Sparse Low-Rank Fusion based Deep Features for Missing Modality Face Recognition

Ming Shao[1] Zhengming Ding[1] and Yun Fu[1,2]
[1] Department of Electrical and Computer Engineering, Northeastern University, USA
[2] College of Computer and Information Science, Northeastern University, USA

***Abstract*— Multi-modality data recently attract more and more research attention. In this paper, we concentrate on a very interesting problem—image classification with missing modality. Specifically, only images in one modality as well as a relevant auxiliary database are accessible during the training phase, which is significantly different from general image classification under the same modality. To this end, we propose a novel framework integrating multiple deep autoencoders with bagging strategy. For each autoencoder, we generate its input by randomly sampling data from other modality and the auxiliary database, and enforce its output to lie in a common feature space through Robust PCA. Finally, a novel sparse low-rank feature fusion approach is proposed in the test phase to integrate multiple features learned from different autoencoders, followed by a decision voting. Extensive experiments on two databases, i.e., BUAA-NIRVIS, Oulu-CASIA NIRVIS databases demonstrate the effectiveness of the proposed framework when there is only one modality available for training.**

I. INTRODUCTION

Multi-modal data are common in the era of multimedia. For a single object, people can record/describe in different ways—we can take a picture, shoot a video, or record the sound. Furthermore, some spectrals are well utilized to feature the object under different scenarios. For example, near infrared (NIR) images are able to provide better visual feature when visible light (VIS) condition is poor for recognition task [16]. There are two typical problems related to the multi-modal data. One is how to better utilize available data from different modalities for better performance, leading to the study of "feature fusion" [1]; the other is how to use one modality to assist the model training of another modality when labeled data in the second modality are few. In fact, this problem is similar to transfer learning [21], but manipulates between different modalities and passes on the knowledge.

The above feature fusion and transfer learning problems have been widely discussed, and assume that both modalities are complete and available during the training phase. However, there is often the case that one modality is missing in the training phase, and it is hard to build correspondences between two modalities. For example, for real-time recognition system, test data (target modality) are only available during the running time. Fortunately, we might recover the correspondence between modalities through an extra auxiliary database which contains similar modalities, although the content of the database is different from the original one. To make it clear, suppose we only have corresponding training samples in VIS, but our recognition task is in NIR where no data are available during the training. In the meanwhile, we have an extra NIR-VIS database with both modalities available. However, this extra database is usually captured under different environments, or by different devices. So, either the quality, or the resolution of images is different from our target database.

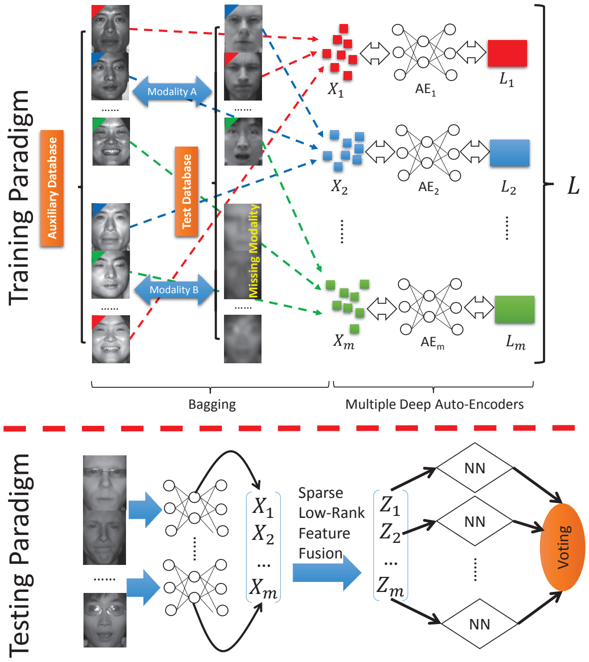


Fig. 1: Framework of the proposed method, which contains two modalities: "Modality-A" as VIS, "Modality-B" as NIR. It uses two databases: an auxiliary database with complete multi-modal data, a test/target database with missing modality. We first use the bagging strategy to sample data from both auxiliary and test databases, where a color (on the up-left corner of each face) represents a sampling. Then these sampled data can train m autoencoders $AE_{1\sim m}$ and yield m decisions which will be fused by a voting scheme. Note X_i and L_i represent a sampled dataset, and its low-rank recovery, respectively in the training phase. X_i and Z_i are deep features, and the new representation of X_i after sparse low-rank feature fusion in the test phase, respectively.

