

# Learning to Authenticate with Deep Multibiometric Hashing and Neural Network Decoding

Veeru Talreja, Sobhan Soleymani, Matthew C. Valenti, and Nasser M. Nasrabadi

Lane Department of Computer Science and Electrical Engineering

West Virginia University, Morgantown, WV, USA

Email: vtalreja@mix.wvu.edu, ssoleyma@mix.wvu.edu, valenti@ieee.org, and nasser.nasrabadi@mail.wvu.edu

**Abstract**—In this paper, we propose a novel multimodal deep hashing neural decoder (MDHND) architecture, which integrates a deep hashing framework with a neural network decoder (NND) to create an effective multibiometric authentication system. The MDHND consists of two separate modules: a multimodal deep hashing (MDH) module, which is used for feature-level fusion and binarization of multiple biometrics, and a neural network decoder (NND) module, which is used to refine the intermediate binary codes generated by the MDH and compensate for the difference between enrollment and probe biometrics (variations in pose, illumination, etc.). Use of NND helps to improve the performance of the overall multimodal authentication system. The MDHND framework is trained in 3 steps using joint optimization of the two modules. In Step 1, the MDH parameters are trained and learned to generate a shared multimodal latent code; in Step 2, the latent codes from Step 1 are passed through a conventional error-correcting code (ECC) decoder to generate the ground truth to train a neural network decoder (NND); in Step 3, the NND decoder is trained using the ground truth from Step 2 and the MDH and NND are jointly optimized. Experimental results on a standard multimodal dataset demonstrate the superiority of our method relative to other current multimodal authentication systems.

**Index Terms**—deep hashing, multibiometric, error-correcting codes, authentication

## I. INTRODUCTION

Multimodal biometric authentication systems use a combination of different biometric traits such as face and iris, or face and fingerprint to authenticate the user. Multimodal systems have the advantage of lower error rates and higher accuracy when compared to unimodal systems [1]. Additionally, multimodal systems are generally more resistant to spoofing attacks, and can be made to be more universal than unimodal systems, since the use of multiple modalities can compensate for missing modalities in a small portion of the population [2].

One of the major challenges of multimodal systems is the selection of the fusion algorithm required to combine the multiple modalities. The fusion of the modalities can be performed at sensor, feature, score, or decision levels [2]–[5]. However, fusion-level authentication systems have better performance because they leverage the richer information available about the biometric data in the feature-set [6]. Feature-level fusion involves combining of feature vectors that are obtained from multiple feature sources. Multibiometric systems can combine biometrics in many ways, including: (i) feature vectors obtained from different sensors for the same biometric; (ii) feature vectors obtained from different entities

for the same biometric, such as combination of feature vectors obtained from left and right eyes; or (iii) feature vectors obtained from multiple biometric traits, such as face and iris. Fusion at the feature level is relatively difficult to achieve in practice due to incompatibility of feature sets extracted from multiple modalities and because the relationship between different feature spaces may be unknown [7]. To overcome this issue, we propose a feature-level fusion of multiple biometrics using convolutional neural networks (CNNs).

In this paper, we have combined multiple modalities at feature-level using two different CNN architectures. Similar work has been done in [8], however, the CNN features are high-dimensional and real-valued, and usually require high computational cost [9]. In order to reduce the computational complexity, in this paper, we have also integrated a novel deep hashing algorithm into a feature-level fusion CNN architecture to design a multimodal deep hashing (MDH) network for combining multiple modalities. Deep hashing is the application of deep learning to generate compact binary vectors from raw image data [10], [11]. In addition to using deep hashing for feature-level fusion, we have also used error correcting codes (ECC) as an additional component to compensate for the difference in enrollment and probe biometrics (arising from variation in pose, illumination, noise in biometric capture). Due to this difference, enrollment and probe biometrics lead to different hash code at the output of the MDH network and therefore a failure to authenticate. ECC can compensate for this difference by forcing the enrollment and probe biometric to decode to the same message, then using that message for authentication, making the system more robust to distortion in the biometric measurements, which helps in improving the authentication performance.

Recent work has shown that the same kinds of neural network architectures used for classification can also be used to decode ECC codes [12]. In this work, we integrate a neural network decoder (NND) [12] into our deep hashing architecture as an ECC component to improve the authentication performance. The NND is a formulation of the belief propagation (BP) algorithm as a neural network. The input to the NND can be considered to be a corrupted codeword of an ECC and this corrupted codeword is within a certain distance of a correct codeword of an ECC. The corrupted codeword can be decoded using the NND to generate the correct codeword. It can be argued that a conventional ECC decoder can be used

instead of using a NND, however, implementing the decoder as a neural network has the benefit of providing the same architecture as the classifier (MDH network) that extracts the hashing code, and hence, it can be more efficiently jointly optimized and implemented within a common framework. Another advantage with NND is that it provides an opportunity to jointly learn and optimize with respect to biometric datasets, which are not necessarily characterized simply by Gaussian noise as is assumed by a conventional decoder. Motivated by this, in this paper, we have integrated our MDH network with NND and a joint optimization process to formulate our novel MDHND framework for an end-to-end multimodal biometric authentication system.

