Deep Hashing for Secure Multimodal Biometrics

Veeru Talreja, *Student Member, IEEE,* Matthew Valenti, *Fellow, IEEE,* and Nasser Nasrabadi, *Fellow, IEEE*

Abstract—When compared to unimodal systems, multimodal biometric systems have several advantages, including lower error rate, higher accuracy, and larger population coverage. However, multimodal systems have an increased demand for integrity and privacy because they must store multiple biometric traits associated with each user. In this paper, we present a deep learning framework for feature-level fusion that generates a secure multimodal template from each user's face and iris biometrics. We integrate a deep hashing (binarization) technique into the fusion architecture to generate a robust binary multimodal shared latent representation. Further, we employ a hybrid secure architecture by combining cancelable biometrics with secure sketch techniques and integrate it with a deep hashing framework, which makes it computationally prohibitive to forge a combination of multiple biometrics that passes the authentication. The efficacy of the proposed approach is shown using a multimodal database of face and iris and it is observed that the matching performance is improved due to the fusion of multiple biometrics. Furthermore, the proposed approach also provides cancelability and unlinkability of the templates along with improved privacy of the biometric data. Additionally, we also test the proposed hashing function for an image retrieval application using a benchmark dataset. The main goal of this paper is to develop a method for integrating multimodal fusion, deep hashing, and biometric security, with an emphasis on structural data from modalities like face and iris. The proposed approach is in no way a general biometrics security framework that can be applied to all biometrics modalities, as further research is needed to extend the proposed framework to other unconstrained biometric modalities.

Index Terms—Channel coding, hashing, multibiometrics, secure-sketch, template security.

I. INTRODUCTION

BIOMETRICS are difficult to forge, and unlike in traditional password-based access control systems, they do not have to be remembered. As much as these characteristics provide an advantage, they also create challenges related to protecting biometrics in the event of identity theft or a database compromise as each biometric characteristic is distinct and cannot be replaced by a newly generated arbitrary biometric. There are serious concerns about the security and privacy of an individual because of the proliferation of biometric usage. These concerns cannot be alleviated by using conventional cryptographic hashing as in case of alpha-numeric passwords because the cryptographic hashes are extremely sensitive to noise and are not suitable for the protection of biometrics due to inherent variability and noise in biometric measurements.

The leakage of biometric information to an adversary constitutes a serious threat to security and privacy because if an adversary gains access to a biometric database, he can potentially obtain the stored user information. The attacker can use this information to gain unauthorized access to the system by reverse engineering the system and creating a physical spoof. Furthermore, an attacker can abuse the biometric information for unintended purposes and violate user privacy [1].

Multimodal biometric systems use a combination of different biometric traits such as face and iris, or face and fingerprint. Multimodal systems are generally more resistant to spoofing attacks [2]. Moreover, multimodal systems 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. Multimodal systems also have an advantage of lower error rates and higher accuracy when compared to unimodal systems [1]. Consequently, multimodal systems have been deployed in many large scale biometric applications including the FBI's Next Genration Identification (NGI), the Department of Homeland Security's US-VISIT, and the Government of India's UID. However, Multimodal systems have an increased demand for integrity and privacy because the system stores multiple biometric traits of each user. Hence, multimodal template protection is the main focus of this paper.

The fundamental challenge in designing a biometric template protection scheme is to manage the intra-user variability that occurs due to signal variations in the multiple acquisitions of the same biometric trait. With respect to biometric template protection, four main architectures are widely used: *fuzzy commitment, secure sketch, secure multiparty computation, and cancelable biometrics* [3]. Fuzzy commitment and secure sketch are biometric cryptosystem methods and are usually implemented with error correcting codes and provide information-theoretic guarantees of security and privacy (e.g., [4]–[8]). Secure multiparty computation architectures are distance based and use cryptographic tools. Cancelable biometrics use revocable and non-invertible user-specific transformations for distorting the enrollment biometric (e.g., [9]–[12]), with the matching typically performed in the transformed domain.

For a template to be secure, it must satisfy the important properties of *noninvertibility* and *revocability*. Noninvertibility implies that given a template, it must be computationally difficult to recover the original biometric data from the template. Revocability implies that if a template gets compromised, it should be possible to revoke the compromised template and generate a new template using a different transformation. Moreover, it should be difficult to identify that the new template and the old compromised template are generated from the same underlying biometric data.

One important issue for multimodal systems is that the multiple biometric traits generally do not have the same feature-

level representation. Furthermore, it is difficult to characterize multiple biometric traits using compatible feature-level representations, as required by a template protection scheme [1]. To counter this issue there have been many fusion techniques for combining multiple biometrics [1], [13], [14]. One possible approach is to apply a separate template protection scheme for each trait followed by decision-level fusion. However, such an approach may not be highly secure, since it is limited by the security of the individual traits. This issue motivated our proposed approach of using multimodal biometric security to perform a joint feature-level fusion and classification.

Another important issue is that biometric cryptosystem schemes are usually implemented using error control codes. In order to apply error control codes, the biometric feature vectors must be quantized, for instance by binarizing. One method of binarizing the feature vectors is thresholding the feature vectors, for example, by thresholding against the population mean or thresholding against zero. However, thresholding causes a quantization loss and does not preserve the semantic properties of the data structure in Hamming space. In order to avoid thresholding and minimize the quantization loss, we have used the idea of *hashing* [15], [16], which is used in the image and data retrieval literature to achieve fast search by binarizing the real-valued image features. The basic idea of hashing is to map each visual object into a compact binary feature vector that approximately preserves the data structure in the original space. Owing to its storage and retrieval efficiency, hashing has been used for large scale visual search and image retrieval.

Recent progress in image classification, object detection, face recognition, speech recognition and many other computer vision tasks demonstrates the impressive learning ability of *convolutional neural networks* (CNN). The robustness of features generated by the CNN has led to a surge in the application of deep learning for generating binary codes from raw image data. Deep hashing [17]–[20] is the technique of integrating hashing and deep learning to generate compact binary vectors from raw image data. There is a rich literature related to the application of optimized deep learning for converting the raw image data to binary hash codes.

Inspired by the recent success of deep hashing methods, the objective of this work is to examine the feasibility of integrating deep hashing with a secure architecture to generate a secure multimodal template for face and iris biometrics. Contributions include:

- We use deep hashing to generate a binary latent shared representation from a user's face and iris biometrics.
- We combine cancelable biometrics and secure sketch schemes to create a hybrid secure architecture.
- We integrate the hybrid secure architecture with the deep hashing framework to generate a multimodal secure sketch, which is cryptographically hashed to generate the secure multimodal template.
- We analyze the trade-off between genuine accept rate (GAR) and security for the proposed secure multimodal scheme using an actual multimodal database.
- Additionally, we also perform an information-theoretic privacy analysis, and unlinkability analysis for the proposed secure system.

The proposed approach represents a biometric security framework integrated with multimodal fusion and deep hashing, and is particularly well suited for structural data from modalities like face and iris. Our approach is not a general biometric security framework that can be applied to all biometric modalities, but rather a proposal that needs further study and validation.

The rest of the paper is organized as follows. Section II provides a background on deep hashing techniques and the various multibiometric template security schemes proposed in the literature. The proposed framework and the associated algorithms are introduced in Section III. Implementation details are presented in Section IV. In Section V, we present a performance evaluation of the cancelable biometric module, which is a part of the overall proposed system. The performance evaluation of the overall proposed system is discussed in Section VI. The conclusions are summarized in Section VII.

II. RELATED WORK

A. Deep Learning

Deep learning has emerged as a new area of machine learning and is being extensively applied to solve problems that have resisted the best attempts of the machine learning and artificial intelligence community for many years. It has turned out to be very good at discovering intricate structures in high-dimensional data and is therefore applicable to many domains of science, business, and government.

Deep learning has been extensively implemented and applied to image recognition tasks. Krizhevsky *et al.* [21] provided a breakthrough in the field of object recognition and ImageNet classification by applying a CNN for object recognition. They were able to reduce the error rate by almost half. The neural network implemented in [21] is currently known as *AlexNet* and triggered the rapid endorsement of deep learning by the computer vision community. Simonyan *et al.* [22] increased the depth of the convolutional network but reduced the size of the filters being used for convolution. The main contribution in [22] was a thorough evaluation of networks of increasing depth using an architecture with very small 3×3 convolution filters, which represented a compelling advancement over the prior-art configurations.

Szegedy *et al.* [23] advanced the architecture of CNN by making it deeper, similar to [21], and wider by introducing a CNN termed *inception*. One particular incarnation of this architecture is known as *GoogleNet* which is 22 layers deep. He *et al.* [24] developed a very deep 152 layer convolutional neural network architecture named *ResNet*. The novelty of *ResNet* lies not only in creating a very deep network but also in the use of a residual architecture to reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions.

In addition to improving performance in image and speech recognition [21], [22], [24], [25], deep learning has produced extremely promising results for various tasks in natural language understanding, particularly topic classification, sentiment analysis, question answering, and language translation.

B. Deep Hashing

Many hashing methods [16], [26]–[30] have been proposed to enable efficient approximate nearest neighbor search due to low space and time complexity. These traditional hashing methods can be categorized into data-independent or data-dependent methods. A comprehensive survey of hashing techniques is presented in [31]. Initial research on hashing was mainly focused on data-independent methods, such as locality sensitive hashing (LSH). LSH methods [15] generate hashing bits by using random projections. However, LSH methods demand a significant amount of memory as they require long codes to achieve satisfactory performance.

To learn compact binary codes, data-dependent hashing methods have been proposed in the literature. Data-dependent methods learn similarity-preserving hashing functions from a training set. Data-dependent hashing methods can be categorized as *unsupervised* [16], [32], [33] or *supervised* [28], [29]. These methods have achieved success to some extent by using handcrafted features for learning hash functions. However, the handcrafted features do not preserve the semantic data similarities of image pairs and non-linear variation in real-world data [20]. This has led to a surge of deep hashing methods [17]–[20], [34], [35] where deep neural networks encode non-linear hash functions. This leads to an effective end-to-end learning of feature representation and hash coding.

Xia *et al*. [17] adopted a two-stage learning strategy wherein the first stage computes hash codes from the pairwise similarity matrix and the second stage trains a deep neural network to fit the hash codes generated in the first stage. The model proposed by Lai *et al*. [18] simultaneously captures the intermediate image features and trains the hashing function in a joint learning process. The hash function in [18] uses a divide-and-encode module, which splits the image features derived from the deep network into multiple blocks, each block encoded into one hash bit. Liu *et al*. [20] present a deep hashing model that learns the hash codes by simultaneously optimizing a contrastive loss function for input image pairs and imposing a regularization on the real-valued outputs to approximate the binary values. Zhu *et al*. [36] proposed a deep hashing method to learn hash codes by optimizing a pairwise cross-entropy quantization loss to preserve the pairwise similarity and minimize the quantization error simultaneously.

C. Secure Biometrics

The leakage of biometric template information to an adversary constitutes a serious threat to security and privacy of the user because if an adversary gains access to the biometric database, he can potentially obtain the stored biometric information of a user. To alleviate the security and privacy concerns in biometric usage, secure biometric architectures have been developed to allow authentication without requiring that the reference biometric template be stored in its raw format at the access control device. Secure biometric architectures include *biometric cryptosystems* (e.g., fuzzy commitment and secure sketch) [4], [5], [7], [8] and *transformation based methods* (e.g., cancelable biometrics) [3].