The problem discussed above was formally introduced in [9] as `missing modality problem`. In this paper,

This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, U.S. Army Research Office Young Investigator Award W911NF-14-1-0218, and Air Force Office of Scientific Research award FA9550-12-1-0201.

978-1-4799-6026-2/15/$31.00 ©2015 IEEE

we present a novel approach towards this problem, and the entire framework is illustrated in Figure 1. Our method takes advantage of one auxiliary database as well as the relevant modality in the target database to address the missing modality problem. Essentially, our method targets at simultaneous feature learning in two directions: (1) feature learning between target and auxiliary databases; (2) feature learning between different modalities.

Apparently, features from auxiliary database and target database might be very diverse, and lie in different feature spaces. Direct combination of these features will hinder the discriminative information. Recently, deep autoencoder has attracted lots of research attention due to its appealing performance in feature and representation learning [4]. Generally, the hidden layer of the autoencoder is used as the new representation since it can reconstruct the input by a few pairs of linear projections, and nonlinear activations, meaning the hidden layer is a rich basis for the input.

Based on these observations, in this paper, we propose to mix data from auxiliary and target databases as the input/target of the autoencoder, which potentially improves the expressiveness of the hidden layer. In addition, to avoid over-fitting problem, we propose to generate multiple deep autoencoders with bagging to mitigate the feature space gap. That is, we randomly and uniformly sample data from the existing modality and the auxiliary database several times to train several autoencoders. In addition, different from the typical setting of autoencoder with the same input and output, we intentionally enforce the outputs of these autoencoders in a common feature space learned with Robust PCA. Finally, a novel sparse low-rank feature fusion approach is proposed in the test phase to integrate multiple features learned from different autoencoders, followed by a decision voting. Extensive evaluations on VIS-NIR missing modality problems demonstrate that the proposed method is superior to the state-of-the-art methods.

A. Related Work

Multi-modality image classification have been widely discussed in three lines: (1) robust feature, (2) common subspace, (3) image synthesis. Robust feature strategy usually adapts off-the-shelf visual descriptors to the multi-modality problem [17], [19]. Common subspace methods prefer to mitigate divergence between different feature spaces in a learned subspace [29], [15]. Finally, image synthesis methods focus on the appearance model and attempt to simulate the facial appearance in specific modality [8], [23]. However, all these methods need the complete multi-modal data to train the learning system, which is apparently not satisfied in our problem. Therefore, these methods cannot be directly applied on the missing modality problem.

The auxiliary database discussed is highly related to transfer learning [21], specifically, visual domain adaptation that includes two categories of methods. The first category attempts to find a better representation in the subspace where the domain shift is mitigated [24], [11], [13]. In addition, it can easily adapt to labeled/unlabeled source/target data by using supervised/unsupervised subspace learning. On the other hand, classifier adaptation [10] uses both labeled source and target data to adjust the classifier to the new problem in the target domain, which has been extensively discussed in video analysis. However, none of them can perfectly solve the missing modality problem mentioned in this paper, since the target data are inaccessible during the training stage.

Low-rank matrix analysis recently has been introduced for subspace recovery or segmentation [7], [18], matrix completion [14] where corrupt or missing data undermine the structure of whole data. Among them, low-rank representation (LRR) [18] is able to identify the subspace structure, which makes it very suitable for structure dependent vision problems such as image segmentation, visual domain adaptation. In addition, it has been extended to robust late fusion which aims at integrating multiple scores from classifiers [28]. Different from these methods, the proposed sparse low-rank fusion scheme still works on the feature level, and treats all the features from different autoencoders as distributed in a common feature space.

Autoencoder [3] aims at learning self-expressive descriptors from the hidden layer of neural networks. It provides appearance honestly reflected feature by minimizing the reconstruction loss. Its recent success on deep structure can be attributed to the layer-wise training strategy [4] and some extra constraint [26] on the input/output. In this work, we propose a parallel structure for multiple autoencoders, which are appropriate for mitigating feature discrepancy incurred by the missing modality problem. To the best of our knowledge, this is the first time when multiple autoencoders are learned simultaneously based on bagging strategy to build common feature space, and avoid the over-fitting problem.

