2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Multi-Level Feature Abstraction from Convolutional Neural Networks for Multimodal Biometric Identification

Sobhan Soleymani, Ali Dabouei, Hadi Kazemi, Jeremy Dawson, and Nasser M. Nasrabadi, *Fellow, IEEE*
West Virginia University
{ssoleyma, ad0046, hakazemi}@mix.wvu.edu, {jeremy.dawson, nasser.nasrabadi}@mail.wvu.edu

***Abstract*—In this paper, we propose a deep multimodal fusion network to fuse multiple modalities (face, iris, and fingerprint) for person identification. The proposed deep multimodal fusion algorithm consists of multiple streams of modality-specific Convolutional Neural Networks (CNNs), which are jointly optimized at multiple feature abstraction levels. Multiple features are extracted at several different convolutional layers from each modality-specific CNN for joint feature fusion, optimization, and classification. Features extracted at different convolutional layers of a modality-specific CNN represent the input at several different levels of abstract representations. We demonstrate that an efficient multimodal classification can be accomplished with a significant reduction in the number of network parameters by exploiting these multi-level abstract representations extracted from all the modality-specific CNNs. We demonstrate an increase in multimodal person identification performance by utilizing the proposed multi-level feature abstract representations in our multimodal fusion, rather than using only the features from the last layer of each modality-specific CNNs. We show that our deep multi-modal CNNs with multimodal fusion at several different feature level abstraction can significantly outperform the unimodal representation accuracy. We also demonstrate that the joint optimization of all the modality-specific CNNs excels the score and decision level fusions of independently optimized CNNs.**

I. Introduction

Feature representation in the biometric systems can be unimodal or multimodal, where the unimodal schemes use a single biometric trait and the multimodal schemes combine the features extracted from multiple biometric feature representations. Benefiting from fusion of features, multimodal biometric models have demonstrated more robustness to noisy data, non-universality and category-based variations [1], [2]. However, one of the major challenges in multimodal biometric systems is the fusion algorithm [3]. The fusion algorithm can be performed at signal, feature, score, rank or decision levels [4], [5], [6], using different schemes such as feature concatenation [7], [8] and bilinear feature multiplication [9], [10], [11].

Compared to score, rank, and decision level fusions, feature level fusion results in a better discriminative classifier [12], [13], due to preservation of raw information [14]. The feature level fusion integrates different features extracted from different modalities into a more abstract and compact feature representation, which can be further used for classification, verification, or identification [15], [16]. Recently several authors have exploited feature level fusion for multimodal biometric identification. Among them the serial feature fusion [17], the parallel feature fusion [18], the CCA-based feature fusion [19], JSRC [14], SMDL [20], and DCA/MDCA [16] are the most prominent techniques.

To integrate features from different modalities, several fusion methods have been considered in the literature [21], [22]. One of the major challenges in multimodal fusion is managing the large dimensionality of the fused feature representations, which highlights the importance of the fusion algorithm. The prevalent fusion method in the literature is feature concatenation, which is very inefficient as the feature space dimensionality increases [3], [8]. Also it does not explore features at different levels of representation and abstraction. To overcome this shortcoming, the weighted feature fusion and multi-level abstract feature representations of individual modalities are proposed in this paper. Using multi-level feature abstraction, feature descriptors at different feature resolutions and abstractions are considered in the proposed classification algorithms. Our proposed fusion method also enforces the higher level dependencies between the modalities through the joint optimization of modality-specific CNNs and backpropagation algorithm. Similar fusion methods have outperformed the conventional feature fusion methods in applications such as multi-task learning [23] and gesture recognition [24], [25].

Convolutional neural networks have been used as classifiers, but they are also efficient tools to extract and represent discriminative features from the raw data at different levels of abstraction. Compared to hand-crafted features, employing CNN as domain feature extractor demonstrated to be more promising when facing different modalities such as face [26], [27], [28], [29], iris [30] and fingerprint [31], [32]. However, the effects of the fusion at different levels of feature resolution and abstraction and joint optimization of the architecture are not investigated for multimodal biometric identification.

In this paper, we make the following contributions: (i) rather than fusing the networks at the softmax layer, the optimally compressed feature representations of all modalities are fused at the fully-connected layers without loss of any performance accuracy, but with a significant reduction in the number of network parameters; (ii) instead of spatial fusion at

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

the convolutional layer, modality-dedicated layers are designed to represent the features for later fusion; (iii) the fully data-driven architecture using fused CNNs, is optimized for joint domain-specific feature extraction and representation with the application of person identification; (iv) in the proposed architecture all the CNNs, the joint representation, and the classifier are jointly-optimized. Therefore, a jointly optimized multimodal representation of all the modalities is constructed. In the previous multi-stream biometric state of the art CNN architectures, the modality-dedicated networks are optimized separately, and the classifier is independent of the modality-dedicated networks, finally (v) multi-level abstract feature fusion for biometric person identification is studied.

