mathbf{W}_x, \mathbf{W}_y \in \mathbb{R}^{k\times d}$ and $\mathbf{b}_x, \mathbf{b}_y \in \mathbb{R}^k$ are the encoder parameters (i.e., weights and biases), and $f(\cdot)$ is the nonlinear activation function. In this work, the activation function is $f(u) = (1 + \exp(-u))^{-1}$. The decoders, which are defined as

$$\hat{\mathbf{x}} = f(\mathbf{b}'_x + \mathbf{W}'_x\mathbf{z}_x), \tag{3}$$

and

$$\hat{\mathbf{y}} = f(\mathbf{b}'_y + \mathbf{W}'_y\mathbf{z}_y), \tag{4}$$

are used in the formulation of the objectives for training a CpAE. The decoder weights $\mathbf{W}'_x, \mathbf{W}'_y \in \mathbb{R}^{d\times k}$ in this work are not explicitly tied to the encoder weights, meaning $\mathbf{W}_x^T \neq \mathbf{W}'_x$ and $\mathbf{W}_y^T \neq \mathbf{W}'_y$. We did not observe any substantial benefits from using tied weights.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ represent the corresponding sets of faces from the visible and thermal domains, where each image pair $(\mathbf{x}_i, \mathbf{y}_i)$ is given a subject label $c_i \in \{1, \ldots, N\}$.

The two types of errors that we are interested in minimizing over the training data ($\mathbf{X}$ and $\mathbf{Y}$) are the reconstruction error and the coupling error. The reconstruction error is given by $R_x(\Omega_x) = \|\mathbf{X} - \hat{\mathbf{X}}(\Omega_x)\|_F^2 = \sum_i \|\mathbf{x_i} - \hat{\mathbf{x}}_\mathbf{i}(\Omega_x)\|^2$ for the source domain, and $R_y(\Omega_y) = \|\mathbf{Y} - \hat{\mathbf{Y}}(\Omega_y)\|_F^2 = \sum_i \|\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}(\Omega_y)\|^2$ for the target domain. Note that Ω_x and Ω_y are compact notations for the model parameters (weights and biases) for source and target domains, respectively. Then, the total reconstruction error is given as

$$R(\Omega_x, \Omega_y) = R_x(\Omega_x) + R_y(\Omega_y). \tag{5}$$

The coupling error is the difference between the visible and thermal latent features, $\mathbf{z}_x$ and $\mathbf{z}_y$,

$$C(\Omega_x, \Omega_y) = \sum_i \|\mathbf{z}_{xi} - \mathbf{z}_{yi}\|^2. \tag{6}$$

Minimizing the reconstruction error (5) attempts to learn the latent feature representations $\mathbf{z}_x$ and $\mathbf{z}_x$ that are sufficiently representative of the inputs. Whereas, minimizing (6) forces the the latent features to become more similar. In principle, we want the $\mathbf{z}_x \approx \mathbf{z}_y$ which can be learned by only minimizing (6), but this most likely will lead to a degenerate solution. Therefore, combining the reconstruction error (5), the coupling constraint (6), and a sparsity regularization term yields the following objective function to be minimized when training a CpAE

$$J(\Omega_x, \Omega_y) = R(\Omega_x, \Omega_y) + \gamma C(\Omega_x, \Omega_y) + \lambda S(\Omega_x, \Omega_y). \tag{7}$$

Similar to [31] and [47], we want to encourage a sparse solution to prevent the model from over-fitting and to make the coupling process easier. Therefore, the objectives are regularized using $S(\cdot)$. The regularization function, $S(\cdot)$, used is the Kullback-Leibler divergence (KL-divergence) [15] criteria given by $\sum_j KL(\rho||\hat{\rho}_j)$, where

$$KL(\rho||\hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}. \tag{8}$$

The user-defined parameter ρ represents the desired mean activation (over the training set) for every hidden unit, and $\hat{\rho}_j$ is the average activation of the j^{th} hidden unit ($\mathbf{z}_x$ or $\mathbf{z}_y$). Also, the parameter λ in (7) controls how much KL-divergence affects the optimization.

C. OPTIMIZATION

The parameters of CpAEs are incrementally updated using backpropagation, which uses gradient descent to update model parameters. The gradients of the reconstruction error (5) with respect to source encoder parameters, $\mathbf{W}_x$, $\mathbf{b}_x$, and target encoder parameters, $\mathbf{W}_y$, $\mathbf{b}_y$, are

$$\nabla_{\mathbf{W}_x} R = \delta_x \mathbf{X}^T, \tag{9}$$
$$\nabla_{\mathbf{b}_x} R = \delta_x, \tag{10}$$
$$\nabla_{\mathbf{W}_y} R = \delta_y \mathbf{Y}^T, \tag{11}$$
$$\nabla_{\mathbf{b}_y} R = \delta_y. \tag{12}$$

Here, $\delta_x = ((\mathbf{W}'_x)^T(X - \hat{X})) \odot f'(\cdot)$ represents the local reconstruction error term for the source domain, and $\delta_y = ((\mathbf{W}'_y)^T(Y - \hat{Y})) \odot f'(\cdot)$ is the local reconstruction error term for the target domain.