Fuzzy commitment, a classical method of biometric protection, was first proposed in 1999 [5]. Forward error correction (FEC) based fuzzy commitment can also be viewed as a method of extracting a secret code by means of polynomial interpolation [6]. An implementation example of such a fuzzy commitment scheme appears in [8], wherein a BCH code is employed for polynomial interpolation; experiments show that when the degree of the interpolated polynomial is increased, the matching becomes more stringent, reducing the false accept rate (FAR), but increasing the false reject rate (FRR).

Cancelable biometrics was first proposed by Ratha *et al*. [9], after which, there have been various different methods of generating cancelable biometric templates. Some popular methods use non-invertible transforms [9], bio-hashing [10], salting [11] and random projections [12]. Literature surveys on cancelable biometrics can be found in [3], and [37].

D. Secure Multimodal Biometrics

The secure biometric frameworks have been extended to include multiple biometric traits of a user [1], [13], [14], [38]. In [13] face and fingerprint templates are concatenated to form a single binary string and this concatenated string is used as input to a secure sketch scheme. Kelkboom *et al*. [39] provided results for decision-level, feature-level, and score-level fusion of templates by using the number of errors corrected in a biometric cryptosystem as a measure of the matching score.

Nagar *et al*. [1] developed a multimodal cryptosystem based on feature-level fusion using two different security architectures, *fuzzy commitment* and *fuzzy vault*. Fu *et al*. [40] theoretically analyzed four different versions of the multibiometric cryptosystem: *no-split*, *MN-split*, *package*, and *biometric model*, using template security and recognition accuracy as performance metrics. In the first three versions, the biometric templates are secured individually with a decision-level fusion, while the last version is a feature-level fusion.

Research has also been directed towards integrating cancelable biometric techniques into multimodal systems. Canuto *et al*. [38] combined voice and iris using cancelable transformations and decision level fusion. Paul and Gavrilova [41] used random projections and transformation-based feature extraction and selection to generate cancelable biometric templates for face and ear. There are some studies related to the use of multi-feature biometric fusion, which involves combining different features of the same biometric trait [42].

However, none of the above papers present a secure architecture that combines multiple secure schemes to protect multiple biometrics of a user. In this paper, we have integrated a deep hashing framework with a hybrid secure architecture by combining cancelable biometric templates and secure sketch, which makes it computationally prohibitive to forge a combination of multiple biometrics that passes the authentication.

III. PROPOSED SECURE MULTIBIOMETRIC SYSTEM

A. System Overview

In this section, we present a system overview including descriptions of the enrollment and authentication procedures. We propose a feature-level fusion and hashing framework for the secure multibiometric system. The general framework for the proposed secure multibiometric system is shown in Fig. 1. During enrollment, the user provides their biometrics (e.g.,

face and iris) as an input to the deep feature extraction and binarization (DFB) block. The output of the DFB block is an J-dimensional binarized joint feature vector $\mathbf{e}$. A random selection of feature components (bits) from the binarized joint feature vector $\mathbf{e}$ is performed. The number of random components that are selected from the binarized joint feature vector $\mathbf{e}$ is G. The indices of these randomly selected G components forms the enrollment key $\mathbf{k_e}$, which is given to the user. The cancelable multimodal template $\mathbf{r_e}$ is formed by selecting the values from the vector $\mathbf{e}$ at the corresponding location or indices as specified by the user-specific key $\mathbf{k_e}$.

This random selection of G components from the binarized joint feature vector $\mathbf{e}$ helps in achieving revocability, because if a key is compromised, a new key can be issued with a different set of random indices. In the next step, $\mathbf{r_e}$ is passed through a forward error correction (FEC) decoder to generate the multimodal sketch $\mathbf{s_e}$. The cryptographic hash of this sketch $f_{\mathsf{hash}}(\mathbf{s_e})$ is stored as a secure template in the database.

During authentication, the probe user presents the biometrics and the key $\mathbf{k_p}$ where $\mathbf{k_p}$ could be same as the enrollment key $\mathbf{k_e}$ in the case of a genuine probe or it could be a synthesized key in case of an impostor probe. The probe biometrics are passed through the DFB block to obtain a binary vector $\mathbf{p}$, which is the joint feature vector corresponding to the probe. Using the key $\mathbf{k_p}$ provided by the user, the multimodal probe template $\mathbf{r_p}$ is generated by selecting the values from $\mathbf{p}$ at the locations given by the key $\mathbf{k_p}$. In the next step, $\mathbf{r_p}$ is passed through a FEC decoder with the same code used during enrollment to generate the probe multimodal sketch $\mathbf{s_p}$. If the cryptographic hash of the enrolled sketch $f_{\mathsf{hash}}(\mathbf{s_e})$ matches the cryptographic hash of the probe sketch $f_{\mathsf{hash}}(\mathbf{s_p})$, then the access is granted, otherwise the access is denied.

The proposed secure multibiometric system consists of two basic modules: *Cancelable Template Module (CTM)* and *Secure Sketch Template Module (SSTM)*, which are described more fully in the following subsections.

B. Cancelable Template Module

The cancelable template module (CTM) consists of two blocks: DFB block and random-bit selection block. The primary function of CTM is non-linear feature extraction, fusion, and binarization using the proposed DFB architecture shown in Figs. 2 and 3. The DFB consists of two layers: domain-specific layer (DSL) and joint representation layer (JRL).

1) Domain-Specific Layer: The DSL consists of a CNN for encoding the face ("Face-CNN") and a CNN for encoding the iris ("Iris-CNN"). For each CNN, we use VGG-19 [22] pre-trained on ImageNet [43] as a starting point and then fine-tune it with an additional fully connected layer *fc3* as described in Sec. IV-B and IV-C. There are multiple reasons for using VGG-19 pre-trained on the ImageNet dataset for encoding the face and iris. In the proposed method, the VGG-19 is only used as feature-extractor for face and iris modalities. It can be seen from the previous literature [44]–[49] that the features provided by a VGG-19 pre-trained on ImageNet and fine-tuned on face/iris images are very discriminative and therefore can be used for face/iris recognition. Moreover, starting with a well-known architecture and using the same architecture for both modalities makes the work highly reproducible.
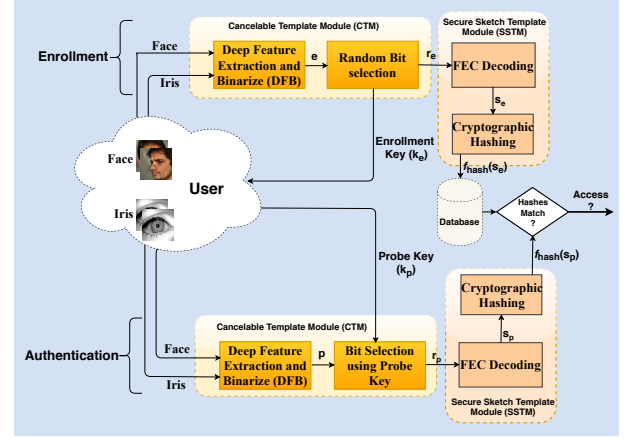

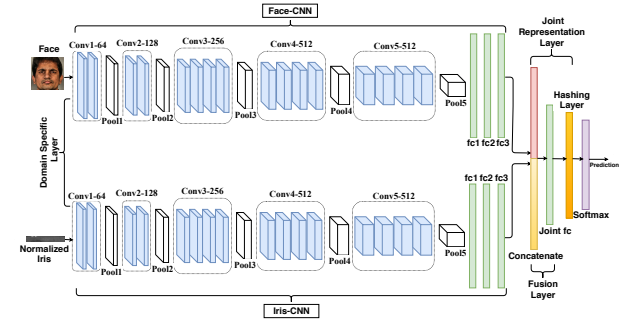Fig. 1: Block diagram of the proposed system.

Fig. 2: Proposed deep feature extraction and binarization (DFB) model for the fully concatenated architecture (FCA).

2) Joint Representation Layer: The output feature vectors of the Face-CNN and Iris-CNN are fused and binarized in the JRL, which is split into two sub-layers: fusion layer and hashing layer. The main function of the fusion layer is to fuse the individual face and iris representations from domain-specific layers into a shared 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: (1) fully concatenated architecture (FCA), and (2) bilinear architecture (BLA). These two architectures differ in the way the face and iris feature vectors are fused together to generate the joint feature vector.

In the FCA shown in Fig. 2, the outputs of the Face-CNN and Iris-CNN are concatenated vertically using a concatenation layer. The concatenated feature vector is passed through a fully connected layer (hereon known as *joint fully connected layer*) which reduces the feature dimensionality (i.e., the number of dimensions is reduced) and also fuses the iris and face features. In the FCA, the concatenation layer and the joint fully connected layer together constitute the fusion layer.

In the BLA shown in Fig. 3, the outputs of the Face-CNN and Iris-CNN are combined using the matrix outer product; i.e., the bilinear feature combination of column face feature vector $\mathbf{f_{\mathsf{face}}}$ and column iris feature vector $\mathbf{f_{\mathsf{iris}}}$ given by $\mathbf{f_{\mathsf{face}}}\mathbf{f_{\mathsf{iris}}}^{T}$. Similar to the FCA, the bilinear feature vector is also passed through a joint fully connected layer. In the BLA, the outer product layer and the joint fully connected layer together
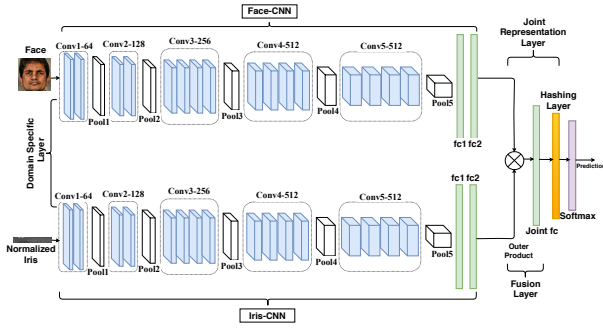
Fig. 3: Proposed deep feature extraction and binarization (DFB) model for the bilinear architecture (BLA).

constitute the fusion layer.

In addition to the two techniques (FCA, BLA) used in this paper, there could be other fusion techniques for combining multiple modalities [50]. The rationale behind implementing FCA is that we wanted to use a fusion technique that involves just simple concatenation where there is no interaction between the two modalities being fused before the joint fully connected layer (Joint f_c). As evident from Fig. 2, the iris and face extracted features do not interact with each other and have their own network parameters before passing through the joint fully connected layer. On the other hand, we also wanted to test a fusion technique that involves high interactions between the two modalities feature vectors at every element before being passed through the joint fully connected layer. That is the reason we have used BLA, which is based on *bilinear fusion* [51]. Bilinear fusion exploits the higher-level dependencies of the modalities being combined by considering the pairwise multiplicative interactions between the modalities at each feature element (i.e., matrix outer product of modalities feature vector). Moreover, bilinear fusion is widely being used in many CNN applications such as fine-grained visual recognition and video action recognition [50], [51].

Hashing layer: The output of the fusion layer produces a J-dimensional shared multimodal feature vector of real values. We can directly binarize the output of the fusion layer by thresholding at any numerical value or 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 and 3). The main function of the hashing layer is to binarize (hash) the shared multimodal feature representation generated by the fusion layer.