To summarize the main contributions of this paper include:

- 1) Conversion of the involved biometric traits into a common feature space by using modality-specific CNNs.
- 2) Fusion and binarization of the individual modality features to generate a robust binary latent shared representation.
- 3) Inclusion and optimization of the neural network based decoder to compensate for the distortion in biometric measurements.
- 4) End-to-end joint optimization of the overall system.

## II. MULTIMODAL DEEP HASHING NEURAL DECODER

In this section, we present a system overview of the multimodal deep hashing neural decoder (MDHND) shown in Fig. 1. MDHND consists of two important modules: multimodal deep hashing (MDH) module and neural network decoder (NND) module. We have considered face and iris as the two modalities for this authentication system. However this system could be used with other modalities and can be extended to more than two modalities.

### A. Multimodal Deep Hashing Module

The multimodal deep hashing (MDH) module, which is shown in Fig. 2, consists of a domain-specific layer containing face and iris CNNs and the joint representation layer. The primary functions of the MDH module are the non-linear feature-level fusion and binarization of the fused features.

The domain-specific layer of the MDH module consists of a CNN for encoding the face (“Face-CNN”) and another CNN for encoding the iris (“Iris-CNN”). The output feature vectors of the face and iris CNNs are fused and binarized in the joint representation layer (JRL), which is split into two sub-layers: a fusion layer and a hashing layer. The main function of the fusion layer is to fuse the individual face and iris representations from domain-specific layers into a non-linear multimodal feature embedding. The hashing layer binarizes the shared multimodal feature representation that is generated by the fusion layer.

**Fusion layer:** We have implemented two different architectures for the fusion layer: Fully Connected architecture (FCA) and Bilinear architecture (BLA). In the FCA, the outputs of the Face-CNN and Iris-CNN are concatenated vertically in the concatenation layer and passed through a fully connected layer

to fuse the iris and face features. In FCA, the concatenation layer and the fully connected layer together constitute the fusion layer. In the BLA (Fig. 2), the outputs of the Face-CNN and Iris-CNN are combined using the matrix outer product of the face and iris feature vectors to create bilinear feature vector. Similar to FCA, the bilinear feature vector is also passed through a fully connected layer. In BLA, the outer product layer and the fully connected layer together constitute the fusion layer.

**Hashing layer:** The output of the fusion layer is a shared multimodal feature vector of unquantized values. The output of the fusion layer can be directly binarized by thresholding at any numerical value or thresholding at the population mean. However, this kind of thresholding leads to a quantization loss, which results in sub-optimal binary codes. To account for this quantization loss, we have included another latent layer after the fusion layer, which is known as the hashing layer (shown in orange in Fig. 2). The main function of the hashing layer is to capture the quantization loss incurred while converting the shared multimodal representation (output of fusion layer) into binary codes.

To generate the binary hash codes, we can directly use the sign activation function  $h = \text{sgn}(z)$  for the hashing layer. However, the use of the non-smooth sign-activation function makes standard back-propagation impracticable as the gradient of the sign function is zero for all non-zero inputs. We have used a continuation method to overcome this zero-gradient problem by starting with a smooth activation function  $y = \tanh(\beta x)$  and making it gradually sharper by increasing the bandwidth  $\beta$  as the training proceeds. This continuation utilizes the relationship between the sign activation function and the scaled tanh function:

$$\lim_{\beta \rightarrow \infty} \tanh(\beta x) = \text{sgn}(x), \quad (1)$$

where  $\beta > 0$  is a scaling parameter. For training the network, we start with a  $\tanh(\beta x)$  activation for the hashing layer with  $\beta = 1$  and continue training until the network converges to zero loss. We then increase the value of  $\beta$  while holding other

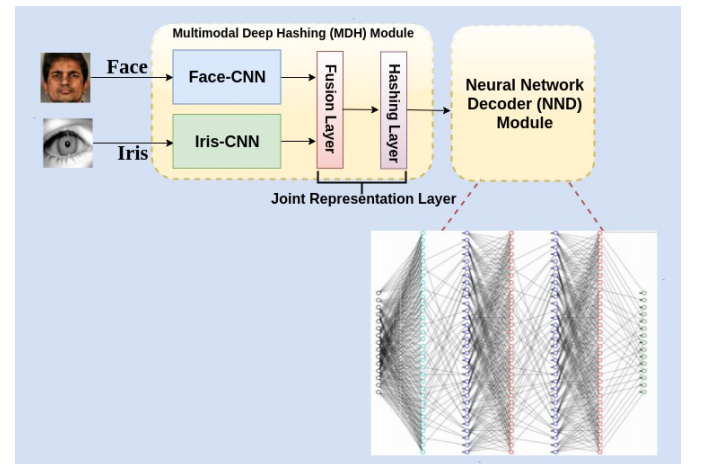


Fig. 1. Block diagram of the proposed system. The NND is shown in the figure. The MDH module is shown in Fig. 2