Note that similar gradients exist for updating decoder parameters. Note that, in this paper, the element-wise multiplication operator represented by the $\odot$ symbol. In the absence of a $\odot$, matrix or scalar multiplication should be inferred from the context.

Similarly, the gradients of the coupling error (6) are computed with respect to source parameters (13) and (14) and target parameters (15) and (16). By incorporating these gradients into the descent algorithm, the latent feature representation between corresponding inputs from source and target domains become more similar.

$$\nabla_{\mathbf{W}_x} C = \bar{\delta}_x \mathbf{X}^T, \tag{13}$$

$$\nabla_{\mathbf{b}_x} C = \bar{\delta}_x, \tag{14}$$

$$\nabla_{\mathbf{W}_y} C = \bar{\delta}_y \mathbf{Y}^T, \tag{15}$$

$$\nabla_{\mathbf{b}_y} C = \bar{\delta}_y. \tag{16}$$

For the coupling objective, the local error terms are $\bar{\delta}_x = -(\mathbf{Z}_x - \mathbf{Z}_y) \odot f'(\cdot)$ and $\bar{\delta}_y = -(\mathbf{Z}_x - \mathbf{Z}_y) \odot f'(\cdot)$ for source and target domains, respectively.

Next, we need to incorporate the gradient of the sparsity term $S(\cdot)$ in order to properly constrain the solution. The gradients of the sparsity term source and target parameters are

$$\nabla_{\mathbf{W}_x} S = \left\{ \left(-\frac{\rho}{\hat{\rho}_x} + \frac{1-\rho}{1-\hat{\rho}_x} \right) \odot f'(\cdot) \right\} \mathbf{X}^T, \tag{17}$$

$$\nabla_{\mathbf{b}_x} S = \left(-\frac{\rho}{\hat{\rho}_x} + \frac{1-\rho}{1-\hat{\rho}_x} \right) \odot f'(\cdot), \tag{18}$$

$$\nabla_{\mathbf{W}_y} S = \left\{ \left(-\frac{\rho}{\hat{\rho}_y} + \frac{1-\rho}{1-\hat{\rho}_y} \right) \odot f'(\cdot) \right\} \mathbf{Y}^T, \tag{19}$$

$$\nabla_{\mathbf{b}_y} S = \left(-\frac{\rho}{\hat{\rho}_y} + \frac{1-\rho}{1-\hat{\rho}_y} \right) \odot f'(\cdot), \tag{20}$$

where $\hat{\rho}_x$, $\hat{\rho}_y$ are the average latent features over the training set, ρ is the desired response for every latent feature.

Therefore, combining the reconstruction gradients (9)–(12) with the coupling gradients (13)–(16) and the sparsity gradients (17)–(20) yields the gradients of the objective function (7), which are

$$\nabla_{\mathbf{W}_x} J = \nabla_{\mathbf{W}_x} R + \nabla_{\mathbf{W}_x} C + \nabla_{\mathbf{W}_x} S, \tag{21}$$

$$\nabla_{\mathbf{b}_x} J = \nabla_{\mathbf{b}_x} R + \nabla_{\mathbf{b}_x} C + \nabla_{\mathbf{b}_x} S, \tag{22}$$

$$\nabla_{\mathbf{W}_y} J = \nabla_{\mathbf{W}_y} R + \nabla_{\mathbf{W}_y} C + \nabla_{\mathbf{W}_y} S, \tag{23}$$

$$\nabla_{\mathbf{b}_y} J = \nabla_{\mathbf{b}_y} R + \nabla_{\mathbf{b}_y} C + \nabla_{\mathbf{b}_y} S. \tag{24}$$

These gradients are used to update the model parameters, as shown in Algorithm 1.

Algorithm 1 Stacked CpAE Training

```
Initialize x, y
for layer = 1 ... N do
    while not converged do
        W_x  ← W_x  + η∇_{W_x} J
        b_x  ← b_x  + η∇_{b_x} J
        W_y  ← W_y  + η∇_{W_y} J
        b_y  ← b_y  + η∇_{b_y} J
        W'_x ← W'_x + η∇_{W'_x} J
        b'_x ← b'_x + η∇_{b'_x} J
        W'_y ← W'_y + η∇_{W'_y} J
        b'_y ← b'_y + η∇_{b'_y} J
    end while
    x ← z_x (from layer)
    y ← z_y (from layer)
end for
Fine-Tune
```

D. STACKED COUPLED AUTOENCODERS

In recent years, research has shown that deep neural networks (DNNs) such as stacked AEs and RBMs are capable of learning hierarchies of latent features that provide robust and invariant high-level descriptors that produce state-of-the-art recognition results in single modality applications. Our aim is to learn deep, nonlinear common latent features for heterogeneous face recognition.