One key challenge of implementing deep learning to hash end-to-end is converting deep representations, which are real-valued and continuous, to exactly binary codes. The sign activation function $h = \text{sgn}(z)$ can be used by the hashing layer to generate the binary hash codes. 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. The problem of zero gradient at the hashing layer due to a non-smooth sign activation can be diminished by using the idea of continuation methods [52].

We circumvent the zero-gradient problem by starting with a smooth activation function $y = \tanh(\beta x)$ and making it sharper by increasing the bandwidth β as the training proceeds. We have utilized a key relationship between the sign activation function and the scaled tanh function using limits:

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

where $\beta > 0$ is a scaling parameter. The scaled function $\tanh(\beta x)$ will become sharper and more saturated as we increase β during training. Eventually, this non-smooth tanh function with $\beta \to \infty$ converges to the original, difficult to optimize, sign activation function. 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 β while holding other training parameters equal to the previously converged network parameters, and start retraining the network for convergence. This process is repeated several times by increasing the bandwidth of the tanh activation as $\beta \to \infty$ until the hashing layer can generate binary codes. In addition to using this continuation method for training the network, we have used additional cost functions for efficient binary codes. The overall objection function used for training is discussed in Sec. IV-A

*3) **Random-Bit Selection***: One of the most prevalent methods for generating cancelable template involves random projections of the biometric feature vector [12], in which the random projection is a revocable transformation. Similarly, the DFB architecture is considered to be the projection of the biometric images in a J-dimensional space. The randomness and revocability is added by performing a random bit selection of G bits from the J-dimensional output vector $\mathbf{e}$ of the DFB. After the selection, these random bits are then arranged in descending order of reliability. The reliability of each bit is computed as $((1 - p_g^e)p_i^e)$, where p_i^e and p_g^e are the impostor and genuine bit error probabilities, respectively [1]. A different set of random bits is selected for every user and these randomly selected G bits form the cancelable multimodal template $\mathbf{r_e}$ and the indices of the selected bits forms the key for that user $\mathbf{k_e}$. This key is revocable and a new set of random bits can be selected in case the key gets compromised. Selecting a new set of bits requires that either the original vector $\mathbf{e}$ be retrieved from a secure location or else the user is re-enrolled, thereby presenting a new instance of $\mathbf{e}$. This method of using the DFB architecture with a random bit selection is analogous to a random projection as a revocable transformation to generate a cancelable template [12].

It is important to note that even if multiple users end up having the same key $\mathbf{k_e}$ (i.e., same indices of G random bits), their final templates will still be distinct because the template depends on the values at those G bits (i.e., $\mathbf{r_e}$) from the enrollment vector $\mathbf{e}$, and not only on the indices of the G bits. A second user having the same key $\mathbf{k_e}$ is equivalent to the stolen key scenario, which is analyzed in Sec. V-B.

C. Secure Sketch Template Module

As shown in Fig. 1, the cancelable template (output of CTM) $\mathbf{r_e}$ is an intermediate template and is not stored in the database. The cancelable template is passed through the

SSTM to generate the secure multimodal template, which is stored in the database. As the name suggests, the SSTM module is related to the secure sketch biometric template protection scheme. The SSTM contains two important blocks: FEC decoding and cryptographic hashing. The main function of the SSTM is to generate a multimodal secure sketch by using the cancelable template as an input to the FEC decoder. This multimodal secure sketch (output of the FEC decoder) is cryptographically hashed to generate the secure multimodal template, which is stored in the database.

The FEC decoding implemented in our framework is the equivalent of a secure-sketch template protection scheme. In a secure-sketch scheme, sketch or helper data is generated from the user's biometrics and this sketch is stored in the access-control database. A common method of implementing secure sketch is to use error control coding. In this method error control coding is applied to the biometrics or the feature vector to generate a sketch which is stored in the database. Similarly, in our framework, the FEC decoding is considered to be the error control coding part required to generate the secure sketch. Our approach is different from other secure sketch approaches using error correcting codes (ECC) as we do not have to present any other side information to the decoder like a syndrome or a saved message key [53].

The cancelable template $\mathbf{r_e}$ from the CTM is considered to be the noisy codeword of an ECC that we can select. This noisy codeword is decoded with a FEC decoder and the output of the decoder is the multimodal secure sketch $\mathbf{s_e}$ that corresponds to the codeword closest to the cancelable template. This multimodal sketch $\mathbf{s_e}$ is cryptographically hashed to generate $f_{\text{hash}}(\mathbf{s_e})$ stored in the database.

During authentication, the same process is performed. The probe user provides the biometrics and the key which are used to generate the probe template $\mathbf{r_p}$. The probe template $\mathbf{r_p}$ is passed through an FEC decoder for the same error correcting code used during the enrollment. The output of the FEC decoder is the probe multimodal sketch $\mathbf{s_p}$ which is cryptographically hashed and access is granted only if this hash matches the enrolled hash. During authentication, if it is a genuine probe, the enrollment $\mathbf{r_e}$ and the probe vector $\mathbf{r_p}$ would usually decode to the same codeword in which case the hashes would match and access would be granted.

IV. IMPLEMENTATION

A. Objective Function for Training the Deep Hashing Network

In this section, the objective function used for training the deep hashing network is described.

Semantics-preserving binary codes: In order to construct semantics-preserving binary codes, we propose to model the relationship between the labels and the binary codes. Every input image is associated with a semantic label, which is derived from the hashing layer's binary-valued outputs, and the classification of each image is dependent on these binary outputs. Consequently, we can ensure that semantically similar images belonging to the same subject are mapped to similar binary codes through an optimization of a loss function defined on the classification error. The classification formulation has been incorporated into the deep hashing framework by adding the *softmax* layer as shown in Fig. 2 and Fig. 3. Let E_1 denote the objective function required for classification formulation:

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

where the first term $L_n(.)$ 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 λ governs the relative importance of the regularization.

The choice of the loss function $L_n(.)$ depends on the application itself. We use a classification loss function that uses softmax outputs by minimizing the cross-entropy error function. Let the predicted softmax output $f(x_n, \mathbf{w})$ be denoted by $\hat{y}_n$. The classification loss for the nth training instance is:

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

where $y_{n,m}$ and $\hat{y}_{n,m}$ is the ground truth and the prediction result for the mth unit of the nth training instance, respectively and M is the number of output units.

Additional cost constraints for efficient binary codes: The continuation method that has been described in III-B2 forces the activations of the hashing layer closer to -1 and 1. However, we need to include additional cost constraints to obtain more efficient binary codes.

Let the J-dimensional vector output of the hashing layer be denoted by $\mathbf{o}_n^H$ for the n-th input image, and let the i-th element of this vector be denoted by $o_{n,i}^H (i = 1, 2, 3, \cdots, J)$. 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 maximizing the sum of squared errors 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. However, this is equivalent to maximizing the square of the length of the vector formed by the hashing layer activations, that is $\sum_{n=1}^{N} ||\mathbf{o}_n^H - \mathbf{0}||^2 = \sum_{n=1}^{N} ||\mathbf{o}_n^H||^2$. Let $E_2(\mathbf{w})$ denote this constraint to boost the activations of units in hashing layer to be closer to -1 or 1:

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

In addition to forcing the codes to become binarized, we also require that the codes satisfy a balance property whereby they produce an equal number of -1's and 1's, which maximizes the entropy of the discrete distribution and results in hash codes with better discrimination. To achieve the balance property, we want each bit to fire 50% of the time by minimizing the sum of the squared error between the mean of the hashing layer activations and 0. This is given by $\sum_{n=1}^{N}(\text{mean}(\mathbf{o}_n^H) - 0)^2$, which is equivalent to $\sum_{n=1}^{N}(\text{mean}(\mathbf{o}_n^H))^2$ where mean(.) computes the average of the elements of the vector. This criterion helps to obtain binary

codes with an equal number of -1's and 1's. 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:

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

Combining the above two constraints (binarizing and balance property constraints) makes $\mathbf{o}_n^H$ close to a length J binary string with a 50% chance of each bit being -1 or 1.

Overall objective function: The overall objective function to be minimized for a semantics-preserving efficient binary codes is given as:

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

where α, β, and γ are the tuning parameters of each term. The optimization to be performed to minimize the overall objective function is given as:

$$\mathbf{w} = \arg\min_{\mathbf{w}} (\alpha E_1(\mathbf{w}) + \beta E_2(\mathbf{w}) + \gamma E_3(\mathbf{w})) \quad (7)$$

The optimization given in (7) is the sum of the losses form and can be performed via the stochastic gradient descent (SGD) efficiently by dividing the training samples into batches. For training the JRL we adopt a two-step training procedure where we first train only the JRL using the objective function in (6) greedily with softmax by freezing the Face-CNN and Iris-CNN. After training the JRL, the entire model is fine-tuned end-to-end using the same objective function with back-propagation at a relatively small learning rate.

For tuning the hyper-parameters α, β, and γ of the objective function (6), we have utilized an iterative grid search. To start, consider a cubic grid with all possible values for each parameter. Each point on this grid (α,β,γ) represents a combination of the three hyper-parameters. Because exhaustively searching over all combinations is computationally expensive, we adopted an iterative and adaptive grid search.

In the iterative and adaptive grid search, for each hyper-parameter, we considered the set of values $\mathcal{S} = \{1, 2i\}$ for $i = \{1, ..., 15\}$; i.e., the set containing 1 and all positive even integers from 2 to 30. This grid search is performed iteratively, where each iteration is a combination of 3 steps. In the first step, we fixed α, and γ to be 1 and β is chosen from the set $\mathcal{S}$. Therefore the set of points considered for this step is:

$$(\alpha, \beta, \gamma) = (1, \beta_i, 1), \text{where } \beta_i \in \mathcal{S}. \quad (8)$$

For each point in the above set $(1, \beta_i, 1)$, we trained our DFB network and calculated the genuine accept rate (GAR) for the overall system for a security of 104 bits using a 5-fold cross validation. Using this method, we found the best value for hyper-parameter β that gave us the highest GAR with the values of α and γ as 1. This best value of β will be denoted as β^t where the superscript t signifies the iteration number.

In the second step, we repeated the same process with α and β fixed at 1 and choosing γ from the set $\mathcal{S}$:

$$(\alpha, \beta, \gamma) = (1, 1, \gamma_i), \text{where } \gamma_i \in \mathcal{S}. \quad (9)$$

Again using a 5-fold cross validation, we found the best value for hyper-parameter γ, which is denoted by γ^1, that gave us the highest GAR with the values of α and β fixed as 1. In the third step, the same procedure was performed by keeping β,

and γ fixed at 1 and found the best value for hyper-parameter α, which is denoted by α^1, from the set $\mathcal{S}$. These three steps together complete one iteration of the iterative grid search.

In the next iteration, we again performed the above 3 steps but instead of fixing the values of the two parameters to 1, we fixed the value of the two parameters to be the best value found in the previous iteration for those parameters. To explain this, consider the best value of the 3 parameters found in the first iteration, denoted by $\alpha^1, \beta^1, \gamma^1$. In the first step of the second iteration, we fixed α, and γ to be α^1 and γ^1 respectively and chose β from the set $\mathcal{S}$. Therefore the set of points are:

$$(\alpha, \beta, \gamma) = (\alpha^1, \beta_i, \gamma^1), \text{where } \beta_i \in \mathcal{S}. \quad (10)$$

Again, using a 5-fold cross validation, we found the best value for hyper-parameter β with the other parameters set to α^1 and γ^1. This best value of β will be denoted as β^2 since this is the second iteration. Similarly, we performed the second and third steps of the second iteration to find the γ^2 and α^2 respectively.