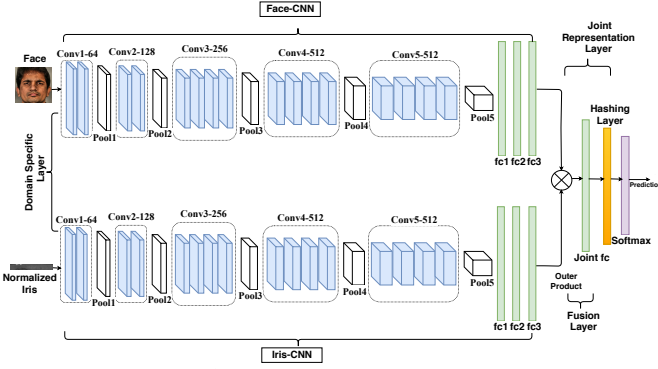


Fig. 2. Proposed multimodal deep hashing (MDH) framework for the bilinear architecture (BLA).

training parameters equal to the previously converged network parameters, and start retraining the network. This process is repeated several times by increasing the bandwidth of the tanh activation allowing  $\beta \rightarrow \infty$  until the hashing layer can generate an output very close to binary values. In addition to using this continuation method, the overall objection function used for training the deep hashing network is discussed in Sec. III-A

### B. Neural Network Decoder Module

The output of the MDH module after training can be binarized by directly using a sign function. Henceforth, we will refer to the output of the MDH network as *intermediate binary code*. Even though we still use a threshold of 0.5, the quantization loss is much lower and authentication performance is improved when compared to values with no hashing layer. However, there is still room to make the system more robust and improve the performance. This is achieved by using ECC. There could be distortion in biometric measurements such as variations in pose, illumination, or noise due to the biometric capturing device, which leads to difference in enrollment and probe biometrics. Due to these differences, enrollment and probe biometrics may lead to different hash codes at the output of the MDH network and therefore a failure to authenticate. ECC can compensate for this difference by forcing the enrollment and probe biometric to decode to the same message, then using that message for authentication, making the system more robust to noise in the biometric measurements, which helps in improving the authentication performance.

Recent research in the field of error-correcting codes (ECC) has focused on designing a neural network architecture as an ECC decoder [12], [13]. We can adapt such a neural network based ECC decoder, train it, and use it as an ancillary component to refine the intermediate binary codes generated by MDH. The advantage of using NND instead of a conventional decoder is that it allows for a common architectural framework to be used for both the classifier (i.e., the MDH) and the decoder and it provides an opportunity to jointly learn and optimize with respect to biometric datasets, which are not necessarily characterized simply by Gaussian noise as is assumed by conventional decoders. For our application, we

have chosen the NND described in [12] as the basis for our neural network ECC decoder to be integrated with the MDH network. Due to space limitations, we do not provide full details on the operation of NND, details can be found in the original paper. Rather, we focus the discussion on the changes in how the decoder is trained relative to [12].

## III. TRAINING STEPS FOR MULTIMODAL DEEP HASHING NEURAL DECODER

The MDHND framework described in Sec. II is trained in 3 steps. In Step 1, we use a novel loss function to train and learn the MDH parameters to generate binary shared multimodal latent code; in Step 2, the latent binary codes from Step 1 are passed through an ECC decoder to generate the ground truth, which will be used to fine-tune a neural network decoder (NND); in Step 3, the NND decoder is trained using the ground truth from Step 2 and this NND is then integrated with the MDH network followed by a joint optimization of the overall system MDHND. In this section, we discuss the 3 steps used during training of the proposed MDHND system.

### A. Step 1: Training of the Multimodal Deep Hashing Module

In this step, the MDH network is trained for feature-level fusion and binarization of the multiple biometrics. For modality-specific CNNs, VGG-19 pre-trained on ImageNet dataset is used as a starting point followed by fine-tuning the entire VGG-19 with additional fully connected layer  $fc3$ . The Face-CNN is fine-tuned end-to-end with the CASIA-Webface [14], which contains 494,414 facial images corresponding to 10,575 subjects. The Iris-CNN is fine-tuned end-to-end using the combination of the CASIA-Iris-Thousand [15] and ND-Iris-0405 [16] datasets with about 84,000 iris images corresponding to 1355 subjects. The Face-CNN and Iris-CNN are also fine tuned on the 2013 face and iris subsets of the WVU-Multimodal 2012-2013 datasets [17] respectively. The WVU-Multimodal dataset for the year 2012 and 2013 together contain a total of 119,700 facial images and 257,800 iris images corresponding to 2263 subjects with 294 common subjects. For fine-tuning the Face-CNN and Iris-CNN, we have used 58,200 facial images and 121,200 iris images, respectively, corresponding to 1,060 subjects from the WVU-Multimodal 2012-2013 dataset. For fine-tuning the CNNs, we have not used any of the common subjects from 2012-2013 WVU dataset, so that these common subjects can be used for fine-tuning the NND and also for testing the overall system.