Stacking CpAEs is very similar to the way "regular" AEs are stacked, the input to the subsequent AE is the code (i.e., output from the hidden layer) from the previous AEs. Therefore, a stacked CpAE (shown in Fig. 6) may be viewed as a pair of stacked AEs, where at least one of the hidden layers is coupled (usually the last AE). The motivation for stacking CpAEs come from [42], which demonstrated that stacked AEs generally provide significantly better initialization for deep networks (like face recognition) resulting in

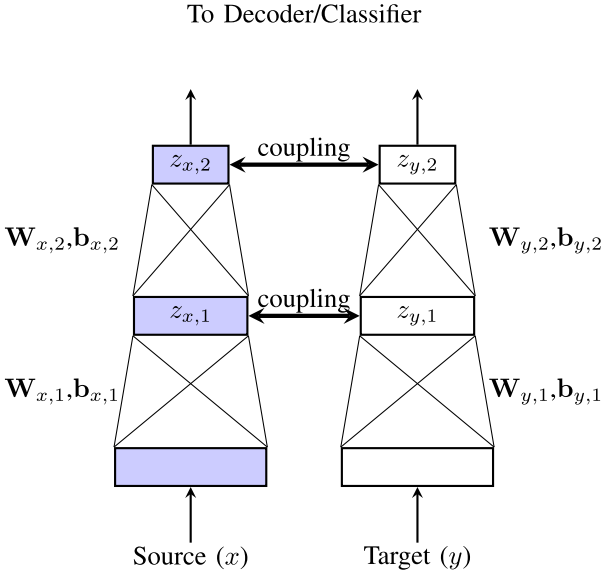

FIGURE 6. A stacked CpAE is a pair of stacked AEs with one (or more) coupled layers of hidden units. As shown, a subsequent CpAE is trained using the hidden units from the previous CpAE. For convenience, we have dropped the decoders.

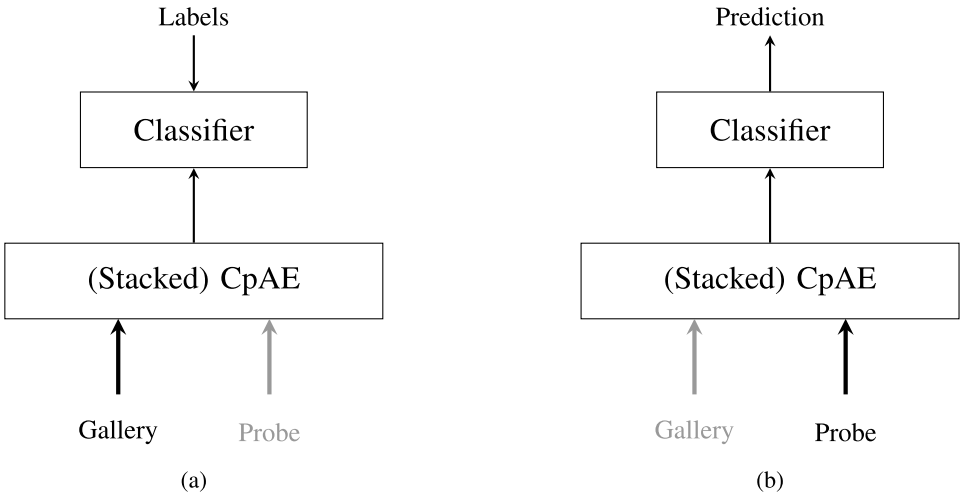

FIGURE 7. A classifier is (a) trained using the latent features from a pre-trained CpAE-based model given only the visible gallery images, and (b) tested using the latent features from the CpAE-based model given only probe IR images.

TABLE 1. Summary of heterogeneous face recognition databases used for comparing models. The Casia database, despite have the most correlation between source and target modalites, varies greatly in terms of pose (P), expression (E), glasses (G), and distance (D). Whereas, the WSRI and UND X1 databased primarily vary only in expression.

Database	Source	Target	# subjects	Variations
WSRI	visible	MWIR	64	E
UND X1	visible	LWIR	241	E
NVESD	visible	MWIR & LWIR	50	E,D
Casia[1]	visible	NIR	725	P,E,G,D

improved overall performance. Therefore, we stacked CpAEs so that a deeper, nonlinear representation may be learned in a layer-wise greedy fashion, so that the latent features between corresponding source and target domains are better correlated.

E. CLASSIFICATION

In this section, we discuss the proposed approach, depicted in Fig. 7 used for classifying an individual from a probe image from the target domain out of a gallery of images from the source domain. After training the CpAE model to extract the common latent features, we train a softmax classifier on the coupled gallery set (i.e., $\mathbf{z}_x$) and then test using the probe set (i.e., $\mathbf{z}_y$).

In practice, due to the limited amount of available data, the optimization problem does not generalize well when coupling entire image representations. Therefore, the common latent features are extracted from smaller patches (or windows) within the larger images. The primary advantage of processing the images at the patch-level rather than the image-level is reduced impact from the *curse of dimensionality* [1]. The impact is reduced in two ways: the dimensionality of input is reduced, and the number of samples is increased. Additionally, to achieve better overall performance we artificially expand each dataset by adding shifted/noisy versions of existing samples. We also utilize dropout [36] when training the classifier in order to improve generalization.