We continued performing these iterations until the parameters converged, which implies that the best value of each parameter did not change from one iteration to the other; i.e., $\alpha^t = \alpha^{t-1}, \beta^t = \beta^{t-1}, \gamma^t = \gamma^{t-1}$.

Using the above procedure for hyperparameter tuning, we have found the values of α^t, β^t, and γ^t to be 8, 2, 2 for FCA and 6, 4, 2 for BLA respectively. The importance of each term will be further discussed in the ablation study in Section VI-D.

B. Network parameters for the Face-CNN

The network used for the Face-CNN is the VGG-19 with an added fully connected layer *fc3* (shown in Fig. 2). The Face-CNN is fine-tuned end-to-end with the CASIA-Webface [54], which contains 494,414 facial images corresponding to 10,575 subjects. After fine-tuning with CASIA-Webface, the Face-CNN is next fine-tuned with the 2013 session of the WVU-Multimodal face 2012-21013 dataset [55]. The WVU-Multimodal face dataset for the year 2012 and 2013 together contain a total of 119,700 facial images corresponding to 2263 subjects with 294 common subjects. All the raw facial images are first aligned in 2-D and reduced to a fixed size of 224×224 before passing through the network [56]. 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 batch size was set to 40, and the momentum to 0.9. The training was regularized by weight decay (the L2 penalty multiplier set to 0.0005) and dropout regularization for the first three fully-connected layers (dropout ratio set to 0.5). We used batch normalization for fast convergence. The learning rate was initially set to 0.1, and then decreased to 90% of its value every 10 epochs. The number of nodes in the last fully connected layer *fc3* before the softmax layer is 1024 for the FCA and 64 for the BLA. This implies that the feature vector extracted from Face-CNN and fused with the feature vector from Iris-CNN has 1024 dimensions for the FCA and 64 for the BLA.

C. Network parameters for the Iris-CNN

The network used for the Iris-CNN is the VGG-19 with an added fully connected layer *fc3*. First, the Iris-CNN has been fine-tuned end-to-end using the combination of CASIA-Iris-Thousand [57] and ND-Iris-0405 [58] with about 84,000 iris images corresponding to 1355 subjects. Next, the Iris-CNN is fine-tuned using the 2013 session of the WVU-Multimodal iris 2012-21013 dataset [55]. The WVU-Multimodal iris dataset for the year 2012 and 2013 together contain a total of 257,800 iris images corresponding to 2263 subjects with 294 common subjects. All the raw iris images are segmented and normalized to a fixed size of 64×512 using Osiris (Open Source for IRIS) which is an open source iris recognition system developed in the framework of the BioSecure project [59]. There is no other pre-processing for the iris images. The other hyper-parameters are consistent with the fine-tuning of the Face-CNN. The iris network has an output of 1024 for FCA and 64 for BLA.

D. Network parameters for the Joint Representation Layer

The details of the network parameters for the two JRL architectures are discussed in this subsection:

1) Fully Concatenated Architecture: In the FCA, the 1024-dimensional outputs of the Face-CNN and Iris-CNN are concatenated vertically to give a 2048-dimensional vector. The concatenated feature vector is then passed through a fully connected layer which reduces the feature dimensionality from 2048 to 1024 and also fuses the iris and face features. The hashing layer is also a fully connected layer that outputs a 1024-dimensional vector and includes a tanh activation.

For the training of the DFB model, we have used a two-step training procedure. First, only the JRL was trained for 65 epochs, a batch size of 32. The learning rate initially set to 0.1, and then decreased to 90% of its value every 20 epochs. The other hyperparameters are consistent with the fine-tuning of the Face-CNN. After training of the joint representation layer, the entire DFB model was fine-tuned end-to-end for 25 epochs on a batch size of 32. The learning rate initialized to 0.07 which is the final learning rate in the training process of the joint fully connected layer in the first step. The learning rate was decreased to 90% of its value every 5 epochs. For this two-step training process, we have used the 2013 session of the overlap subjects in the 2012 and 2013 sessions from the WVU-Multimodal dataset. This common subset consists of 294 subjects with a total of 18700 face and 18700 iris images with the same number of face and iris images per subject.

2) Bilinear architecture: For the BLA, we do not add *fc3* (i.e., the additional fully connected layer) to either the Face-CNN or the Iris-CNN. In addition, the number of nodes in the first and second fully connected layers $fc1$ and $fc2$ are reduced to 512 and 64, respectively. This means that the output feature vector of the face and iris networks have 64 dimensions rather than the 1024 dimensions of the FCA. The 64-dimensional outputs of the Face-CNN and Iris-CNN are combined in the bilinear (outer product) layer using the matrix outer product as explained in Sec. III-B2. The bilinear layer produces an output of dimension $64 \times 64 = 4096$ fusing the iris and face features. The bilinear feature vector is then passed through a fully connected layer, which reduces the feature dimension from 4096 to 1024 followed by a hashing layer which produces a binary output of 1024 dimensions.

In the first step of the two-step training process, only the joint representation layer was trained for 80 epochs on a batch size of 32. The momentum was set to 0.9. The learning rate was initially set to 0.1, and then decreased by a factor of 0.1 every two epochs. The other hyperparameters and the input image sizes are consistent with the training process used in FCA. After training of the joint representation layer, the entire DFB model was fine-tuned for 30 epochs on a batch size of 32. The learning rate was initialized to 0.0015 which is the final learning rate in the training process of the joint representation layer in the first step. The learning rate was decreased by a factor of 0.1 every five epochs. The other hyper-parameters are consistent with the training of the JRL in FCA.

E. Parameters for the FEC Decoding

The cancelable template generated from the CTM is considered to be the noisy codeword of some error correcting code that we can select. Due to its maximum distance seperable (MDS) property, we have selected Reed-Solomon (RS) codes and used RS decoder for FEC decoding in SSTM. The G-dimensional cancelable template is passed through a Reed-Solomon (RS) decoder to identify the closest codeword, which is the multimodal secure sketch.

RS codes use symbols of length m bits. The input to the RS decoder is of length $N' = 2^{m-1}$ in symbols, which means the number of bits per input codeword to the decoder is $n' = mN'$. For example, if the symbol size $m = 6$ then $N' = 63$ is the codeword length in symbols and $n' = 378$ is the codeword length in bits. Let's assume the size of the cancelable template is $G = 378$ bits, which is the number of bits at the input to the RS decoder. This 378-dimensional vector is decoded to generate a secure sketch whose length is K' symbols or, equivalently, $k' = mK'$ bits. K' can be varied depending on the error correcting capability required for the code and k' also signifies the security of the system in bits [60].

We have used *shortened* RS codes. A shortened RS code is one in which the codeword length is less than 2^{m-1} symbols. In standard error control coding, the shortening of the RS code is achieved by setting a number of data symbols to zero at the encoder, not transmitting them, and then re-inserting them at the decoder. A shortened $[N, K]$ Reed-Solomon code essentially uses an $[N', K']$ encoder, where $N' = 2^m - 1$, where m is the number of bits per symbol (symbol size) and $K' = K + (N' - N)$. In our experiments we have used $m = 8$ and $N' = 255$. In the case of using shortened RS codes, the size of the cancelable template is considered equal to N symbols rather than N' symbols. For example, the output of the cancelable template block could be 768 bits which equals to $N = 768/8 = 96$ symbols. The security of the secure multimodal template depends on the selected value of K, implying that the security of the system is k bits, where $k = mK$. The output of the decoder is a length-k binary message, which is cryptographically hashed and stored as the secure multimodal template in the database. When a query is presented for authentication, the system approves the authentication only if the cryptographic hashes of the query matches with the specific enrolled identity.

V. EXPERIMENTAL RESULTS FOR THE CANCELABLE MULTIMODAL TEMPLATE

We have evaluated the matching performance and the security of our proposed secure multibiometric system using the WVU multimodal database [55] containing images for face and iris modalities. Note that all the experiments have been performed with optimized hyper-parameters. We have used $\{\alpha, \beta, \gamma\}$ as $\{8, 2, 2\}$ for FCA and $\{6, 4, 2\}$ for BLA, respectively.

In this section, we analyze the cancelable multimodal template, which is the output of the CTM. Analyzing the output of the CTM helps us to gain insight into the requirements and the strength of the error correcting code to be used in the secure sketch template module (SSTM). In the next section, we analyze the secure multimodal template, which is the output of the overall secure multimodal system.

A. Evaluation Protocol

For the cancelable multimodal template, equal error rate (EER) has been used as one of the metrics to evaluate the matching performance for various levels of random bit selection (values of G). EER indicates a value that the proportion of false acceptances is equal to the proportion of false rejections. The lower the equal error rate value, the higher the accuracy of the biometric system. We have also used the genuine and impostor distribution curves along with the receiver operating characteristic (ROC) curves to evaluate the matching performance of the cancelable template.

B. Performance Evaluation

After fine-tuning the entire DFB, we test this network by extracting features using the JRL of the DFB. In both the FCA and BLA architectures, the output is a 1024-dimensional joint binarized feature vector. For testing, we have used 50 subjects from the WVU-Multimodal 2012 dataset. The training and testing set are completely disjoint which means these 50 subjects have never been used in the training set. 20 face and 20 iris images are chosen randomly for each of these 50 subjects. This will give 20 pairs (face and iris) per subject with no repetitions. These 1,000 pairs (50×20) are forward passed through the DFB and 1024-dimensional 1,000 fused feature vectors are extracted. A user-specific random-bit selection is performed using the fused feature vector to generate the cancelable multimodal template. The number of randomly selected bits G that we have used in our experiments is equal to 128, 256, 512, 768 bits out of the 1024 dimensional binary fused vector to generate the cancelable multimodal template.

In this section, we present the results for the statistical analysis of the cancelable multimodal template, using two different architectures (FCA and BLA) for fusing the face and iris features. The performance evaluation for each architecture is also discussed here.