For the Face-CNN, all the raw facial images are first aligned in 2-D and cropped to a size of  $224 \times 224$  before passing through the network [18]. The only other pre-processing is subtracting the mean RGB value, computed on the training set, from each pixel. The training is carried out by optimizing the multinomial logistic regression objective using mini-batch gradient descent with momentum. The number of nodes in the last fully connected layer  $fc3$  before the softmax layer is 1,024 for the FCA and 64 for the BLA. This implies that the feature vector that is extracted from the Face-CNN and fused with the feature vector from Iris-CNN has 1,024 dimensions

for the FCA and 64 for the BLA. For the iris-CNN, all the raw iris images are segmented and normalized to a fixed size of  $64 \times 512$  using Osiris (Open Source for IRIS) which is an open source iris recognition system developed in the framework of the BioSecure project [19]. As with the Face-CNN, the iris network has an output of 1,024 for FCA and 64 for BLA.

Next, we train the joint representation layer, which is a combination of the fusion layer and the hashing layer. For training the joint representation layer we have used a novel objective function, which helps in reducing the quantization loss and also maximize the entropy to generate optimal and discriminative binary shared multimodal representation. For training the joint representation layer, we adopt a two-step training procedure where we first train only the joint representation layer greedily by freezing the face and iris CNNs followed by fine tuning the entire MDH module end-to-end using back-propagation at a relatively small learning rate.

The objective function used for training the joint representation layer is a combination of classification loss, quantization loss and entropy maximization loss. The classification loss has been added into the MDH network by using the *softmax* layer as shown in Fig. 2. Let  $E_1(\mathbf{w})$  denote the objective function required to fulfill the classification task:

$$E_1(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N L_n(f(x_n, \mathbf{w}), y_n) + \lambda \|\mathbf{w}\|^2, \quad (2)$$

where the first term  $L_n(\cdot)$  is the classification loss for a training instance  $n$  and is described below,  $N$  is the number of training images in a mini-batch.  $f(x_n, \mathbf{w})$  is the predicted softmax output of the network and is a function of the input training image  $x_n$  and the weights of the network  $\mathbf{w}$ . The second term is the regularization function where  $\lambda$  governs the relative importance of the regularization. Let the predicted softmax output  $f(x_n, \mathbf{w})$  be denoted by  $\hat{y}_n$ . The classification loss for the  $n^{\text{th}}$  training instance is given as:

$$L_n(\hat{y}_n, y_n) = - \sum_{m=1}^M y_{n,m} \ln \hat{y}_{n,m}, \quad (3)$$

where  $y_{n,m}$  and  $\hat{y}_{n,m}$  is the ground truth and the prediction result for the  $m^{\text{th}}$  unit of the  $n^{\text{th}}$  training instance, respectively and  $M$  is the number of output units. In addition to using classification loss and the continuation method described in Sec. II-A, we add constraint of quantization loss and entropy maximization to generate efficient binary codes at the output of the MDH module. The output of the hashing layer is a  $J$ -dimensional vector denoted by  $\mathbf{o}_n^H$ , corresponding to the  $n$ -th input image. The  $i$ -th element of this vector is denoted by  $o_{n,i}^H$  ( $i = 1, 2, 3, \dots, J$ ). The value of  $o_{n,i}^H$  is in the range of  $[-1, 1]$  because it has been activated by the tanh activation. To make the codes closer to either -1 or 1, we add a constraint of quantization loss between the hashing layer activations and 0, which is given by  $\sum_{n=1}^N \|\mathbf{o}_n^H - \mathbf{0}\|^2$ , where  $N$  is the number of training images in a mini-batch and  $\mathbf{0}$  is the  $J$ -dimensional vector with all elements equal to 0. Let  $E_2(\mathbf{w})$  denote this constraint to boost the activations of the units in the hashing layer to be closer to -1 or 1:

$$E_2(\mathbf{w}) = -\frac{1}{J} \sum_{n=1}^N \|\mathbf{o}_n^H - \mathbf{0}\|^2 = -\frac{1}{J} \sum_{n=1}^N \|\mathbf{o}_n^H\|^2. \quad (4)$$

In addition to forcing the codes to become binarized, we also require that the binary codes have equal number of -1's and 1's, which maximizes the entropy of the discrete distribution and results in hash codes with better discrimination. Let  $E_3(\mathbf{w})$  denote this constraint that forces the output of each node to have a 50% chance of being -1 or 1:

$$E_3(\mathbf{w}) = \sum_{n=1}^N (\text{mean}(\mathbf{o}_n^H))^2. \quad (5)$$

Therefore, the overall objective function that is required to be minimized to generate discriminative efficient binary codes is given as:

$$\alpha E_1(\mathbf{w}) + \beta E_2(\mathbf{w}) + \gamma E_3(\mathbf{w}), \quad (6)$$

where  $\alpha$ ,  $\beta$ , and  $\gamma$  are the tuning parameters of each term.

### B. Step 2: Generating the Ground Truth for Training the Neural Network Decoder

In [12], the goal is to optimize the NND for a gaussian noise channel and the database used reflects various channel output realizations when the zero codeword has been transmitted. However, for our proposed system, we want to optimize the NND to be used with biometric data, where the channel noise for our database is characterized by the image distortions (e.g., pose variations, illumination variations and noise due to biometric capturing device) in different biometric images of the same subject. We can create the input dataset for training the NND by using the different facial and iris images for the same subject. However, we do not have the labels or the ground truth codewords that these input images need to be mapped to. An external conventional ECC decoder can be used to generate the ground truth codewords for this input dataset.