Given that the common latent features are extracted locally at the patch-level, the global image representations are formed by dividing the gallery and probe images into non-overlapping patches, and concatenating the latent features of every patch for every images. Note that the softmax classifier is trained using the image representations, namely the concatenated latent features, from the gallery. Then, similarly, the concatenated features for the probe images are used to obtain the recognition rates at test time.

IV. HETEROGENEOUS FACE RECOGNITION DATASETS

In order to evaluate the proposed CpAE based model, we use four different heterogeneous face recognition databases:

1) Wright State (WSRI)
2) Notre Dame X1 (UND X1)
3) Night Vision (NVESD)
4) Casia NIR-VIS 2.0 (Casia) [20]

to test NIR-to-visible, MWIR-to-visible, LWIR-to-visible, sketch-to-visible face recognition applications. Each of these databases is briefly described below, and an overview is provided in Table 1.

[1]The images are not aligned since images from each modality are not taken simultaneously.

A. WSRI

The WSRI dataset is composed of 1615 visible and 1615 MWIR images from 64 different subjects, which have been prepocessed (i.e., aligned, cropped, and DOG filtered) according to section III-A. There are approximately 25 images per subject (on average), where the subjects vary facial expression between image captures. The original resolutions of these images are 1004×1004 pixels for the visible domain and 512×640 pixels for the MWIR domain, but after preprocessing the images (visible and MWIR) are resampled to 235×295 pixels. This database is split into a random training set composed of 10 subjects and a testing set of the remaining 54 subjects.

B. UND X1

The UND X1 dataset is composed of 2451 visible and 2451 LWIR images from 241 different subjects. These images have also been preprocessed in a similar way as the WRSI dataset. The original resolutions of the images are 1600×1200 pixels for the visible domain and 320×240 pixels for the LWIR domain, but both are resampled to 150×110 pixels.

The UND X1 dataset is split into training and testing sets. The training set contains 159 subjects captured in the visible and LWIR domains, with only one image pair per subject. The testing set contains 82 different subjects with multiple images per subject. Note that this is very challenging dataset because of the low resolution and noise present in the LWIR imagery. In this case, there is significantly less correlation between the two domains.

C. NVESD

The NVESD database was collected by U.S. Army CERDEC-NVESD in 2012. The entire database contains visible, SWIR, MWIR, and LWIR images that were captured simultaneously from 50 different subjects. The original resolutions of the images, before preprocessing, are 640×480 pixels respectively. After preprocessing, as in [9], the images are aligned, filtered, cropped to 174×174 pixels, and split into training and testing sets.

D. CASIA

The Casia dataset consists of both visible and NIR images from 725 different subject, that were not captured concurrently. For each subject there are 1–22 visible images and 5-50 NIR images, each of which varies pose, expression, glasses, and distance to camera/sensor. The original image sizes (NIR and visible) are 640×480 pixels, and the cropped image sizes are 128×128 pixels. This databases provides a sub-collection of images for parameter tuning, and 10 different sub-collections for reporting experimental results.

V. EXPERIMENTS AND RESULTS

The experiments conducted for evaluating the CpAE model span four different datasets: WSRI, UND X1, Casia, and NVESD (as noted in Section IV). The CpAE is compared with other existing cross-modality models in terms cross-modal recognition rate (Sections V-A–V-C). Also, we show how this performance changes as the number of features, patch size, and sparsity parameters vary.

Before presenting quantitative results, we provide some qualitative results that visually demonstrate what the CpAE model has learned. Fig. 8 shows some reconstructed visible images from probe thermal images in order to demonstrate that a CpAE can effectively map inputs from one domain to the other. For purposes of visual comparison, we also show (see Fig. 8) the original visible and thermal images along with the visible-to-visible reconstructions and the thermal-to-visible reconstructions. As expected, the thermal-to-visible reconstructions are sufficiently similar to the original visible images. More importantly, the thermal-to-visible reconstructions are acceptable for visualizing the corresponding faces in the visible domain. However, reconstruction is only a surrogate objective for preserving information, the true objective is to sufficiently classify probe images in the test set. Therefore, we report the rank-1 cross-modal identification rates using the CpAE model and existing models on the different datasets.

FIGURE 8. Visible-to-visble reconstruction and thermal-to-visible reconstruction using CpAE. For each subject (a) and (b), there are four images shown: visible input (bottom left), thermal input (top left), visible-to-visible reconstruction (bottom right) and thermal-to-visible reconstruction (top right).

A. WSRI AND UND RESULTS

The goal of cross-modal face recognition is to be able to identify individuals given a single probe image. For each model (the baseline encoder (Enc) in Fig. 1(a), the shared representation (SR) in Fig. 1(b), the deep shared representation (DSR) in Fig. 1(c), and the stacked CpAE in Fig. 6, the training set is used to learn the optimal model parameters given corresponding visible and thermal images.