Two scenarios have been considered for the evaluation of the secure templates. One is the unknown key scenario. In this scenario, the impostor does not have access to the key of the legitimate user. The impostor tries to break into the system by posing as a genuine user by presenting an artificially synthesized key (which is different from the actual key of the genuine user) and also presenting impostor biometrics. This

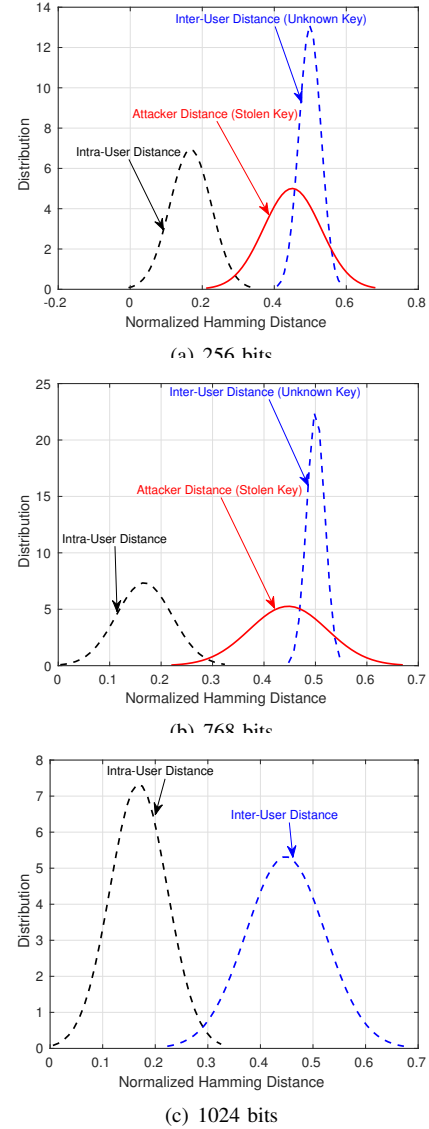(a) 256 bits

(b) 768 bits

(c) 1024 bits

Fig. 4: Genuine and impostor distribution of cancelable template distances using FCA for varying number of random bits.

means that the impostor will try to present random indices for our random-bit selection method in the CTM. These random indices are different from the actual indices that were selected during the enrolment for the legitimate user. The other scenario is the stolen key scenario. In this scenario the impostor has access to the actual key of the genuine user and tries to break the system by presenting actual key with impostor biometrics.

The genuine and impostor distributions for the cancelable template for FCA in the unknown key and stolen key scenarios generated by varying the number of random bits selected by the CTM is given in Fig. 4. The genuine and impostor distributions shown in Fig. 4 have been generated by fitting a normal distribution curve to the histogram. We first observe that there is no overlap between the inter-user (impostor) and intra-user (genuine) distributions. These distributions assume that every user employs his own key. Also plotted is an attacker (stolen key) distribution in which a user (attacker) uses the key of another user (victim). In this case, the attacker distribution slightly overlaps with the genuine distribution,
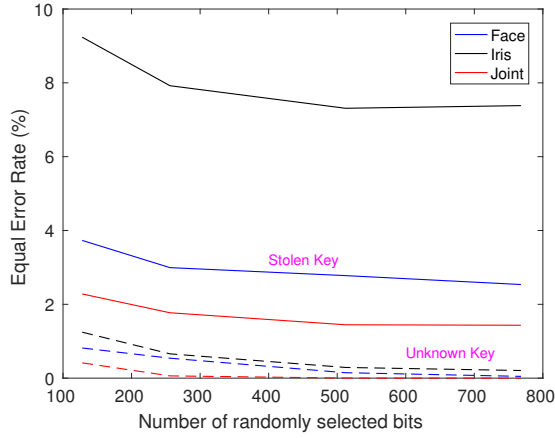
Fig. 5: EER curves for face, iris, joint-FCA modalities in unknown key (dashed lines) and stolen key (solid lines) scenarios using different sizes of cancelable template.
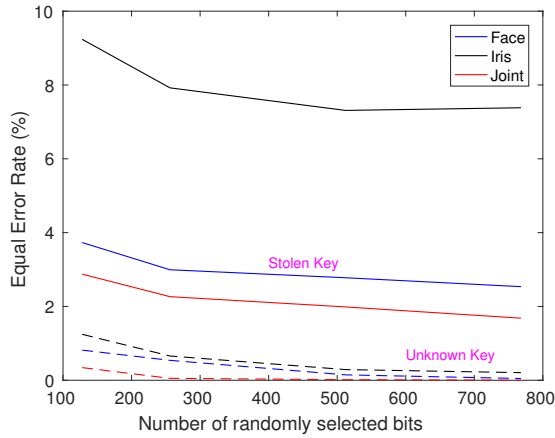

Fig. 6: EER curves for face, iris, and joint-BLA modalities in unknown key (dashed lines) and stolen key (solid lines) scenarios for different sizes of cancelable template.

but the overlap between the two is still reasonably small. In addition, observe that as the number of random bits selected grows from 256 to 768, the overlap between the genuine and impostor distributions reduces in both the scenarios. However, when all the 1024 bits are used, the overlap again is increased. This clearly shows the trade-off between the security (selection of 'G' random bits) and the matching performance (overlap of the distributions). Notice that there is no "stolen key" curve in Fig. 4(c) as all the 1024 bits are used with no down-selection of bits, and hence, no key.

The EER plots for FCA and BLA are given in Fig. 5 and Fig. 6, respectively. EER plot is obtained by calculating the value of EER by varying the length of the cancelable template (number of randomly selected bits). In general, it can be observed from the EER plots that there is an increase in performance by using additional biometric features and the multimodality (joint) template performs better than the individual modalities (face and iris). As seen from the curves, the EER for the joint modality is lower than the EER for face or iris. For example, the EER for joint modality using FCA and BLA at 512 bits for stolen key scenario is 1.45% and 1.99%, respectively. Using the same settings, the EER for face and
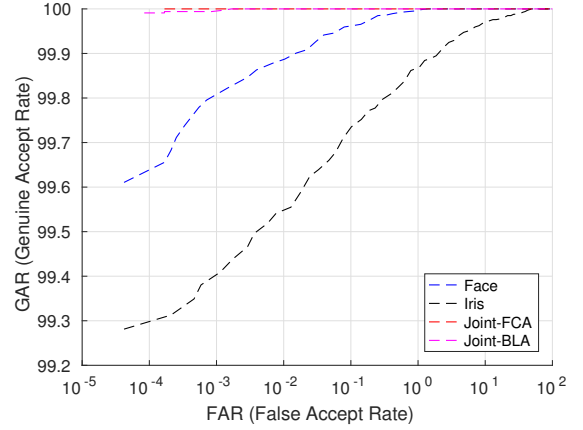

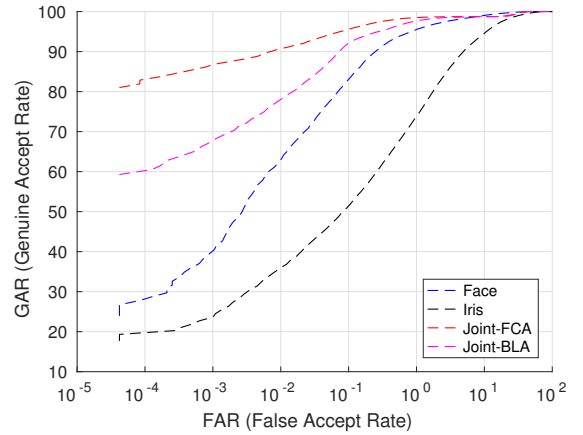Fig. 7: ROC curves for face, iris, joint-FCA, and joint-BLA in unknown key scenario for a random selection of 768 bits.

Fig. 8: ROC curves for face, iris, joint-FCA, and joint-BLA in stolen key scenario for a random selection of 768 bits.

iris is 2.6% and 7.4%, respectively. This clearly shows that there is an improvement by fusing multiple modalities.

The ROC curves for both the architectures have been compared in Fig. 7 and 8 for unknown and stolen key scenarios, respectively, when the number of randomly selected values (security) is 768 bits. Again, we can clearly observe that the joint modality performs better than the individual modality. For a false accept rate (FAR) of 0.5%, the genuine accept rate (GAR) for stolen key scenario using FCA and BLA is 98.25% and 96.33%, respectively. For face and iris, the GAR is 90.8% and 62.5%, respectively at an FAR of 0.5%.

As observed from the plots, the matching performance is not compromised for high security and the multimodality gives us better performance than unimodality.

VI. EXPERIMENTAL RESULTS FOR THE OVERALL SYSTEM

In this section, we analyze the performance at the output of the overall system, where the output of the overall system is the secure multimodal template that is stored in the database.

A. Evaluation Protocol

We evaluate the trade-off between the matching performance and the security of the proposed secure multimodal system using the curves that relate the GAR to the security in bits (i.e., the G-S curves). The G-S curve is acquired by
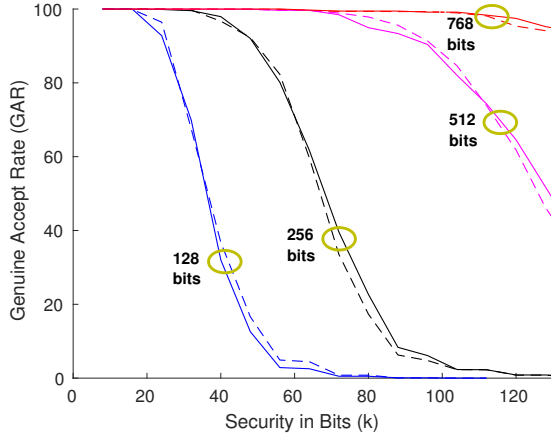
Fig. 9: G-S curves using FCA in unknown key (dashed) and stolen key (solid) scenarios for different values of n bits.
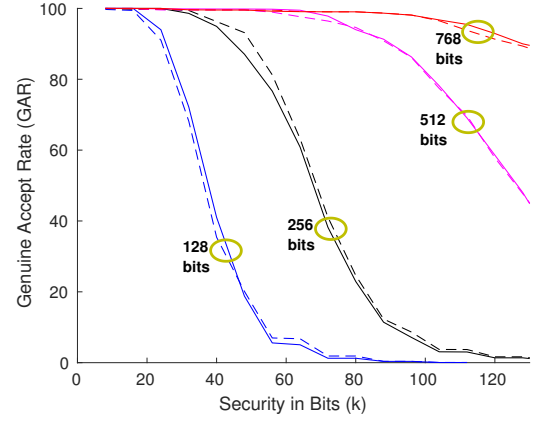

Fig. 10: G-S curves using BLA in unknown key (dashed) and stolen key (solid) scenarios for different values of n bits.
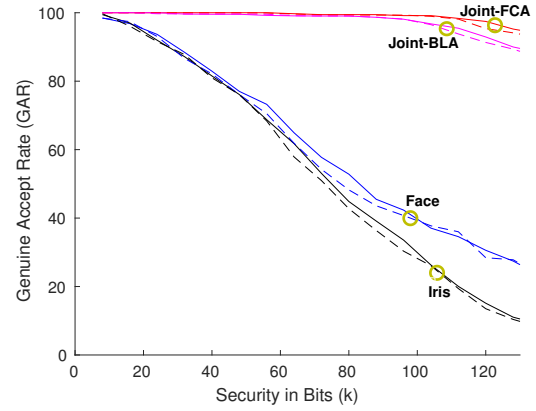

Fig. 11: G-S curves for face, iris, joint-FCA, and joint-BLA modalities in unknown key (dashed lines) and stolen key (solid lines) scenarios for $n = 768$ bits.

varying the error correcting capability of the Reed-Solomon code used for FEC decoding in the SSTM. The error correcting capability of a code signifies the number of bits (or symbols) that a given ECC can correct. The error correcting capability of a Reed-Solomon code is given by $\frac{(N-K)}{2}$ symbols or $\frac{(n-k)}{2}$ bits. We vary the error correcting capability of the code by using different code rates (K/N).

B. Performance Evaluation

As explained in Sec. IV-E, the output of the cancelable template block (n bits) is decoded in order to generate a multimodal secure sketch of length k bits, where k also represents the security of the proposed secure multibiometric system. This multimodal sketch is cryptographically hashed and stored as the secure multimodal template in the database. When a query is presented for authentication, the system authenticates the user only if the cryptographic hash of the query matches that of the specific enrolled identity.