After training the MDH network in Step 1, facial and iris image pairs for a different set of subjects disjoint from the training dataset are used for generating the ground truth for training the NND. We use the MDH network to extract the joint binary features for this disjoint dataset. These extracted features are used as input to a conventional ECC decoder for soft-decision decoding. Hard-limiting the output of the ECC decoder generates ground truth codewords that are used as labels for optimizing the NND in Step 3.

For this step, we have used the facial and iris image pairs of 294 common subjects from the 2013 year of the WVU multimodal 2012-2013 dataset. This implies that we have extracted the features for all the face and iris image pairs of the 294 subjects using the MDH network and decoded it using soft-decision decoding with an ECC decoder to generate the ground truth to be used in Step 3. While usually all the joint feature vectors of a given subject are mapped by the decoder to the same codeword, it is possible that the vectors could be mapped to different codewords. This is especially the case when there are substantive differences in the latent subspaces of the different face/image pairs for the subject or

when the ECC code is not sufficiently strong. In the case that the feature vectors get mapped to a plurality of codewords, the most common of these codewords is used as the ground truth for that subject.

### C. Step 3: Joint Optimization of the Multimodal Deep Hashing Neural Decoder

In Step 3 of the training, first the NND is trained using the procedure and database from the original paper [12]. In the next step, the NND is fine-tuned for our multibiometric data. For fine-tuning the NND, we use the same database used for the ECC decoder in Step 2, where the input to NND is given by the feature vectors generated by the MDH network and the labels are provided by the decoded codewords generated by the conventional ECC decoder in Step 2. Using this database, the NND is fine-tuned for our joint biometric data. We have used sigmoid activation for the last layer of NND. The sigmoid is added so that the final network output is in the range  $[0, 1]$ . This makes it possible to train and fine-tune the NND using cross-entropy loss function:

$$L(o, y) = -\frac{1}{N} \sum_{i=1}^N y_i \log(o_i) + (1 - y_i) \log(1 - o_i), \quad (7)$$

where  $o_i, y_i$  are the deep neural network output and the actual  $i$ th component of the ground truth codeword (label), respectively.

After fine-tuning the NND, we integrate MDH and NND by discarding the softmax layer in the MDH network and connecting the output of the hashing layer from MDH as input to NND to create an end-to-end MDHND network. This MDHND is optimized end-to-end using the same dataset as used for fine-tuning NND and also the same cross-entropy loss function given in (7). For fine-tuning the NND and end-to-end optimization of the MDHND, we use the same 294 subjects from the 2013 year of the WVU multimodal dataset used in Step 2. The total number of facial and iris images corresponding to the 294 subjects is equal to 15,500 and 33,200, respectively. We use the Adam optimizer [20] with the default hyper-parameter values ( $\epsilon = 10^{-3}$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ) to train all the parameters. The batch size in all the experiments is fixed to 32. Our MDHND is implemented in TensorFlow with python API and all the experiments are conducted on two GeForce GTX TITAN X 12GB GPUs.

## IV. PERFORMANCE EVALUATION

### A. Details of the Code and Decoder

The intermediate binary code generated from the MDH module is considered to be the noisy codeword of some error correcting code that we can select and this noisy codeword can be decoded using the NND. We have used BCH code for our experiments. The size of the intermediate binary codes, which is the output of the MDH network depends upon the size of the error correcting code being used for the NND. We have experimented with a few different sizes of the BCH code for NND including BCH(63,45), BCH(127,85), BCH(255,187),

and BCH(511,376). We have tried to keep the code rate to be around 0.7. For the external ECC decoder, we have used the BCH decoder from the communication toolbox in MATLAB®.

### B. Evaluation Protocol for Authentication

For evaluation of authentication performance, we use a disjoint dataset of 70 subjects, which have never been seen during training with 20 facial and 20 iris images per subject leading to a total of 1400 face and iris image pairs with no repetitions. For authentication, all the 1400 pairs are forward passed through our proposed MDHND system and genuine and impostor scores are calculated using Hamming distance as our score measure. Based on the number of subjects ( $N = 70$ ) and the number of image pairs ( $t = 20$ ) per subject we obtain  $Nt(t-1)/2 = 13,300$  genuine scores and  $(N(N-1)t^2)/2 = 966,000$  imposter scores. We have used the ROC curve and the EER as our performance metric for authentication. The ROC curve is plotted between the genuine accept rate (GAR) and the false acceptance rate (FAR) at all possible values of score thresholds. EER indicates a value that the proportion of false acceptances is equal to the proportion of false rejections.