II. Deep Autoencoder with Bagging

A. Deep Autoencoder

Let $X = \{x_1, x_2, \ldots, x_n\}$, $x \in \mathbb{R}^D$ represent features from n images, where D is the dimensionality of the visual descriptor. The feed-forward process of the autoencoder includes two steps: (1) "inputs→hidden units", (2) "hidden units→outputs", and we denote them as f_1 and f_2, respectively. Then the first part of feed-forward process can be expressed in: $a_i = f_1(x_i) = \sigma(W_1 x_i + b_1)$, where $W_1 \in \mathbb{R}^{d\times D}$, $b_1 \in \mathbb{R}^d$, and σ is the sigmoid function which has the form of $\sigma(x) = (1 + e^{-x})^{-1}$. Similarly, the second part of the feed-forward process takes hidden units as the input to generate hypothesis $h(x_i) = f_2(a_i) = \sigma(W_2 a_i + b_2)$, where $W_2 \in \mathbb{R}^{D\times d}$, $b_2 \in \mathbb{R}^D$. The objective of autoencoder is to minimize the following loss function:

$$\min_{W_1,b_1,W_2,b_2} \mathcal{L}(X) = \min_{W_1,b_1,W_2,b_2} \frac{1}{2n}\sum_{i=1}^{n} \|x_i - h(x_i)\|_2^2, \quad (1)$$

where n is the number of features. In this way, the neurons in the hidden layer are good representations for the input, since the reconstruction process enable them to capture the intrinsic structure of the input.

Next, we introduce two regularizers: matrix Frobenius norm and KL divergency, to address the over-fitting problem of the above model. Specifically, the KL divergency regularizer allows the model to generate higher dimensional features in the hidden layer. We therefore are able to reformulate Eq. (1) as:

$$\min_{\substack{W_1,W_2,\\b_1,b_2}} \mathcal{L}(X) + \lambda_1(\|W_1\|_{\mathrm{F}}^2 + \|W_2\|_{\mathrm{F}}^2) + \lambda_2 \sum_{i=1}^{d} \mathrm{KL}(\rho\|\hat{\rho}_i),$$

where $\|\cdot\|_{\mathrm{F}}$ is the matrix Frobenius norm, $\mathrm{KL}(\cdot)$ is the KL divergence, $\hat{\rho}_i$ is the average of the activation of the i-th hidden unit ($\hat{\rho}_i = \frac{1}{n}\sum_j a_j^{(i)}$), and ρ is a small real number. Such unconstraint optimization problem can be effectively solved by L-BFGS optimizor [20], which is able to tackle large-scale data with limited memory. Note that the deep structure is trained in a layer-wise fashion, meaning we can train the i-th layer of autoencoder by using the hidden layer of the (i-1)-th autoencoder as both the input and target.

B. *Bagging with Multiple Autoencoders*

Although conventional autoencoder can find a good identity representation, it does not consider the feature space shift between different modalities or databases. To be concrete, it does not intentionally compute the common feature space. To this end, we propose two methods to construct common representation among different modalities and databases: (1) picking data from different databases or modalities to train the autoencoder; (2) setting the common feature space as the output to guide the autoencoder.

The first method trains the autoencoder with the capability of reconstructing data in any modality or database. We can implement this by simply sampling data uniformly from the auxiliary database and the existing modality of the target database. As to the second method, finding a common feature space for the arbitrary mixture of data is a non-trivial task. The challenge is these data lie in different feature spaces far away from each other.

Bagging or bootstrap aggregating [5] has long been discussed for computer vision and data mining problems, which is able to ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms. In brief, bagging generates m different sub-training sets which are uniformly sampled from the original training sets with replacement. Then either weak feature extractors, or classifiers can be trained independently based on these sub-training sets. The final step is to aggregate these weak extractors/classifers to build a strong one. In our problem, to learn a feature space that treats NIR and VIS images from both target and auxiliary databases with no difference, we utilize deep autoencoder to learn the hidden subspace, which is able to represent both NIR and VIS from different databases well. Therefore, we could make this feature learning process more general, and robust to over-fitting. The bagging strategy is described as:

1) Pick a shared sampling rate α for both target and auxiliary databases.
2) Uniformly sample images from one modality in target database, and from two modalities in the auxiliary one.
3) Repeat Step (2) m times and train m deep autoencoders with m sets of sampled data.