We have experimented with different values of N symbols with $m = 8$ and $N' = 255$ symbols using shortened RS codes. The G-S curves for different values of n bits (equivalent to N symbols) for unknown and stolen key scenarios using FCA and BLA are given in Fig. 9 and Fig. 10, respectively. We can observe from the curves that as the size of the cancelable template in bits (n) increases, the GAR for a given level of security in bits (k) also increases.

For example at a security (k) of 104 bits (equivalent to $K = 13$ symbols) using FCA with the stolen key scenario, the GAR for n=128, 256, 512, and 768 bits is equal to $0.62\%, 4.33\%, 82.05\%$, and 99.16%, respectively. Similarly for the unknown key scenario and FCA, the GAR for n=128, 256, 512, and 768 bits is equal to $0.74\%, 4.3\%, 84.63\%$, and 98.95%, respectively. It can be observed that the use of a larger cancelable template results in better performance. This performance improvement can be attributed to the fact that an increase in n at a fixed value of k (security) improves the error correcting capability of the RS codes which is given by $\frac{(n-k)}{2}$ and hence a better matching performance.

Table I summarizes the GAR for different values of n at security levels of 56, 80, and 104 bits using both FCA and BLA. The error correcting capabilities in bits $\left(\frac{(n-k)}{2}\right)$ for

the RS codes at different security levels are also given in the table. From the Table I, it can be observed that for a given size of the cancelable template in bits (n), the error correcting capability decreases with an increase in the required security levels in bits (k) of the system, which results in a decrease in GAR. This implies that the code cannot correct the intra-class variations at high code rates (k/n) (higher value of k), which results in a reduced GAR. This is the trade-off between the matching performance (GAR) and the security (k) of the system. We have chosen a minimum security level of 56 bits for comparison in Table I which is higher when compared to those reported in the literature [1].

The plot in Fig. 11 gives a comparison of G-S curves for face, iris, joint-FCA, and joint-BLA modalities using $m = 8, N' = 255$ and $n = 768$ bits (equivalent to $N = 96$ symbols) for unknown and stolen key scenario, respectively. The security for the iris modality in stolen key scenario at a GAR of 95% is 20 bits. However, by incorporating additional biometric features (face), the security of the multibiometric system using FCA increases to 128 bits at the same GAR.

C. Comparison with State-of-the-Art Hashing Techniques

As a further experiment, we compare the proposed hashing technique with other hashing techniques. This is done by replacing our hashing method with two other hashing methods

TABLE I: GARs of FCA and BLA in unknown and stolen key scenarios at a security level of 56, 80 and 104 bits using different cancelable template size (N).

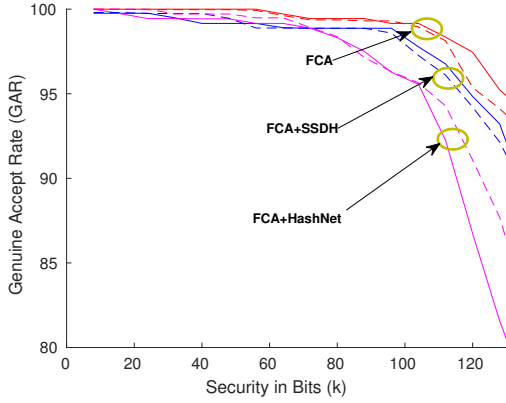| N (symbols)	n (bits)	Security (K) (symbols)	Security (k) (bits)	$\frac{(n-k)}{2}$	FCA-GAR		BLA-GAR	
					Unknown	Stolen	Unknown	Stolen
32	256	7	56	100	82.30%	82.15%	82.25%	80.66%
		10	80	88	31.32%	32.68%	36.67%	35.92%
		13	104	76	4.3%	4.33%	6.77%	6.07%
64	512	7	56	228	99.65%	99.68%	98.95%	99.77%
		10	80	216	97.85%	94.95%	94.63%	94%
		13	104	204	84.63%	82.05%	84.41%	85.15%
96	768	7	56	356	99.93%	99.99%	99.55%	99.22%
		10	80	344	99.37%	99.44%	99.04%	99.04%
		13	104	332	98.95%	99.16%	96.51%	96.75%

Fig. 12: G-S curves to compare performance of the proposed hashing with two other hashing techniques for FCA in unknown (dashed lines) and stolen key (solid lines) scenarios.

[61] and [52], and then training and testing the multimodal authentication system using the same WVU multimodal dataset. The rest of the system is kept the same for comparison purposes. We have compared our hashing technique with supervised semantics-preserving deep hashing (SSDH) [61], and HashNet [52] and evaluated the overall system to produce G-S curves. We have used the FCA system for comparison. We denote the system with our proposed hashing technique as "FCA", use "FCA+SSDH" to denote the FCA architecture with our hashing function replaced by the SSDH hashing, and use "FCA+HashNet" to denote our FCA architecture with the HashNet hashing function. Fig. 12 shows G-S curves for stolen key and unknown key scenarios. It can clearly be seen that our proposed hashing method performs better than the other two deep hashing techniques for the given multimodal biometric security application. Compared to the other two hashing techniques, our proposed method improves the GAR by at least 1.15% at a high security of 104 bits. A comparison of our hashing technique against others for an image-retrieval application can be found in the Appendix.

D. Ablation Study

The objective function defined in (6) contains 3 constraints, one for the semantics-preserving binary codes (i.e., for classification) and two constraints for efficient binary codes (i.e., for binarization and entropy maximization). In this section, we study the relative importance of each of these terms.

First, we measure the influence of the classification term E_1 by setting $\alpha = 1$, $\beta = 0$, and $\gamma = 0$. Using this setting, we train our DFB model and evaluate the overall system by calculating the GAR for a security of $k = 56$, 80, and 104 bits

for $n = 768$ bits (similar to Table I) on the test data for the WVU-Multimodal 2012 dataset. We also study the effect of the binarization constraint along with classification term by setting $\alpha = 1$, $\beta = 1$, and $\gamma = 0$, train our DFB model and again evaluate the overall system by calculating the GARs. Finally, we set $\alpha = 1$, $\beta = 1$, and $\gamma = 1$, and train the DFB model and evaluate the overall system. We performed this experiment for both FCA and BLA architectures only for stolen key scenario because we can see from Table I that unknown key and stolen key scenarios give very similar results. The GAR results for this experiment are shown in Table II.

It can be observed from Table II that the classification term E_1 is the most important term. However, adding the binarization and the entropy constraints E_2 and E_3 (i.e., $\alpha = 1, \beta = 1, \gamma = 1$) definitely help to improve the matching performance (i.e., GAR) by at least 1.25% at a high security of 104 bits in our proposed system. We also note that this performance improvement is evident for both FCA and BLA architectures. Therefore, using all the terms proves beneficial to improve the matching performance evident at higher level of security for both FCA and BLA architectures.

TABLE II: GARs of FCA and BLA in stolen key scenario showing the influence of each term in the objective function.

Hyper-parameters	n (bits)	Security (k) (bits)	FCA-GAR	BLA-GAR
$\alpha = 1, \beta = 0, \gamma = 0$	768	56	99.16%	98.71%
		80	98.32%	96.87%
		104	95.26%	93.29%
$\alpha = 1, \beta = 1, \gamma = 0$	768	56	99.73%	98.76%
		80	98.8%	97.14%
		104	95.72%	94.72%
$\alpha = 1, \beta = 0, \gamma = 1$	768	56	99.52%	98.70%
		80	98.41%	97.02%
		104	95.43%	93.98%
$\alpha = 1, \beta = 1, \gamma = 1$	768	56	99.9%	99%
		80	99.8%	97.6%
		104	96.5%	95.6%

E. Privacy Analysis

The objective of our work is to design a multimodal authentication system that maximizes the matching performance while keeping the biometric data secure. However, the problem is complicated by the possibility that the adversary may gain access to the enrollment key $\mathbf{k_e}$, the multimodal secure sketch $\mathbf{s_e}$, the enrollment feature vector $\mathbf{e}$, or any combination thereof. Using this information, the adversary could not only compromise the authentication integrity of the system, but may also extract information about the biometric data. The system should be robust in these scenarios and the system design should minimize the privacy leakage, which is the leakage of the user biometric information from the compromised data, and preserve authentication integrity of the system.

The G-S curves which have been discussed in Sec. VI-B quantify the security of the system. In this subsection, we will quantify the privacy leakage of the user's biometric information for our proposed system. The privacy of the user is compromised if the adversary gains access to the enrollment feature vector $\mathbf{e}$ as we assume that the enrollment feature vector can be de-convolved to recover the biometric data of the user. The information leaked about the user's enrollment feature vector $\mathbf{e}$ can be quantified as mutual information:

$$I(\mathbf{e}; \mathbf{V}) = H(\mathbf{e}) - H(\mathbf{e}|\mathbf{V}), \qquad (11)$$

where $\mathbf{e}$ represents the enrollment feature vector, and $\mathbf{V}$ represents the information that adversary has access to. $\mathbf{V}$ could be the enrollment key $\mathbf{k_e}$ and/or the multimodal secure sketch $\mathbf{s_e}$. $H(\mathbf{e})$ represents entropy of $\mathbf{e}$ and quantifies the number of bits required to specify $\mathbf{e}$. In particular, $H(\mathbf{e}) = J$ because the optimization described in Sec. IV-A is designed to ensure that the J bits in the encoded template are independent and equally likely to be 0 or 1. $H(\mathbf{e}|\mathbf{V})$ is the entropy of $\mathbf{e}$ given $\mathbf{V}$ and quantifies the remaining uncertainty about $\mathbf{e}$ given knowledge of $\mathbf{V}$. The mutual information $I(\mathbf{e}; \mathbf{V})$ is the reduction in uncertainty about $\mathbf{e}$ given $\mathbf{V}$ [3].

Let's assume that the adversary gains access to the enrollment key $\mathbf{k_e}$. In this case $\mathbf{V} = \mathbf{k_e}$ and mutual information is:

$$I(\mathbf{e}; \mathbf{k_e}) = H(\mathbf{e}) - H(\mathbf{e}|\mathbf{k_e}) = 0, \qquad (12)$$

because $H(\mathbf{e}|\mathbf{k_e}) = H(\mathbf{e}) = J$ as the key $\mathbf{k_e}$ does not give any information about the enrollment feature vector $\mathbf{e}$. $\mathbf{k_e}$ just gives the indices of the random values selected from $\mathbf{e}$ but does not provide values at those indices.

The information leakage when $\mathbf{s_e}$ or the pair $(\mathbf{k_e}, \mathbf{s_e})$ is compromised can be quantified using the conditional mutual information because $\mathbf{s_e}$ is dependent on $\mathbf{r_e}$ which is driven by $\mathbf{k_e}$. Hence, the information leakage when the secure sketch is compromised is conditionally dependent on $\mathbf{k_e}$ and given as:

$$I(\mathbf{e}; \mathbf{s_e}|\mathbf{k_e}) = H(\mathbf{e}|\mathbf{k_e}) - H(\mathbf{e}|\mathbf{s_e}, \mathbf{k_e}), \qquad (13)$$

where $H(\mathbf{e}|\mathbf{k_e})$ quantifies the remaining uncertainty about $\mathbf{e}$ given knowledge of $\mathbf{k_e}$ and $H(\mathbf{e}|\mathbf{s_e}, \mathbf{k_e})$ quantifies the remaining uncertainty about $\mathbf{e}$ given knowledge of $\mathbf{k_e}$ and $\mathbf{s_e}$. This conditional mutual information is measured under two scenarios discussed below.