### C. Evaluation Protocol for Identification

We have also evaluated our system for identification application. The proposed system can also be used for identification as long as the database of individuals is not too large because if we make the ECC too strong, it may lead to ambiguity, where multiple users are resolved to the same codeword and the system may not be too discriminative and may not provide good performance. For identification performance, MDHND is trained on 294 mutual subjects from year 2013, and is tested on the same subjects from year 2012 of the WVU multimodal dataset. The total number of image pairs in the test set for identification is equal to 2940 corresponding to 294 subjects. For identification, the score can be calculated against each of the 294 stored codes and the class yielding the best score can be identified as the subject class. We have used the classification accuracy as our performance metric for identification. The classification accuracy is defined as the ratio of the correct predictions made to the total number of predictions for a given testing set.

### D. Baselines

We have compared our algorithms with some of the state-of-the-art score, decision, and fusion-level algorithms. CNN-Sum is a score-level fusion algorithms, which use the probability outputs for the test sample of each modality, added together to give the final score vector. CNN-Major is a decision-level fusion algorithm, which chooses the maximum number of modalities taken to be from the correct class. The feature level fusion techniques include serial feature fusion [21], parallel feature fusion [22], CCA-based feature fusion [23], and discriminant correlation analysis (DCA/MDCA) [24] methods. For the serial, parallel and CCA fusion techniques, we have used principal component analysis (PCA) [25] and

linear discriminant analysis (LDA) for dimensionality reduction followed by K-nearest neighbor (KNN) classifier [26]. To compare the results for the proposed system, with the state-of-the-art algorithms, we extract Gabor features in five scales and eight orientations for face and iris modalities. For each face, and iris image 31, 360, and 36, 630 features are extracted respectively. These hand-crafted Gabor features are used only for serial, parallel, CCA and DCA fusion techniques. For all the other algorithms being compared we have used CNN features.

We have also compared our multibiometric system with single modality system denoted as Iris-CNN and Face-CNN. As stated previously, we have implemented two different fusion architectures FCA and BLA in our MDH network. MDH-FCA/BLA denotes a stand-alone MDH network with feature level fusion using FCA/BLA and with no ECC decoder of any kind being used. For such cases in comparison, when no error correcting code is used, the length of the output feature vector is equal to the codeword length  $n$  of the  $BCH(n,k)$  code. MDH-FCA+Ext.Decoder/MDH-BLA+Ext.Decoder denotes feature level fusion using FCA/BLA in the MDH network integrated with an external conventional decoder. MDH-FCA+NND/MDH-FCA+NND denotes feature level fusion using FCA/BLA in the MDH network integrated with a non optimized NND (NND is just trained as in [12] with AWGN channel but not fine-tuned with biometric data) and also with no joint optimization. MDHND-FCA/MDHND-BLA denotes our proposed overall system with feature level fusion using FCA/BLA in the MDH network integrated and jointly optimized with a NND for biometric data.

### E. Authentication Results

Table I shows a comparison of authentication performance for our proposed system with other methods in terms of EER. Our proposed systems (MDHND-FCA/BLA) are shown in bold in the last 2 rows of the table. As it can be seen that our proposed MDHND framework outperforms single modality by at least 1.5%. It can be observed that there is an improvement in performance by using additional biometric features and the multimodality system performs better than the unimodal systems. We can also observe that there is an improvement in performance when ECC decoding is used, and that this improvement improves when the NND decoder is used. For example, for  $BCH(255,187)$  and the BLA architecture, the EER decreases from 1.45% without ECC decoding to 1.32% when a conventional decoder is used. The EER further decreases

Algorithm	BCH(63,45)	BCH(127,85)	BCH(255,187)	BCH(511,376)
CNN-Sum	1.71%	1.64%	1.32%	1.29%
CNN-Major	2.11%	2.19%	1.61%	1.53%
Iris-CNN	2.74%	2.34%	2.10%	2.04%
Face-CNN	1.91%	1.72%	1.63%	1.58%
MDH-FCA	1.84%	1.67%	1.48%	1.43%
MDH-BLA	1.76%	1.66%	1.45%	1.42%
MDH-FCA+Ext.Decoder	1.73%	1.64%	1.41%	1.39%
MDH-BLA+Ext.Decoder	1.62%	1.58%	1.32%	1.28%
MDH-FCA+NND	1.39%	1.34%	1.30%	1.23%
MDH-BLA+NND	1.30%	1.26%	1.24%	1.19%
<b>MDHND-FCA</b>	<b>1.05%</b>	<b>0.99%</b>	<b>0.90%</b>	<b>0.86%</b>
<b>MDHND-BLA</b>	<b>1.01%</b>	<b>0.94%</b>	<b>0.84%</b>	<b>0.79%</b>

TABLE I

EER FOR DIFFERENT ALGORITHMS USING DIFFERENT BCH CODES.