Suppose $X_1, X_2, \ldots, X_m$ are m different training sets sampled by bagging, then they are also the output/target values of each autoencoder, based on the definition of autoencoder. Since we would explore a common feature space rather than diverse ones, it is better to explicitly guide the output/target values during the training of multiple autoencoders. We therefore resort to Robust PCA [7] that is able to discover a sum of subspaces free of noise. In brief, Robust PCA decomposes the original data into a low-rank common feature space L, and a sparse matrix E compensating for noises and outliers. For theoretical background, please refer to [7]. If we first transform the whole target values $X = [X_1, X_2, \ldots, X_m]$ into recovered, noise free subspace L, then we can modify the target values of each autoencoder as: L_i, where $L_i = X_i - E_i$, and L_i, E_i are outputs (low-rank, and noisy parts, respectively) of Robust PCA corresponding to the i-th sub-training set.

C. *Sparse Low-Rank Feature Fusion*

Recent research on face recognition suggests that facial images lie in some low-dimensional space although the ambient dimension may be very high. Recently, low-rank representation (LRR) has been widely discussed in subspace segmentation/recovery and feature learning, which can be formulated as:

$$\min_{Z,E} \|Z\|_* + \lambda_1\|E\|_{2,1}, \quad \text{s.t.} \quad X = XZ + E, \tag{2}$$

where $X \in \mathbb{R}^{D\times n}$ is the training data, Z is the affinity matrix that needs to be recovered under the constraint of self-reconstruction plus an error term E, $\|\cdot\|_*$ is the nuclear norm of a matrix used as the convex surrogate of the original rank constraint, λ_1 is the balancing parameter, and $\|\cdot\|_{2,1}$ means the $L_{2,1}$-norm. Intuitively, LRR attempts to find the correlations between different data samples under the matrix low-rank constraint, which can be immediately applied to spectral clustering for subspace segmentation, or feature learning.

However, LRR treats all features as they are sampled from a union of feature spaces based on the underlying structure of the data, which may ignore or even conflict with the knowledge we have from multiple deep autoencoders. To make this concrete, suppose we have m autoencoders and n feature vectors, and the features generated by the m autoencoders are $X = [X_1, X_2, \ldots, X_m]$, in the ideal case, LRR will rebuild a single feature vector $x_i \in X_j$ by the feature space of X_j. This is highly preferred for the problem of subspace segmentation, but not enough for our feature fusion problem. Since $X_1, X_2, \ldots, X_m$ lie in different feature spaces but for the same data samples, if x_i picks a few samples in X_j for reconstruction, it should also pick the samples in X_k where $j \neq k$. This inspires us to add an extra spatial constraint on the formulation of coefficient matrix Z.

Suppose each feature set[1] X_i from a single deep autoen-

[1] We slightly abuse the symbol X_i here as the feature generated by the i-th autoencoder. Note that it was used as the i-th training set sampled from bagging in the last section.

coder has its own reconstruction coefficients matrix $Z_i (1 \leq i \leq m)$, then the problem above can be written as:

$$\min_{Z,E} \|Z\|_* + \lambda_1 \|E\|_{2,1}, \quad \text{s.t.} \quad [X_1, X_2, \ldots, X_m] = X[Z_1; Z_2; \ldots; Z_m] + E, \tag{3}$$

where "$[\ldots, \ldots]$" and "$[\ldots; \ldots]$" mean the column-wise and row-wise matrix concatenation. To impose the consistence of features among different autoencoders/feature spaces, we explicitly enforce the following structure to be column-wise sparse:

$$\Psi(Z) = \begin{bmatrix} (Z_{11})_1 & (Z_{12})_1 & \ldots & (Z_{nn})_1 \\ (Z_{11})_2 & (Z_{12})_2 & \ldots & (Z_{nn})_2 \\ \vdots & \vdots & \ddots & \vdots \\ (Z_{11})_m & (Z_{12})_m & \ldots & (Z_{nn})_m \end{bmatrix}.$$

Clearly, each column is features for the same dimensionality but from different autoencoders. Since these features are assumed to be aligned well after deep feature learning, the coefficients/entries with the same coordinate over different Z_i should be very close or have similar magnitudes. The column-wise sparse on $\Psi(Z)$ can exactly model such spatial constraint, which leads to a novel feature fusion formulation:

$$\min_{Z,E} \|Z\|_* + \lambda_1 \|E\|_{2,1} + \lambda_2 \|\Psi(Z)\|_{2,1}, \quad \text{s.t.} \quad X = XZ + E. \tag{4}$$

The problem above can be solved by Augmented Lagrangian Multiplier (ALM) method, where each variable is optimized at one time by fixing others. The optimization process will not terminate until all the target variables are converged.