For the encoder model, two patch-based AEs with 25 hidden units, with non-overlapping patches of 10×10 pixels (or 100 inputs/outputs), are trained independently. One is trained using the visible imagery and the other using the thermal imagery. The share representation model is an AE with 25 hidden units where the input/output is the concatenation of corresponding visible and thermal patches

of 10 × 10 pixels. The deep shared representation model constructs two independent AEs (visible and thermal) with 25 hidden units which are trained in the same way as the encoder model. Then, the second layer of the deep shared representation model is trained by constructing a single AE with 25 units where the input is a concatenation of the hidden units from the visible and thermal encoders from the first layer. Lastly, the proposed CpAE model is a two layer network (as described in Section III-D) where the inputs are 10 × 10 non-overlapping patches (for each modality), the number of hidden units in the first layer are 50, and the number units in the second layer are 25 (100-50-25). This network is trained in a greedy layer-wise fashion as outlined in Algorithm 1.

Once the models are trained, the visible probe images are fed forward in order to obtain the latent features for every non-overlapping patch. The latent features from patches are concatenated to form the overall image representations. Then, a softmax classifier is trained only using the visible image representations. After the softmax classifier is trained, the latent features for every non-overlapping patch are obtain from the thermal probe images. Finally, the concatenated latent features are classified using the trained softmax classifier. The identification rate is computed as the ratio of the number of correctly classified individuals and the total number of subjects.

The identification rates for each model is reported in Table 2 using both the WSRI and UND X1 datasets. The reported performances demonstrate that the CpAE model effectively learns common latent subspace in which visible and IR image representations are sufficiently similar.

TABLE 2. Comparison of CpAE with existing cross-modality neural nets.

Method	WSRI	UND X1
Enc	75.4%	20.9%
SR	88.1%	23.2%
DSR	91.5%	40.2%
CpAE	**97.2%**	**51.9%**

Additionally, we compared the performances achieved by the CpAE model with PLS, (linear) bilevel coupled dictionary learning (BCDL) [31], and Kernel bilevel coupled dictionary learning (K-BCDL) [32]. The tabulated results (Table 3) again show improved performance.

B. NVESD RESULTS

Next, we evaluated our proposed CpAE model with the results reported on the NVESD database. For this experiment we trained a 256-200-100 CpAE using data from the training set (10 subjects) from source and target domains. Then, we report the rank-1 classification performance on the test sets (40 subjects) from target domain. In this case, we evaluate two types of heterogeneous face recognition: MWIR-to-visible and LWIR-to-visible. The compiled results

TABLE 3. Comparing CpAE with PLS and DL models on WSRI and UND X1 datasets.

Method	WSRI	UND X1
PLS [5][2]	83.7%	41.0%
BCDL [31]	93.1%	50.5%
K-BCDL [32]	95.9%	**52.0%**
CpAE	**97.2%**	**51.9%**

TABLE 4. Comparison of CpAE, PLS [10], and DL models on NVESD dataset.

Method	MWIR	LWIR
PLS	82.4%	70.4%
BCDL	90.7%	90.6%
K-BCDL	93.3%	92.5%
CpAE	94.4%	89.1%

in Table 4 show that the CpAE is better than PLS. Also, the CpAE results are competitive with the result for the BCDL approach.

TABLE 5. Comparison of CpAE with principle components analysis (PCA) and hetero-component analysis (HCA) methods [20] on the Casia dataset.

Method	Mean Acc.	Std. Dev.
PCA [20]	7.16%	0.52%
PCA + Sym [20]	9.26%	0.66%
PCA + Sym + HCA [20]	23.7%	1.89%
CpAE	33.1%	6.6%

C. CASIA RESULTS

In this experiment, we compared our CpAE model with the results reported in [20]. We trained the system using the Casia data provided (as described in Section IV-D). As outlined in [20], we report the mean accuracy and standard deviation over the 10 sub-experiments in Table 5. The reported performances on the Casia dataset can be improved significantly with better preprocessing (i.e., alignment [48]), compact binary codes (56.6% without LDA, 81.8% with labels) [23], or using labels for discrimination (71.5%) [12]. We simply trained the system in the same manner as with WSRI and UND X1 datasets, and reported the results. It can also be observed from Table 5 the performance variance reported for CpAEs is larger than other methods. This is primarily due to the fact that PCA and HCA are deterministic methods whereas ours method is stochastic, which depends greatly on initialization. This variance usually can be reduced by better initialization.

[2]The results reported here differ from those reported in [5] because the experiment was altered to align with the experiments in [31] and [32]. Note that the authors of [5] provided the code necessary for this comparison.

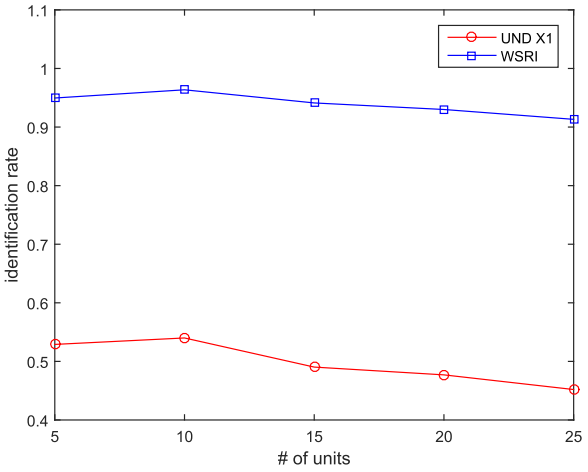

FIGURE 9. Performance with respect to number of features.