To the best of our knowledge this is the first research effort to utilize multi-stream CNNs for joint multimodal fusion person recognition, which deploys multiple abstraction levels of modalities face, iris, and fingerprint.

II. Multi-level feature abstraction and fusion

Our proposed multimodal architecture consists of multiple CNN-based modality-dedicated networks and a joint representation layer, which are jointly optimized. The modality-dedicated networks are trained to extract the modality specific features at different abstract levels, and the joint representation is trained to explore and enforce dependency between different modalities.

In the CNN architecture, each layer represents different abstract feature representation of the input, where deeper levels provide more complex and abstract features. To benefit from different resolutions and abstractions generated by feature maps at different layers of each modality-dedicated CNN, we propose to utilize the information within the feature maps at different layers in our classification algorithm. One generic example for this approach is presented in Figure 1, where deep and shallow level feature maps are contributing in the classification algorithm. In this example, function f maps the feature map space to a one-dimensional feature vector. Then, the combination of the vectors extracted from different levels of abstraction are considered for the classification task.

In this paper, $maxpooling$ followed by a fully-connected layer is considered as the function f. Another example for the function f is $globalpooling$, where each feature map is averaged to construct the representative feature vector. These mappings drastically reduce the number of parameters in the model. In this paper, we focus on the first example where the feature space is mapped to a feature vector through $maxpooling$ and one fully-connected layer, as presented in Figure 2 and Table II.

In the proposed multi-stream CNN, different levels of abstraction from each modality contribute to the decision making algorithm. Table III (b) presents one example for multi-stream multimodal CNN architecture, where each modality is represented in the decision making algorithm by both its deep and shallow feature maps.

network	CNN-Face	CNN-Iris	CNN-Fingerprint
input	$224 \times 224 \times 3$	$64 \times 512 \times 3$	$224 \times 224 \times 3$
layer	kernel	kernel	kernel
conv1 (1-2)	$3 \times 3 \times 64$	$3 \times 3 \times 64$	$3 \times 3 \times 64$
maxpool1	2×2	2×2	2×2
conv2 (1-2)	$3 \times 3 \times 128$	$3 \times 3 \times 128$	$3 \times 3 \times 128$
maxpool2	2×2	2×2	2×2
conv3 (1-4)	$3 \times 3 \times 256$	$3 \times 3 \times 256$	$3 \times 3 \times 256$
maxpool3	2×2	2×2	2×2
conv4 (1-4)	$3 \times 3 \times 512$	$3 \times 3 \times 512$	$3 \times 3 \times 512$
maxpool4	2×2	2×2	2×2
conv5 (1-4)	$3 \times 3 \times 512$	$3 \times 3 \times 512$	$3 \times 3 \times 512$
maxpool5	2×2	2×2	2×2
FC6	$7 \times 7 \times 1024$	$2 \times 16 \times 1024$	$7 \times 7 \times 1024$

TABLE I: The modality-dedicated CNN architectures. Notation conv3 (1-4) represents all four convolutional layers conv3-1,..., conv3-4, where each of these layers includes 256 kernels of size 3×3.

III. Modality-dedicated networks

Each modality-dedicated CNN consists of the first 16 convolutional layers of VGG19 [33] and a fully-connected dimensionality reduction layer (FC6) of size 1024. The conventional VGG19 networks are not practical for this application, since the joint optimization of all the modality-dedicated networks and the joint representation is practically impossible, as the result of massive number of parameters that are needed to be trained. Limitations in the number of training samples, along with large feature dimensionality, result in different training phase complexities which require solutions such as Bayesian controlled sampling [34], imposing common structural assumptions on features [35], and few-shot domain adaptation [36]. In the proposed framework, due to the limited number of training samples, it is not applicable to train a vast number of weights in the last layer of the architecture. Therefore, the number of kernels in the fully connected layer (FC6), compared to the conventional VGG19, is decreased to 1024. The details for each modality-dedicated network can be found in Table I.

IV. Fusion algorithms

In this section, we investigate the fusion of multi-stream CNN architectures. The main goal for the fusion layer is to train the multimodal CNNs such that the ultimate joint feature representation outperforms single modality representations. A recognition algorithm using a multimodal architecture requires selecting the discriminative and informative features at different levels from each modality, as well as exploring the dependencies between different modalities. In addition, the joint optimization should discard the redundant single modality features that are not useful in the joint recognition.