First of all, we transform it into the following optimization problem by introducing two relaxing variables J, S:

$$\arg\min_{Z,E,J} \|J\|_* + \lambda_1 \|E\|_{2,1} + \lambda_2 \|\Psi(Z)\|_{2,1} \quad \text{s.t.} \; X = XS + E, \; Z = J, \; Z = S, \tag{5}$$

whose augmented Lagrangian function is:

$$\begin{aligned} &\|J\|_* + \lambda_1 \|E\|_{2,1} + \lambda_2 \|\Psi(Z)\|_{2,1} + \langle Y_1, Z - J \rangle + \\ &\langle Y_2, Z - S \rangle + \langle Y_3, X - XS - E \rangle + \tfrac{\mu}{2}(\|Z - J\|_{\mathrm{F}}^2 \\ &+ \|Z - S\|_{\mathrm{F}}^2 + \|X - XS - E\|_{\mathrm{F}}^2), \end{aligned} \tag{6}$$

where Y_1, Y_2, Y_3 are the three lagrange multipliers and $\mu > 0$ is a penalty parameter. The operator $\langle \cdot \rangle$ means the inner product of two matrixes. We alternately optimize the variables J, S, E and $\Psi(Z)$ in the $t + 1$ iteration:

$$J_{t+1} = \arg\min_J = \frac{1}{\mu}\|J\|_* + \frac{1}{2}\|J - (Z_t + \frac{Y_1}{\mu})\|_{\mathrm{F}}^2, \tag{7}$$

$$S_{t+1} = (\mathrm{I} + X^{\mathrm{T}}X)^{-1}(Z_t + X^{\mathrm{T}}(X - E_t) + \frac{Y_2 + X^{\mathrm{T}}Y_3}{\mu}), \tag{8}$$

$$E_{t+1} = \arg\min_E \frac{\lambda_1}{\mu}\|E\|_{2,1} + \frac{1}{2}\|E - (X - XS_{t+1} + \frac{Y_3}{\mu})\|_{\mathrm{F}}^2, \tag{9}$$

$$\Psi(Z_{t+1}) = \arg\min_{\Psi(Z)} \frac{\lambda_2}{\mu}\|\Psi(Z)\|_{2,1} + \frac{1}{2}\|\Psi(Z) - \Psi(F_t)\|_{\mathrm{F}}^2, \tag{10}$$

where $F_t = \frac{1}{2}(S_{t+1} + J_{t+1} - \frac{Y_1 + Y_2}{\mu})$. To be specific, we can solve Eq. (7) through Singular Value Thresholding (SVT) [6]. Whilst Eqs. (9) and (10) can be addressed by the shrinkage operator [27]. **Algorithm 1** depicts the details of the solution.

```
Algorithm 1 Solution to Problem (6)
Input: $X, \lambda_1, \lambda_2$
Initialize: $S_0 = J_0 = E_0 = Y_1 = Y_2 = Y_3 = 0$,
        $\mu = 10^{-6}, \rho = 1.2, \max_\mu = 10^6, \epsilon = 10^{-6}, t = 0$.
while not converged do
  1. Optimize $J_{t+1}$ via Eq. (7), when fixing others.
  2. Optimize $S_{t+1}$ via Eq. (8), when fixing others.
  3. Optimize $E_{t+1}$ via Eq. (9), when fixing others.
  4. Optimize $\Psi(Z_{t+1})$ via Eq. (10), when fixing others.
  5. Update those multipliers $Y_1, Y_2, Y_3$ via
       $Y_1 = Y_1 + \mu(Z_{t+1} - J_{t+1})$; $Y_2 = Y_2 + \mu(Z_{t+1} - S_{t+1})$;
       $Y_3 = Y_3 + \mu(X - XS_{t+1} - E_{t+1})$.
  6. Update the parameter $\mu$ via $\mu = \min(\rho\mu, \max_\mu)$
  7. Check the convergence conditions
       $\|X - XS_{t+1} - E_{t+1}\|_\infty < \epsilon, \|Z_{t+1} - J_{t+1}\|_\infty < \epsilon$,
       $\|Z_{t+1} - S_{t+1}\|_\infty < \epsilon$.
  8. $t = t + 1$;
end while
output: $S_{t+1}, E_{t+1}, J_{t+1}, \Psi(Z_{t+1})$
```

Sparse low-rank fusion is able to refine the multiple features from different autoencoders, and its solution $Z = [Z_1; Z_2; \ldots; Z_m]$ is the new representation for the refined features. Therefore, we will have m decision results from m nearest neighbor classifiers. Afterwards, we conduct a simple voting scheme to decide the label for each test sample. Note that a more accurate yet complex late fusion scheme may exist by manipulating the decision scores, e.g., weighted scores fusion; however, its detailed discussion is already beyond the focus of this paper.