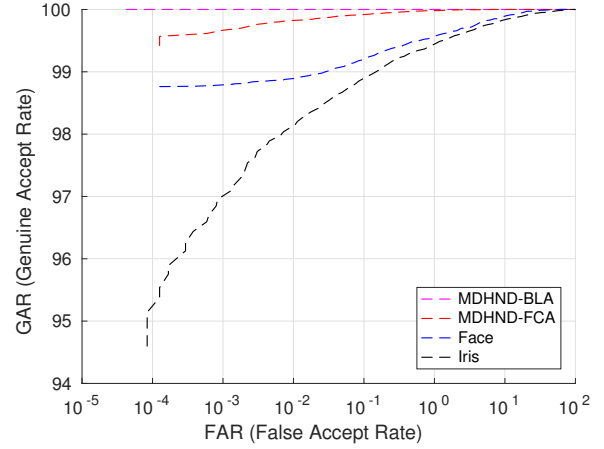


Fig. 3. ROC curves for face, iris, MDHND-FCA, and MDHND-BLA modalities using  $BCH(511,376)$

to 1.24% when an unoptimized NND is used (i.e., trained on AWGN) and to 0.84% when an optimized NND is used (i.e., trained on biometric data) for a net improvement of about 0.5% relative to the conventional decoder.

Fig. 3 shows the comparison of ROC curves for unimodal and a multimodal system. It is evident that the multimodal feature-level fusion using BLA integrated with NND and joint optimization significantly improves unimodal representation accuracy by using a bilinear formulation for exploiting the captured multiplicative interactions of the low-dimensional modality-dedicated feature representations. Fig. 4 shows the comparison of ROC curves showing the improvement in performance by using an optimized NND (MDHND-BLA) relative to a non-optimized NND (MDH-BLA+NND) and an external decoder (MDH-BLA+Ext. Decoder). It can clearly be seen that with a FAR of 0.01% we get an improvement in GAR of about 1.3% by using an optimized NND relative to a conventional external decoder.

We have also evaluated the time/delay required to perform a single authentication. We have measured the time for 1000 authentications using our trained MDHND-BLA with the same hardware as described at the end of section III-C. Based on this evaluation, the total time required for 1000 authentication is equal to 9.625 secs, which equals to an average of 9.625ms per authentication.

### F. Identification Results

Table II shows a comparison of our proposed systems with different state of the art fusion algorithms for identification task. It can clearly be seen that the performance of our proposed systems (MDHND-FCA and MDHND-BLA shown in the last two rows) in terms of classification accuracy is significantly better than previous fusion approaches. It can also be observed that using optimized NND (MDHND-BLA) definitely helps to improve the identification performance and we get an improvement of about 3% when compared to a conventional external ECC decoder (MDH-BLA+Ext.Decoder) or an improvement of about 3.5% with no decoder (MDH-BLA) being used at all.

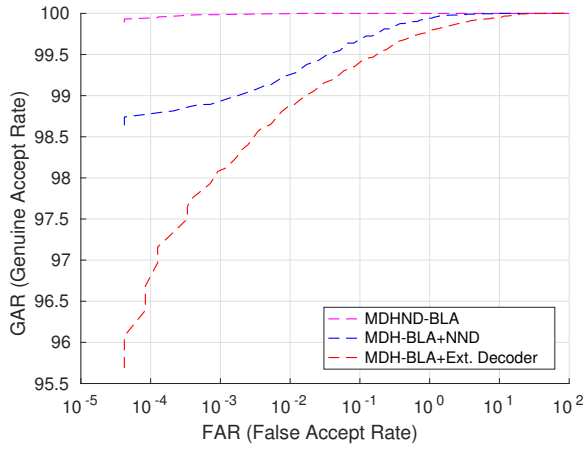


Fig. 4. ROC curves for face, iris, MDHND-FCA, and MDHND-BLA modalities using BCH(255,187) and BLA architecture

Algorithm	BCH(63,45)	BCH(127,85)	BCH(255,187)	BCH(511,376)
Serial + PCA + KNN	68.92%	70.47%	70.83%	71.12%
Serial + LDA + KNN	76.11%	78.00%	80.15%	80.52%
Parallel + PCA + KNN	71.2%	73.21%	74.1%	74.69%
Parallel + LDA + KNN	77.4%	80.19%	82.11%	82.53%
CCA + PCA + KNN	84.34%	85.12%	87.45%	87.21%
CCA + LDA + KNN	87.43%	88.65%	88.98%	89.12%
DCA/MDCa + KNN	79.08%	81.67%	82.09%	83.02%
CNN-Sum	95.13%	95.21%	95.54%	96.11%
CNN-Major	93.34%	94.16%	95.23%	95.71%
Iris-CNN	92.03%	92.39%	93.22%	94.91%
Face-CNN	94.19%	94.97%	95.06%	95.65%
MDH-FCA	94.76%	95.45%	95.82%	96.01%
MDH-BLA	94.88%	95.26%	95.98%	96.36%
MDH-FCA+Ext.Decoder	95.16%	95.88%	96.11%	96.23%
MDH-BLA+Ext.Decoder	95.22%	95.34%	96.28%	96.72%
MDH-FCA+NN	96.32%	97.22%	97.28%	97.83%
MDH-BLA+NN	96.98%	97.11%	97.29%	98.12%
<b>MDHND-FCA</b>	<b>98.02%</b>	<b>98.13%</b>	<b>99.10%</b>	<b>99.11%</b>
<b>MDHND-BLA</b>	<b>98.16%</b>	<b>98.94%</b>	<b>99.13%</b>	<b>99.23%</b>

TABLE II

CLASSIFICATION ACCURACY USING DIFFERENT BCH CODES.