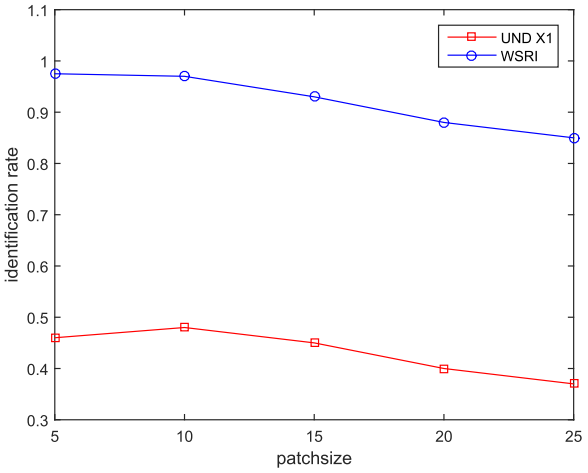

FIGURE 10. Performance with respect to patch size.

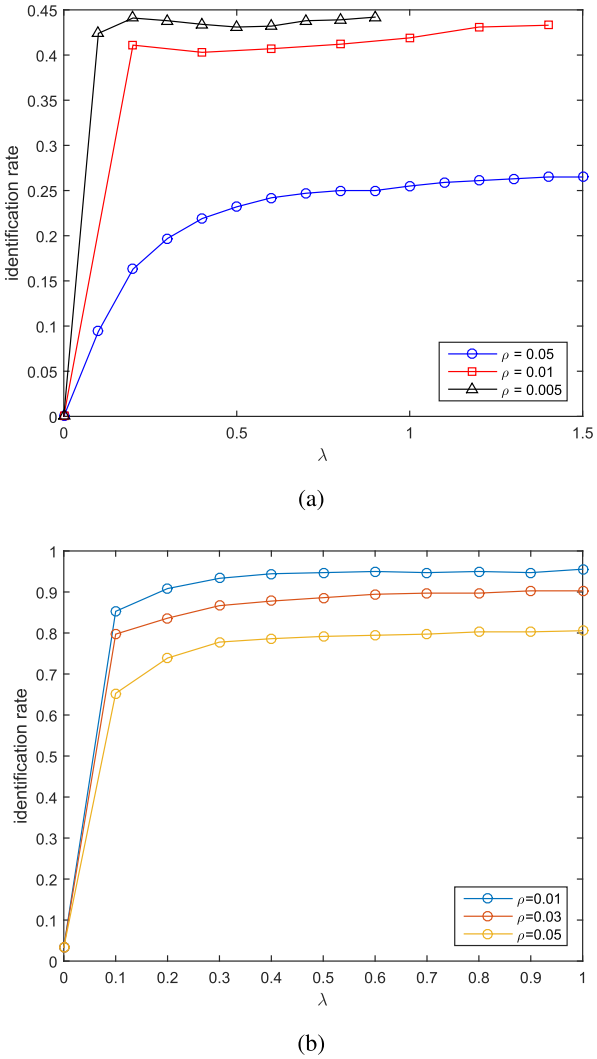

FIGURE 11. Performance with respect to sparsity parameters (ρ and λ) on (a) UND X1 dataset and (b) NVESD dataset (MWIR only).

D. NUMBER OF UNITS

In this experiment, we vary the number of hidden units being coupled between the thermal and visible image representations. As the number of units decreases, the coupling error becomes smaller. However, the reconstruction error increases. From the plot in Fig. 9, it is observed that performance degrades as the number of units increase. This is primarily due to the explicit coupling that takes place. As the number of units increases, the model is to able capture more of the subtle variations in the visible domain. However, the coupling process becomes more difficult because there are more units that are being coupled across domains. With fewer units, we tolerate some of the reconstruction error in order to improve the generalization of the thermal-to-visible coupling.

E. PATCH SIZE

Now, the patch size is varied in order to determine the effect that the size of the patch (i.e., size of the inputs) has on recognition rate. It is well known that AEs generalize better when using patches compared to using entire images. When using images, rather than patches, the increase in the number of model parameters that must be learned and insufficient training data yields poor results. The plot in Fig. 10 shows how changing the patch affects the cross-modal identification rate. As we can see, similar to increasing the number of units, increasing the patch size degrades the performance of CpAEs. As the patch size increases, the pixel variance within a single patch increases. Although the model becomes more complex (i.e.,, more free parameters), it does not maintain the same level of performance. Note that one must also be careful not to make the patch size too small (e.g., 1 pixel) because small patch size may contain little (if any) structure and could be perceived fitting noise.