III. Experiments

A. Data and Experimental Setup

We conduct experiments on one set of multi-modal databases: BUAA[2] and Oulu-CASIA[3] (we use Oulu for short) visible light (VIS) and near infrared (NIR) face databases (Figure 2). BUAA database contains 15 subjects while Oulu database includes 80 subjects, and each of them contains two modalities: NIR and VIS. We crop images to a size of 30×30, and use a subset of BUAA with 75 subjects and their corresponding VIS images as one modality, and the remaining 75 subjects with their NIR images as another modality. Similarly, we select 40 subjects with their VIS images, and remaining 40 subjects with their NIR images from Oulu database. In the following experiments, we will alternate the missing modality in each test to showcase the performance of all methods. For example, the "curve" or "bar" marked by "BUAA NIR" means the NIR modality from BUAA database is used as the missing modality.

In each experiment, we randomly choose one sample per individual from the testing data as the reference. Note that there is no overlap between the testing data and the reference. The nearest-neighbor classifier is used to predict the labels of the testing samples. Without specification, the default setup of our method is: the hidden layer number of the autoencoder is 1; the dimensionality of hidden layer is 100; the number

[2] http://irip.buaa.edu.cn/Research.html

[3] http://www.ee.oulu.fi/∼gyzhao/

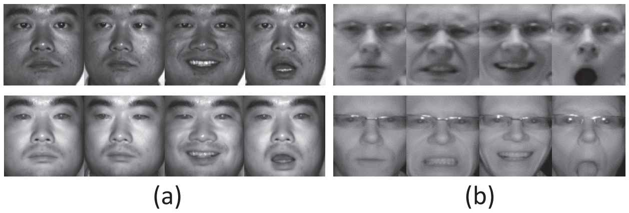


Fig. 2: Sample faces from (a) BUAA and (b) Oulu-CASIA databases. Facial images in the first row are VIS images, and those in the second row are corresponding NIR images. We show two subjects with different poses/expressions from the two databases.

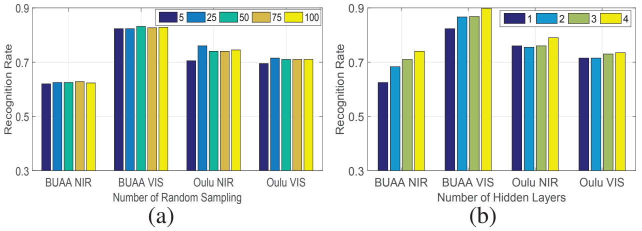


Fig. 3: Recognition rates over (a) different numbers of random samplings (5-100); recognition rates over (b) different numbers of layers (1-4).

of random sampling is 25; the ratio of each random sampling is 0.4. We repeat each experiment five times, and the average results are reported.

B. Experimental Results and Analysis

In the first set of the experiments, we systematically evaluate the proposed method with different parameters on BUAA and Oulu databases.

The performance over different numbers of sampling is illustrated in Figure 3a, showing that the number of random sampling does not affect the performance dramatically. Four datasets: BUAA NIR, BUAA VIS, Oulu NIR and Oulu VIS achieve their best performance when sampling numbers are 75, 50, 25 and 25, respectively. Too few sampling, say 5 leads to the worst performance in each dataset. So we suggest a random sampling number from 25 to 75 in this work. Figure 3b illustrates the performance of our method when the hidden layer number of autoencoder is at 1, 2, 3 and 4. The setting of layers and dimensionality is: one layer (100); two layers (300, 100); three layers (700, 300, 100) and four layers (700, 500, 300, 100). From the results, we learn that the performance is improved as the number of layers increases, especially for BUAA NIR dataset that achieves an improvement from 0.63 to 0.74 when the hidden layer number increases from 1 to 4. Clearly, deep structure could boost the system, but needs more computational cost.