## V. CONCLUSION AND FUTURE WORK

We have presented a feature-level fusion and binarization framework using deep hashing, and integrated a neural network decoder into the framework. The result is a design for a multimodal biometric system that leverages each user's multiple biometrics for authentication. In this framework, we leveraged a neural network based decoder to refine the codes generated by the deep hashing network to improve the authentication performance. We have also implemented multiple architectures for combining the biometrics at feature-level. The experimental results show that the bilinear architecture is better than linear concatenation of features. Additionally, the experimental results show that the optimized neural network decoder decreases the EER of the multimodal biometric system by 0.7%, relative to not using any decoder at all. Also, optimized neural network decoder significantly improves the authentication performance (GAR) of the multimodal biometric system by about 1.3% at an FAR of 0.01%, when compared to using an external conventional ECC decoder. The current work deals with fusion of two modalities and we plan to extend our model and use more than two modalities. Also we intend to use more recent codes such as Turbo and LDPC codes as our error-correcting code.

## REFERENCES

[1] A. Nagar, K. Nandakumar, and A. K. Jain, "Multibiometric cryptosystems based on feature-level fusion," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1, pp. 255–268, Feb. 2012.

[2] A. Ross and A. K. Jain, "Multimodal biometrics: An overview," in *Proc. European Signal Processing Conference*, Sept. 2004, pp. 1221–1224.

[3] K. Nandakumar and A. K. Jain, "Multibiometric template security using fuzzy vault," in *Proc. IEEE International Conference on Biometrics: Theory, Applications and Systems*, Oct. 2008.

[4] F. Taherkhani, N. M. Nasrabadi, and J. Dawson, "A deep face identification network enhanced by facial attributes prediction," *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[5] V. Talreja, M. C. Valenti, and N. M. Nasrabadi, "Multibiometric secure system based on deep learning," in *Proc. IEEE Global Conference on Signal and Information Processing*, Nov. 2017, pp. 298–302.

[6] Y. Xin, L. Kong, Z. Liu, C. Wang, H. Zhu, M. Gao, C. Zhao, and X. Xu, "Multimodal feature-level fusion for biometrics identification system on iomt platform," *IEEE Access*, vol. 6, pp. 21 418–21 426, March 2018.

[7] Y. Shi and R. Hu, "Rule-based feasibility decision method for big data structure fusion: Control method," *International Journal of Simulation-Systems, Science & Technology*, vol. 17, no. 31, 2016.

[8] S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson, and N. M. Nasrabadi, "Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification," in *24th International Conference on Pattern Recognition (ICPR)*, Aug 2018, pp. 3469–3476.

[9] F. Taherkhani and M. Jamzad, "Restoring highly corrupted images by impulse noise using radial basis functions interpolation," *IET Image Processing*, vol. 12, no. 1, pp. 20–30, 2018.

[10] V. Talreja, F. Taherkhani, M. C. Valenti, and N. M. Nasrabadi, "Using deep cross modal hashing and error correcting codes for improving the efficiency of attribute guided facial image retrieval," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov 2018, pp. 564–568.

[11] F. Taherkhani, V. Talreja, H. Kazemi, and N. Nasrabadi, "Facial attribute guided deep cross-modal hashing for face image retrieval," in *Proc. IEEE International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sep. 2018, pp. 1–6.

[12] E. Nachmani, Y. Be'ery, and D. Burshtein, "Learning to decode linear codes using deep learning," in *Proc. 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept. 2016, pp. 341–346.

[13] L. Lugosch and W. J. Gross, "Neural offset min-sum decoding," *Proc. IEEE International Symposium on Information Theory*, pp. 1361–1365, June 2017.

[14] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *CoRR*, vol. abs/1411.7923, Nov. 2014.

[15] "CASIA-iris-thousand." [Online]. Available: <http://biometrics.idealtest.org/>

[16] K. W. Bowyer and P. J. Flynn, "The ND-IRIS-0405 iris image dataset," *CVRL, Dept. Comp. Sci. Eng., Univ Notre Dame*, 2010.

[17] "WVU multimodal dataset." [Online]. Available: <http://biic.wvu.edu/>

[18] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.

[19] G. Sutra, B. Dorizzi, S. Garcia-Salitetti, and N. Othman, "A biometric reference system for iris (osiris)," 2013.

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

[21] C. Liu and H. Wechsler, "A shape- and texture-based enhanced fisher classifier for face recognition," *IEEE Transactions on Image Processing*, vol. 10, no. 4, pp. 598–608, April 2001.

[22] J. Yang, J. Yang, D. Zhang, and J. Lu, "Feature fusion: parallel strategy vs. serial strategy," *Pattern Recognition*, vol. 36, pp. 1369–1381, 2003.

[23] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recogn.*, vol. 38, no. 12, pp. 2437–2448, Dec. 2005.

[24] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1984–1996, Sept. 2016.

[25] A. Mackiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers and Geosciences*, vol. 19, pp. 303–342, 1993.

[26] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, January 1967.