F. SPARSITY PARAMETER

The final experimental results reported in this paper demonstrate the effect that sparsity has on the cross-modal identification rate. In Section III-C, we stated the KL-divergence criteria is used to regularize the optimization. There are two parameters associated with the KL-divergence: a sparsity penalty (λ) and a sparsity target (ρ). Figs. 11(a) and 11(b)

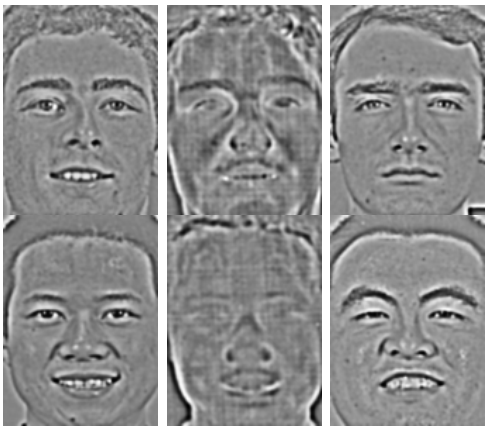

FIGURE 12. Corresponding visible (fist column) and thermal (second column) images for failure cases. Last column shows rank-1 classification.

show that a sparse solution can help improve performance on both UND X1 and NVESD datasets respectively. Note that one must be careful not to impose constraints that are too restrictive, leading to a trivial solution.

G. MISCLASSIFICATIONS

Despite the intuitive simplicity and acceptable performance of CpAEs, the difficulty of the problem means that there are many failures. These failures are fundamentally caused by the differences in the sensing modalities. As a matter of fact, if too many of the identifying features present in source domain are not able to be recovered from the target domain, then cross-modal recognition failure is expected. Fig. 12 shows a few difficult cases where the model was unable to correctly identify the individual. These examples greatly obscure identifying features, such as eyes, nose, and mouth. Also, as in standard face recognition, different facial expressions can be problematic for thermal-to-visible face recognition.

Here, we reiterate the difficulty of the problem at hand. Our CpAE is not trained on any of the images from the test set. Therefore, the test subjects have never been observed in thermal domain. Since the trained model has never learned to couple these unknown subjects, an occasional failure is to be expected.

VI. CONCLUSIONS

A novel method for coupling hidden units between two CpAEs was described for extracting nonlinear common latent features for heterogeneous face recognition. This idea, which utilizes fundamental concepts from machine learning and other cross-modality models, is intuitively simple due to the input-output nature of training neural networks. Additionally, as shown in Section V, the approach works well on a variety of heterogeneous face recognition databases.

The CpAE model was shown to achieve competitive performance on the WSRI and UND X1 databases, achieving near optimal performance on the former and a respectable 52% identification rate on the latter. Moreover, with little human intervention, we were able to achieve a reasonable level performance on the challenging Casia database. Given the difficulty of heterogeneous face recognition, there remains much more to learn regarding CpAEs. Additionally, we evaluated CpAEs on NVESD and Casia datasets. On the NVESD dataset, we achieved comparable level of performance with state-of-the-art for the MWIR-to-visible and LWIR-to-visible face recognition tasks. On the Casia dataset, we achieve a respectable 33% (beating reported PCA and HCA performances) considering that the faces are poorly aligned. Since CpAEs require that the faces between source and target domains should be well aligned, this result is not too surprising. This does seem to suggest, as with visible face recognition [39], that alignment between domains is critical for heterogeneous face recognition.

The bottom line in regards to using CpAEs is there must exist an underlying common subspace between source and target domains such that corresponding samples in that subspace produce similar codes. If this assumption does not hold then the optimization problem is ill-posed. Additionally, hyper-parameter selection, as with most neural network techniques, is critical for learning a good subspace for classification. In general, we found that patch-based training performed better over image-based training due to the limited amount of corresponding cross-domain data available for training our model. Other machine learning techniques like dataset expansion and dropout were also employed in order to reduce over-fitting and improve generalization.

This approach may very well extend to other cross-domain applications and other deep cross-modality models, like RBMs and CNNs. Further experimentation is required to fully comprehend the impact that the coupling objective introduced in this paper has on CpAEs and DNN models.

ACKNOWLEDGMENT

The authors would like to thank Sean Hu for helping provide access to the the UND X1, WSRI, and NVESD databases. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1957.

[2] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. NIPS*, 2007, pp. 153–160.

[3] T. Bourlai, N. Kalka, A. Ross, B. Cukic, and L. Hornak, "Cross-spectral face verification in the short wave infrared (SWIR) band," in *Proc. 20th ICPR*, Aug. 2010, pp. 1343–1347.

[4] T. Bourlai, A. Ross, C. Chen, and L. Hornak, "A study on using mid-wave infrared images for face recognition," *Proc. SPIE*, vol. 8371, pp. 83711K-1–83711K-13, May 2012.

[5] J. Choi, S. Hu, S. S. Young, and L. S. Davis, "Thermal to visible face recognition," *Proc. SPIE*, vol. 8371, pp. 83711L-1–83711L-10, May 2012.