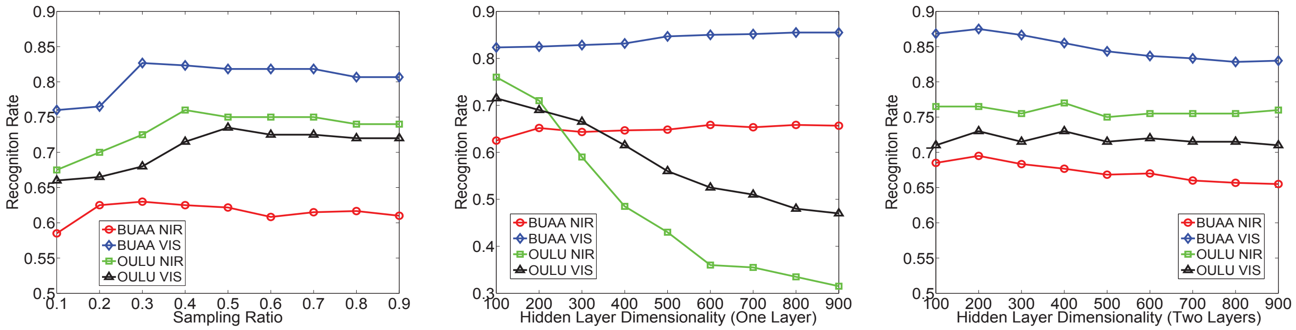
Figure 4a shows the recognition rates of our method with different sampling ratios. For BUAA database, two modalities achieve best results when sampling ratio is 0.3. For Oulu database, NIR modality achieves the best result at sampling ratio of 0.4 while that for VIS is 0.5. From these results, we can see that our method achieves best results when sampling ratio ranges from 0.3 to 0.5, meaning we could use a moderate sampling ratio for the best performance. Figure 4b illustrates the recognition rates at different dimensionality of the hidden layer in a one-hidden-layer autoencoder. Dimensionality does not affect BUAA database much when it changes from 100 to 900, but affects much on Oulu database. Figure 4c shows how the dimensionality affects the recognition ability of a two-layer autoencoder. Since the best performance is achieved when the dimensionality is 100 for an one-layer autoencoder, we fix the dimensionality of the second hidden layer at 100. From Figure 4c, we can see that the dimensionality of the first layer (in a two hidden layers autoencoder) does not affect the recognition rate much. Hence, the dimensionality of the final layer is important for recognition ability of our method.

In the second set of the experiments, we compare our model with transfer learning that can only consider knowledge transfer in one direction assuming both source and target data are available during the training. It includes transfer subspace learning (TSL) [24], low-rank transfer subspace learning (LTSL) [22], robust domain adaptation with low-rank reconstruction (RDALR) [13], geodesic flow kernel (GFK) [11] and L^{2}STL [9] in different subspace settings: PCA[25], LDA [2], Unsupervised LPP (ULPP) and Supervised LPP (SLPP) [12].

We denote our method without bagging as "Ours-I" and that with bagging as "Ours-II". The parameters setting is: the hidden layer number of the autoencoder is 4; the dimensionality of hidden layers is (700,500,300,100); the number of random sampling is 25; the ratio of each random sampling is 0.4. For each test on the transfer learning based methods, e.g., TSL, LTSL, RDALR, GFK, DASA, L^{2}TSL, the source domain is one modality of two databases, and the target domain is set as another modality of the auxiliary database. The missing modality in the objective database provides the testing data. We switch source and target domains by switching modalities. Table I lists the recognition rates of different algorithms. From these results, we can observe that the proposed method achieves good performance even without bagging (in Case 2). This is because our model not only can couple the knowledge between modalities, but also can mitigate the divergence between databases. Furthermore, Ours-II outperforms Ours-I in all cases, which proves that bagging + sparse low-rank fusion is effective on enhancing the performance of conventional deep autoencoders for the missing modality problem.

IV. Conclusions

In this paper, we designed a novel framework to address the image classification with missing modality. First, we proposed to train an autoencoder by a mixture of multi-modal data from the auxiliary database and the existing modality from the target database. Second, a common feature space was intentionally set as the target of the autoencoder to guide the shared feature learning, where bagging strategy was adopted to boost the performance. Finally, in the testing



(a) (b) (c)

Fig. 4: Recognition rates over (a) different sampling ratios ("ratio" means the number of select samples over the total samples); over (b) the dimensionality of one hidden layer; over (c) the dimensionality of two hidden layers.

TABLE I: Average recognition rates (%) of all compared methods on BUAA&Oulu face database, where the test data, respectively, are NIR of BUAA (**Case 1**), VIS of BUAA (**Case 2**), NIR of Oulu (**Case 3**) and VIS of Oulu (**Case 4**).