Both $\mathbf{s_e}$ and $\mathbf{k_e}$ are compromised: In this scenario the adversary gains access to both $\mathbf{s_e}$ and $\mathbf{k_e}$. As previously discussed, $H(\mathbf{e}|\mathbf{k_e}) = H(\mathbf{e}) = J$ because knowing $\mathbf{k_e}$ does not provide any information about $\mathbf{e}$. If the adversary knows $\mathbf{s_e}$, the information leakage of $\mathbf{r_e}$ due to $\mathbf{s_e}$ is equal to the length of $\mathbf{s_e}$ which is k bits. The adversary can use this information of $\mathbf{r_e}$ with the additional knowledge of the enrollment key $\mathbf{k_e}$ and exactly know the indices and the values for the k bits in the enrollment vector $\mathbf{e}$. However, there is still uncertainty about the remaining $J - k$ bits of the enrollment feature vector $\mathbf{e}$, which implies $H(\mathbf{e}|\mathbf{s_e}, \mathbf{k_e}) = J - k$. Therefore, the information leakage about enrollment feature vector when both secure sketch and enrollment key are compromised is:

$$\begin{aligned} I(\mathbf{e}; \mathbf{s_e}|\mathbf{k_e}) &= H(\mathbf{e}|\mathbf{k_e}) - H(\mathbf{e}|\mathbf{s_e}, \mathbf{k_e}) \\ &= J - (J - k) = k. \end{aligned} \qquad (14)$$

Only $\mathbf{s_e}$ is compromised: In this scenario the adversary gains access to only $\mathbf{s_e}$. Even in this case if the adversary knows $\mathbf{s_e}$, the information leakage of $\mathbf{r_e}$ due to $\mathbf{s_e}$ is k bits. However, the adversary does not have any information about the enrollment key $\mathbf{k_e}$ which means that there is added uncertainty in the information about the enrollment feature vector $\mathbf{e}$ as the adversary does not know the exact locations of the k bits given by $\mathbf{s_e}$. This added uncertainty is measured by $H(\mathbf{k_e})$ which is calculated using combinatorics and is:

$$H(\mathbf{k_e}) = \log_2 \binom{J}{n}, \qquad (15)$$

where n is the size of the key and (15) provides all the combinations that n bits could be selected from J. Therefore, the conditional mutual information is given as:

$$\begin{aligned} I(\mathbf{e}; \mathbf{s_e}|\mathbf{k_e}) &= H(\mathbf{e}|\mathbf{k_e}) - H(\mathbf{e}|\mathbf{s_e}, \mathbf{k_e}) \\ &= J - \left(J - k + \log_2 \binom{J}{n} \right) \\ &= k - \log_2 \binom{J}{n} \\ &= \max\left(0, k - \log_2 \binom{J}{n} \right), \end{aligned} \qquad (16)$$

where the max function is applied in the last equation as information leakage cannot be negative. We have evaluated (16) using different values of n and k for $J = 1024$ bits. We know that n ranges from 1 to J depending on the number of random bits selected from the enrollment feature vector $\mathbf{e}$ and k ranges from 1 to n depending on the rate of the error correcting code. We found that information leakage is zero for all the values of k for n ranging from 1 to 792 bits. However, if $n > 792$, there is a positive information leakage for $k > 780$.

From (14) and (16), we can conclude that for $J = 1024$, the ideal value of n should be less than 792 and ideal value of k should be small. This would make the information leakage to be zero or small in case if $\mathbf{s_e}$ or the pair $(\mathbf{s_e}, \mathbf{k_e})$ gets compromised. These values of n and k would also keep the matching performance high as shown in Fig. 11.

F. Unlinkability Analysis

According to ISO/IEC International Standard 24745 [62], transformed templates generated from the same biometric references should not be linkable across applications or databases. By using the protocol defined in [63], we have evaluated the unlinkability of the proposed system. The protocol in [63] is based on mated (H_m) and non-mated (H_{nm}) samples distributions. Mated samples correspond to the templates extracted from the samples of the same subject using different user-specific keys. Non-mated samples correspond to the templates extracted from the samples of different subjects using different keys. For an unlinkable system, there must exist a significant overlap between mated and non-mated score distributions [63].

Using these distributions, two measures of unlinkability are specified: i) Local measure $D_\leftrightarrow(s)$ evaluates the linkability of the system for each specific linkage score s and is dependent upon the likelihood ratio between score distributions. $D_\leftrightarrow(s) \in [0, 1]$ and is defined over the entire score domain. $D_\leftrightarrow(s) = 0$ denotes full unlinkability, while $D_\leftrightarrow(s) = 1$
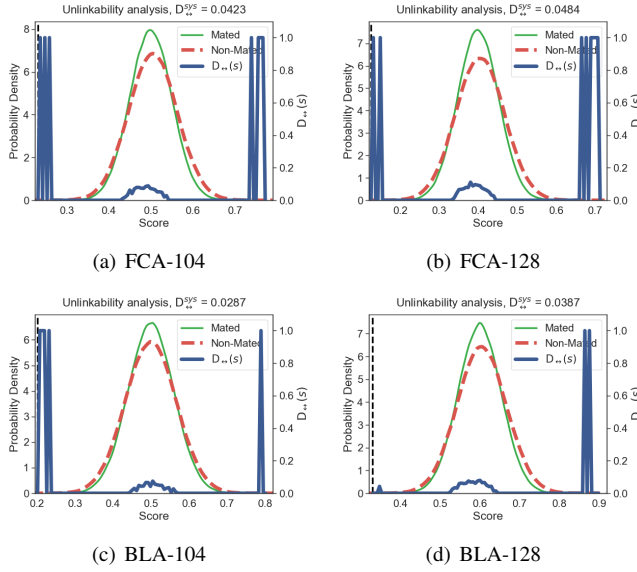
(a) FCA-104

(b) FCA-128

(c) BLA-104

(d) BLA-128

Fig. 13: Unlinkability analysis of the proposed system for FCA and BLA for different quantities of security bits (104, 128).

denotes full linkability of two transformed templates at score s. All values of $D_{\leftrightarrow}(s)$ between 0 and 1 indicate an increasing degree of linkability. ii) Global measure $D_{\leftrightarrow}^{sys}$ provides an overall measure of the linkability of the system independent of the score domain and is a fairer benchmark for unlinkability comparison of two or more systems. $D_{\leftrightarrow}^{sys} \in [0,1]$, where $D_{\leftrightarrow}^{sys} = 1$ indicates full linkability for all the scores of the mated samples distribution and $D_{\leftrightarrow}^{sys} = 1$ indicates full unlinkability for the whole score domain. All values of $D_{\leftrightarrow}^{sys}$ between 0 and 1 indicates an increasing degree of linkability.

According to the benchmark protocol defined in [63], six transformed databases were generated from WVU Multimodal face and iris test dataset by using different set of random bits (enrollment key) in the CTM for each template of a subject. The linkage score we have used is the Hamming distance between the $\mathbf{s_e}$ and $\mathbf{s_p}$. The mated samples distribution and the non-mated samples distribution were computed across these six databases. These score distributions are used to calculate local measure $D_{\leftrightarrow}(s)$, which is further used to compute the global measure $D_{\leftrightarrow}^{sys}$ (overall linkability of the system). Fig. 13 shows unlinkability curves when transformed templates are generated for joint-FCA, and joint-BLA modalities using $m = 8, N' = 255$, and $n = 768$. We have tested with two quantities of security bits $k = 104$ and $k = 128$ bits. With significant overlap, the overall linkability of the system is close to zero for both joint-FCA ($D_{\leftrightarrow}^{sys} = 0.048$) and joint-BLA ($D_{\leftrightarrow}^{sys} = 0.038$). Based on this discussion, the proposed system can be considered to be unlinkable.

VII. CONCLUSION

We have presented a feature-level fusion and binarization framework using deep hashing to design a multimodal template protection scheme that generates a single secure template from each user's multiple biometrics. We have employed a hybrid secure architecture combining the secure primitives of *cancelable biometrics* and *secure-sketch* and integrated it with a deep hashing framework, which makes it computationally

TABLE III: Mean average precision (MAP %) comparison with other hashing methods for 32, 48 and 64 bits.

Methods	ImageNet		
	32	48	64
LSH [15]	25.42	33.74	36.18
ITQ [16]	46.96	53.23	57.05
CCA-ITQ [16]	47.1	55.67	58.80
DHN [36]	49.17	57.19	59.82
HashNet [52]	**63.48**	66.07	68.51
SSDH [61]	63.26	66.34	68.68
Our Hashing Method	63.18	**66.85**	**69.12**

prohibitive to forge a combination of multiple biometrics that passes the authentication. We have also proposed two deep learning based fusion architectures, *fully connected architecture* and *bilinear architecture* that could be used to combine more than two modalities. Moreover, we have analyzed the matching performance and the security, and also performed also unlinkability analysis of the proposed secure multibiometric system. Experiments using the WVU multimodal dataset, which contain face and iris modalities, demonstrate that the matching performance does not deteriorate with the proposed protection scheme. In fact, both the matching performance and the template security are improved when using the proposed secure multimodal system. However, we want to clarify that while the proposed solution is an interesting biometric security framework, in particular for structured data from modalities like face and iris, further validation is required to show how much it can work with other biometric modalities. Finally, the goal of this paper is to motivate researchers to investigate how to generate secure compact multimodal templates.

APPENDIX
IMAGE-RETRIEVAL EFFICIENCY ON IMAGENET DATASET

In order to test the effectiveness of the hashing layer in our proposed methods, we have also tested our deep hashing method for image retrieval on the ImageNet (ILSVRC 2015) [43] dataset and compared the retrieval performance against some baseline hashing methods. The ImageNet dataset contains over 1.2 million images in the training set and about 50 thousand images in the validation set corresponding to 1000 categories. For comparison, we follow the same setting in [52]. We randomly select 100 categories and use all the corresponding training set images as our database and corresponding validation set images as our query points. We select 100 images per category from database as training points.

For evaluation, we use Mean Average Precision (MAP@1000), Precision curves with Hamming radius 2 ($P@r = 2$), and Precision curves for different numbers of top returned samples ($P@K$). We compare our proposed hashing method with 6 state-of-the-art hashing methods including shallow hashing methods LSH [15], ITQ [16], CCA-ITQ [16], and the deep hashing methods DHN [36], HashNet [52] and SSDH [61]. We report results using source code provided by the respective authors except for DHN for which we report result published in [52]. For all the shallow hashing methods, we use VGG-19 fc7 features as input, and for deep hashing methods, we use raw images as input. For fair comparison we use VGG-19 for all the deep hashing methods.