[6] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2005, pp. 539–546.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2005, pp. 886–893.

[8] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. CVPR*, 2015, pp. 2625–2634.

[9] S. Hu, J. Choi, A. L. Chan, and W. R. Schwartz, "Thermal-to-visible face recognition using partial least squares," *J. Opt. Soc. Amer. A*, vol. 32, no. 3, pp. 431–442, 2015.

[10] S. Hu, N. Short, P. Gurram, K. Gurton, and C. Reale, "MWIR-to-visible and LWIR-to-visible face recognition using PLS and dictionary learning," *Face Recognit. Across EM Spectrum*, 2015.

[11] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2013.

[12] Y. Jin, J. Lu, and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 640–652, Mar. 2015.

[13] N. D. 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 *Proc. IJCB*, Oct. 2011, pp. 1–8.

[14] H. Larochelle, D. Erhan, and P. Vincent, "Deep learning using robust interdependent codes," in *Proc. Int. Conf. AISTATS*, 2009, pp. 312–319.

[15] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Proc. Adv. NIPS*, 2011, pp. 1017–1025.

[16] Q. V. Le *et al.*, "Building high-level features using large scale unsupervised learning," in *Proc. 29th ICML*, 2012, pp. 1–11.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[18] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Proc. Adv. NIPS*, 2008, pp. 873–880.

[19] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. 26th Annu. ICML*, 2009, pp. 609–616.

[20] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proc. IEEE Conf. CVPRW*, Jun. 2013, pp. 348–353.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, 2015, pp. 3431–3440.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[23] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2041–2056, Oct. 2015.

[24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th ICML*, 2011, pp. 689–696.

[25] H. V. Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Kernel dictionary learning," in *Proc. IEEE ICASSP*, Mar. 2012, pp. 2021–2024.

[26] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[27] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[28] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th ICML*, 2007, pp. 759–766.

[29] M. Ranzato, Y.-L. Boureau, and Y. LeCun, "Sparse feature learning for deep belief networks," in *Proc. Adv. NIPS*, 2007, pp. 1185–1192.

[30] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.

[31] C. Reale, N. M. Nasrabadi, and R. Chellappa, "Coupled dictionaries for thermal to visible face recognition," in *Proc. IEEE ICIP*, Oct. 2014, pp. 328–332.

[32] C. Reale, N. M. Nasrabadi, and R. Chellappa, "Kernel task-driven coupled dictionaries for thermal to visible face recognition," *Trans. Image Process.*

[33] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, Jun. 2010.

[34] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608014002135

[35] Y. Shi and F. Sha, "Information-theoretical learning of discriminative clusters for unsupervised domain adaptation," in *Proc. 29th ICML*, 2012, pp. 1079–1086.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jun. 2014.

[37] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. NIPS*, 2012, pp. 2222–2230.

[38] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 3476–3483.

[39] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 1701–1708.

[40] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.

[41] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th ICML*, 2008, pp. 1096–1103.

[42] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Mar. 2010.

[43] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 2216–2223.

[44] W. Wang, Z. Cui, H. Chang, S. Shan, and X. Chen. (Feb. 2014). "Deeply coupled auto-encoder networks for cross-view classification." [Online]. Available: http://arxiv.org/abs/1402.2031

[45] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.

[46] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Proc. Adv. NIPS*, 2014, pp. 1790–1798.

[47] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[48] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *Proc. 11th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2015, pp. 1–7.

BENJAMIN S. RIGGAN (S'12–M'14) received the B.S. degree in computer engineering from North Carolina State University, in 2009, and the M.S. and Ph.D. degrees in electrical engineering from North Carolina State University, in 2011 and 2014, respectively. He is currently a Post-Doctoral Fellow with the U.S. Army Research Laboratory. His research interests are in image and video processing, machine learning, and signal reconstruction.

CHRISTOPHER REALE (S'13) received the B.S. degree in electrical engineering from Washing University in St. Louis, in 2009. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Maryland, College Park. He is also a Student Intern with the U.S. Army Research Laboratory.

NASSER M. NASRABADI (S'80–M'84–SM'92–F'01) received the B.Sc. (Eng.) and Ph.D. degrees in electrical engineering from Imperial College London, London, U.K., in 1980 and 1984, respectively. In 1984, he was with IBM, U.K., as a Senior Programmer. From 1985 to 1986, he was with the Philips Research Laboratory, NY, as a member of the Technical Staff. From 1986 to 1991, he was an Assistant Professor with the Department of Electrical Engineering, Worcester Polytechnic Institute, Worcester, MA. From 1991 to 1996, he was an Associate Professor with the Department of Electrical and Computer Engineering, State University of New York at Buffalo, Buffalo, NY. Since 1996, he has been a Senior Research Scientist with the U.S. Army Research Laboratory (ARL) working on image processing and automatic target recognition. His current research interests include hyperspectral imaging, automatic target recognition, statistical machine learning theory, robotics, and neural networks applications to image processing. He is also a fellow of ARL and SPIE. He has served as an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS, SYSTEMS AND VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON NEURAL NETWORKS.