Methods	**Case 1**				**Case 2**				**Case 3**				**Case 4**			
	PCA	LDA	ULPP	SLPP	PCA	LDA	ULPP	SLPP	PCA	LDA	ULPP	SLPP	PCA	LDA	ULPP	SLPP
TSL [24]	35.8	31.3	29.2	36.8	37.0	28.3	38.2	46.8	39.2	42.2	47.3	45.7	31.5	40.3	39.2	36.2
RDALR [13]	40.2	38.5	42.8	47.2	33.7	34.5	39.8	50.2	41.3	36.5	42.3	48.2	39.7	42.3	47.5	49.3
GFK [11]	38.3	12.7	40.2	39.5	42.3	15.8	39.2	48.3	39.5	26.8	28.3	45.3	39.2	38.3	42.8	29.3
LTSL [22]	47.2	42.3	50.8	53.5	38.3	41.3	41.2	56.7	41.8	50.7	48.2	54.7	43.3	48.2	52.3	58.8
L^2TSL [9]	52.3	48.7	59.7	**63.7**	49.8	43.2	49.3	60.7	48.3	**56.8**	50.8	55.7	46.3	67.5	58.2	**68.5**
Ours-I	46.22				**66.12**				50.92				50.72			
Ours-II	**74.03**				**89.83**				**79.06**				**73.47**			

stage, a sparse low-rank feature fusion approach was proposed to further enforce the consistency of features from different autoencoders. Experimental results on BUAA and Oulu-CASIA databases have demonstrated the advantage of the proposed method by comparing with several related state-of-the-art methods.

REFERENCES

[1] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE TPAMI*, 19(7):711–720, 1997.

[3] Y. Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[4] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. *NIPS*, 19:153, 2007.

[5] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[6] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[7] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

[8] J. Chen, D. Yi, J. Yang, G. Zhao, S. Li, and M. Pietikainen. Learning mappings for face synthesis from near infrared to visual light images. In *IEEE CVPR*, pages 156–163, 2009.

[9] Z. Ding, S. Ming, and Y. Fu. Latent low-rank transfer subspace learning for missing modality recognition. In *AAAI*, 2014.

[10] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *IEEE TPAMI*, 34(9):1667–1680, 2012.

[11] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE CVPR*, pages 2066–2073. IEEE, 2012.

[12] X. He and P. Niyogi. Locality preserving projections. In *NIPS*, volume 16, page 153, 2004.

[13] I.-H. Jhuo, D. Liu, D. Lee, and S.-F. Chang. Robust visual domain adaptation with low-rank reconstruction. In *IEEE CVPR*, pages 2168–2175. IEEE, 2012.

[14] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[15] Z. Lei and S. Z. Li. Coupled spectral regression for matching heterogeneous faces. In *IEEE CVPR*, pages 1123–1128, 2009.

[16] S. Z. Li, R. Chu, S. Liao, and L. Zhang. Illumination invariant face recognition using near-infrared images. *IEEE TPAMI*, 29(4):627–639, 2007.

[17] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Z. Li. Heterogeneous face recognition from local structures of normalized appearance. In *Advances in Biometrics*, pages 209–218. Springer, 2009.

[18] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670, 2010.

[19] S. Liu, D. Yi, Z. Lei, and S. Z. Li. Heterogeneous face image matching using multi-scale features. In *IAPR ICB*, pages 79–84. IEEE, 2012.

[20] J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.

[21] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2010.

[22] M. Shao, C. Castillo, Z. Gu, and Y. Fu. Low-rank transfer subspace learning. In *IEEE ICDM*, pages 1104–1109. IEEE, 2012.

[23] M. Shao, Y. Wang, and Y. Wang. A super-resolution based method to synthesize visual images from near infrared. In *IEEE ICIP*, pages 2453–2456. IEEE, 2009.

[24] S. Si, D. Tao, and B. Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE TKDE*, 22(7):929–942, 2010.

[25] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[26] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103. ACM, 2008.

[27] J. Yang, W. Yin, Y. Zhang, and Y. Wang. A fast algorithm for edge-preserving variational multichannel image restoration. *SIAM Journal on Imaging Sciences*, 2(2):569–592, 2009.

[28] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang. Robust late fusion with rank minimization. In *IEEE CVPR*, pages 3021–3028. IEEE, 2012.

[29] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li. Face matching between near infrared and visible light images. In *Advances in Biometrics*, pages 523–530. Springer, 2007.