We can observe from the MAP comparison in Table III that our hashing technique is better than shallow hashing
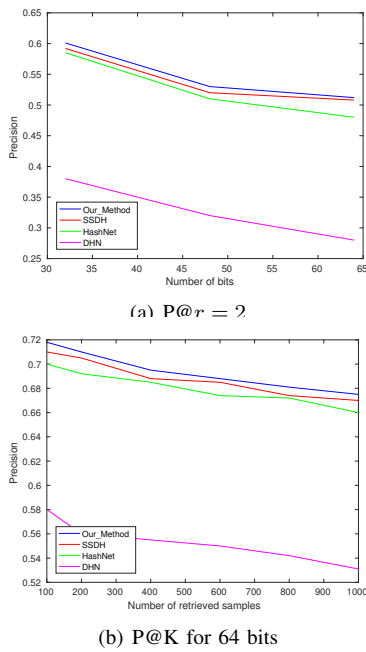
(a) P@r = 2

(b) P@K for 64 bits

Fig. 14: Experimental precision results for the ImageNet dataset for different deep hashing methods.

methods for all hash code lengths. Also, our hashing method is competitive with the other state-of-the-art deep hashing methods when the size of the hash code is 32 bits, at the higher hash code lengths of 48 and 64 bits, our hashing technique is slightly better than other deep hashing methods by 0.35%.

Fig. 14(a) shows the Hamming precision curves for Hamming radius $r = 2$ (P@$r = 2$) for different hash code lengths only for the deep hashing methods. Fig. 14(b) shows the precision for hash code length of 64 bits for different number of top retrieved results (P@K) only for the deep hashing methods. Our hashing technique consistently provides better precision than all the other hashing methods for the same number of retrieved results. Also, it is noted from Fig. 14(a) that precision at 32 bits is better than the precision at 48 and 64 bits. This is because when using longer binary codes, the data distribution in Hamming space becomes progressively sparse and fewer samples fall within the set Hamming radius [34].

ACKNOWLEDGMENT

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] S. Rane, Y. Wang, S. C. Draper, and P. Ishwar, "Secure biometrics: Concepts, authentication architectures, and challenges," *IEEE Signal Processing Magazine*, vol. 30, no. 5, pp. 51–64, Sept. 2013.

[4] Y. Sutcu, Q. Li, and N. Memon, "Protecting biometric templates with sketch: Theory and practice," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 503–512, Sept. 2007.

[5] A. Juels and M. Wattenberg, "A fuzzy commitment scheme," in *6th ACM Conference on Computer and Communications Security*, Nov. 1999, pp. 28–36.

[6] A. Juels and M. Sudan, "A fuzzy vault scheme," in *Proc. IEEE International Symposium on Information Theory*, July 2002, p. 408.

[7] K. Nandakumar, A. K. Jain, and S. Pankanti, "Fingerprint-based fuzzy vault: Implementation and performance," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 4, pp. 744–757, Dec. 2007.

[8] A. Nagar, K. Nandakumar, and A. K. Jain, "Securing fingerprint template: Fuzzy vault with minutiae descriptors," in *Proc. 19th International Conference on Pattern Recognition*, Dec. 2008.

[9] N. K. Ratha, S. Chikkerur, J. H. Connell, and R. M. Bolle, "Generating cancelable fingerprint templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 561–572, April 2007.

[10] A. Kong, K.-H. Cheung, D. Zhang, M. Kamel, and J. You, "An analysis of biohashing and its variants," *Pattern Recognition*, vol. 39, no. 7, pp. 1359–1368, July 2006.

[11] J. Zuo, N. K. Ratha, and J. H. Connell, "Cancelable iris biometric," in *Proc. IEEE International Conference on Pattern Recognition*, Dec. 2008, pp. 1–4.

[12] A. B. Teoh, Y. W. Kuan, and S. Lee, "Cancellable biometrics and annotations on biohash," *Pattern Recognition*, vol. 41, no. 6, pp. 2034–2044, June 2008.

[13] Y. Sutcu, Q. Li, and N. Memon, "Secure biometric templates from fingerprint-face features," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2007.

[14] 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.

[15] A. Gionis, P. Indyk, R. Motwani *et al.*, "Similarity search in high dimensions via hashing," in *Proc. International Conference on Very Large Data Bases*, vol. 99, no. 6, Sept. 1999, pp. 518–529.

[16] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.

[17] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *Proc. AAAI Conference on Artificial Intelligence*, July 2014.

[18] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2015, pp. 3270–3278.

[19] K. Lin, J. Lu, C. S. Chen, and J. Zhou, "Learning compact binary descriptors with unsupervised deep neural networks," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2016, pp. 1183–1192.

[20] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 2064–2072.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, Dec. 2012, pp. 1097–1105.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, Sept. 2014.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 770–778.

[25] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[26] K. He, F. Wen, and J. Sun, "K-means hashing: An affinity-preserving quantization method for learning binary compact codes," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2013, pp. 2938–2945.

[27] P. Jain, B. Kulis, and K. Grauman, "Fast image search for learned metrics," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2008.

[28] B. Kulis and T. Darrell, "Learning to hash with binary reconstructive embeddings," in *Proc. Advances in Neural Information Processing Systems*, Dec. 2009, pp. 1042–1050.

[29] M. Norouzi and D. M. Blei, "Minimal loss hashing for compact binary codes," in *Proc. 28th International Conference on Machine Learning*, July 2011, pp. 353–360.

[30] M. Raginsky and S. Lazebnik, "Locality-sensitive binary codes from shift-invariant kernels," in *Proc. Advances in Neural Information Processing Systems*, Dec. 2009, pp. 1509–1517.

[31] J. Wang, H. T. Shen, J. Song, and J. Ji, "Hashing for similarity search: A survey," *CoRR*, vol. abs/1408.2927, Aug. 2014.

[32] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Advances in Neural Information Processing Systems*, Dec. 2009, pp. 1753–1760.

[33] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[34] X. Yuan, L. Ren, J. Lu, and J. Zhou, "Relaxation-free deep hashing via policy gradient," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[35] Z. Chen, X. Yuan, J. Lu, Q. Tian, and J. Zhou, "Deep hashing via discrepancy minimization," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6838–6847, 2018.

[36] H. Zhu, M. Long, J. Wang, and Y. Cao, "Deep hashing network for efficient similarity retrieval." in *Proc. AAAI Conference on Artificial Intelligence*, Feb. 2016, pp. 2415–2421.

[37] V. M. Patel, N. K. Ratha, and R. Chellappa, "Cancelable biometrics: A review," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 54–65, Sept. 2015.

[38] A. M. Canuto, F. Pintro, and J. C. Xavier-Junior, "Investigating fusion approaches in multi-biometric cancellable recognition," *Expert Systems with Applications*, vol. 40, no. 6, pp. 1971–1980, May 2013.

[39] E. J. C. Kelkboom, X. Zhou, J. Breebaart, R. N. J. Veldhuis, and C. Busch, "Multi-algorithm fusion with template protection," in *Proc. IEEE International Conference on Biometrics: Theory, Applications, and Systems*, Sept. 2009.

[40] B. Fu, S. X. Yang, J. Li, and D. Hu, "Multibiometric cryptosystem: Model structure and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 4, pp. 867–882, Dec. 2009.

[41] P. P. Paul and M. Gavrilova, "Multimodal cancelable biometrics," in *Proc. IEEE International Conference on Cognitive Informatics & Cognitive Computing*, Aug. 2012, pp. 43–49.

[42] C. Rathgeb and C. Busch, "Cancelable multi-biometrics: Mixing iris-codes based on adaptive bloom filters," *Computers & Security*, vol. 42, pp. 1–12, May 2014.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[44] K. D. Nguyen, C. Fookes, A. Ross, and S. Sridharan, "Iris recognition with off-the-shelf CNN features: A deep learning perspective," *IEEE Access*, vol. 6, pp. 18 848–18 855, 2017.

[45] Z. Zhao and A. Kumar, "Towards more accurate iris recognition using deeply learned spatially corresponding features," in *Proc. of the IEEE International Conference on Computer Vision*, 2017, pp. 3809–3818.

[46] S. Minaee, A. Abdolrashidiy, and Y. Wang, "An experimental study of deep convolutional features for iris recognition," in *2016 IEEE signal processing in medicine and biology symposium (SPMB)*, 2016, pp. 1–6.

[47] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

[48] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *CoRR*, vol. abs/1502.00873, 2015.

[49] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *Proc. British Machine Vision Conference (BMVC)*, pp. 41.1–41.12, September 2015.

[50] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[51] T. Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE International Conference on Computer Vision*, Dec. 2015, pp. 1449–1457.

[52] Z. Cao, M. Long, J. Wang, and P. S. Yu, "Hashnet: Deep learning to hash by continuation," in *Proc. IEEE International Conference on Computer Vision*, Oct. 2017, pp. 5609–5618.

[53] Y. Sutcu, S. Rane, J. S. Yedidia, S. C. Draper, and A. Vetro, "Feature extraction for a Slepian-Wolf biometric system using LDPC codes," in *Proc. IEEE International Symposium on Information Theory*, July 2008, pp. 2297–2301.

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

[55] "WVU multimodal dataset." [Online]. Available: http://biic.wvu.edu/

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

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

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

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

[60] 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.

[61] H. Yang, K. Lin, and C. Chen, "Supervised learning of semantics-preserving hash via deep convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 437–451, Feb 2018.

[62] Information Technology — Security Techniques — Biometric Information Protection, ISO/IEC 24745:2011, ISO/IEC JTC1 SC27 Security Techniques, ISO, 2011.

[63] M. Gomez-Barrero, J. Galbally, C. Rathgeb, and C. Busch, "General framework to evaluate unlinkability in biometric template protection systems," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1406–1420, 2018.

Veeru Talreja is a Ph.D. candidate at West Virginia University (WVU), Morgantown, WV, USA. He received the M.S.E.E. degree from West Virginia University and the B.Engg. degree from Osmania University, Hyderabad, India. From 2010 to 2013 he worked as a Geospatial Software Developer with West Virginia University research corporation. His research interests include applied machine learning, deep learning, coding theory, multimodal biometric recognition and security, and image retrieval.

Matthew C. Valenti (M'92 - SM'07 - F'18) received the M.S.E.E. degree from the Johns Hopkins University, Baltimore, MD, USA, and B.S.E.E. and Ph.D. degrees from Virginia Tech, Blacksburg, VA,USA. He has been a Faculty Member with West Virginia University since 1999, where he is currently a Professor and the Director of the Center for Identification Technology Research. His research interests are in wireless communications, cloud computing, and biometric identification. He is the recipient of the 2019 MILCOM Award for Sustained Technical Achievement. He is active in the organization and oversight of several ComSoc sponsored IEEE conferences, including MILCOM, ICC, and Globecom. He was Chair of the ComSoc Communication Theory Technical committee from 2015-2016, was TPC chair for MILCOM'17, is Chair of the Globecom/ICC Technical Content (GITC) Committee (2018-2019), and is TPC co-chair for ICC'21 (Montreal). He was an Electronics Engineer with the U.S. Naval Research Laboratory, Washington, DC, USA. Dr. Valenti is registered as a Professional Engineer in the state of West Virginia

Nasser M. Nasrabadi (S'80 – M'84 – SM'92 – F'01) received the B.Sc. (Eng.) and Ph.D. degrees in electrical engineering from the Imperial College of Science and Technology, University of 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, New York, NY, USA, 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, USA. 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, USA. From 1996 to 2015, he was a Senior Research Scientist with the U.S. Army Research Laboratory. Since 2015, he has been a Professor with the Lane Department of Computer Science and Electrical Engineering. His current research interests are in image processing, computer vision, biometrics, statistical machine learning theory, sparsity, robotics, and neural networks applications to image processing. He is a fellow of ARL and SPIE and has served as an Associate Editor for the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits, Systems and Video Technology, and the IEEE Transactions on Neural Networks.