Fusion can be performed on the feature maps of the CNNs, when the feature maps, corresponding to different modalities have the same spatial dimensions. However, in multimodal architectures, the feature level representations can vary in the spatial dimension, due to different inputs' spatial dimensionality. To handle this issue, instead of utilizing feature

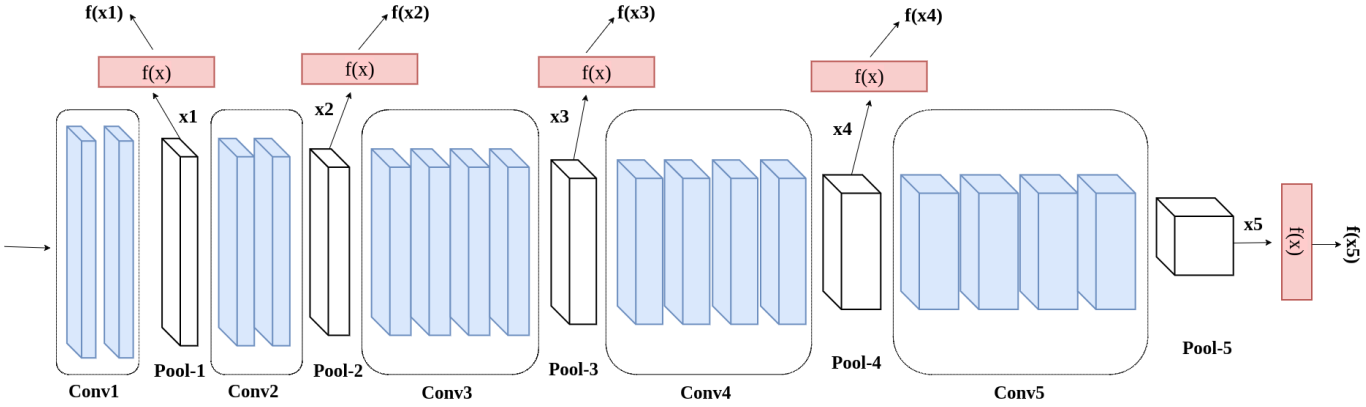


Fig. 1: Our general scheme for the multi-level feature abstraction from a modality-dedicated network, where f maps the feature space to the one-dimensional feature vector. Therefore, the modality-dedicated network is represented by the modality-dedicated embedding layers.

map layers for fusion, fully-connected layers constructed from different CNN feature maps are used in our architecture for ultimate modality-dedicated feature representation. Prior to the fusion, each modality is represented by either the output of its last fully-connected layer, or from multiple CNN layers representing abstract levels. We call these fully connected layers the *modality-dedicated embedding layers*. We demonstrate that the fully-connected representation provides promising results in the case of recognition application. A generic scheme for modality-dedicated embedding layers for a modality-dedicated CNN can be found in Figure 2, where deep and shallow level feature maps are represented by modality-dedicated embedding layers.

A. Weighted feature fusion

In this fusion algorithm, the joint layer is built upon the weighted fusion of the last modality-dedicated embedding layers of the modality-dedicated CNNs. The number of features in this layer is equal to the sum of the number of the output features in the last modality-dedicated embedding layers of the modality-dedicated networks. For instance, for BioCop database [40] which consists of three modalities, the three modality dedicated embedding layers (FC6 layers in Table I) build a layer of size 3072. Then they are fused together using a fully-connected layer of size 1024. The output of this fusion layer is fed to the fully-connected classification layer of size 294, as shown in Table III (a), while no non-linear activation is performed at classification layer. Softmax function is utilized to normalize the outputs of this layer. By training the whole architecture jointly, the first order dependency is enforced between the modalities through backpropagation.

The BIOMDATA database [37], used in our experiments, consists of four fingerprint and two iris modalities. Considering the nature of this database, we propose to use a bi-level weighted feature fusion. In this fusion algorithm, the four fingerprint modalities are fused together using a fully-connected layer of size 1024. Similarly, the two iris modalities are fused together using a fully-connected layer of size 1024. The outputs of these two fully-connected layers are fused using the classification layer of size 219 as presented in Table III (c).

network	CNN-Face	CNN-Iris	CNN-Fingerprint
input(pool3)	$28 \times 28 \times 256$	$8 \times 64 \times 256$	$28 \times 28 \times 256$
layer	kernel	kernel	kernel
pool3x	4×4	4×4	4×4
FC3	$7 \times 7 \times 1024$	$2 \times 16 \times 1024$	$7 \times 7 \times 1024$

TABLE II: Additional layers added to each modality-dedicated network for multi-level feature abstraction fusion.

B. Multi-level feature abstraction and fusion

To benefit from the different resolutions generated by different layers of the modality-dedicated CNN, the $pool_3$ layer is down-sampled using $maxpool$ of size 4×4. Then, a fully-connected layer of size 1024 is considered to represent the shallow level feature maps. This modality-dedicated embedding layer, along with the last layer of the original modality-dedicated network (FC6), are employed as the modality-dedicated embedding layers for the classification task, as presented in Table III (b) and (d). The details for the four architectures considered in this paper are presented in Table III.

V. Experimental Setup

To evaluate the performance of the proposed fusion multimodal architectures, two challenging multimodal biometric databases are considered:

A. Datasets

BioCop multimodal database: The proposed algorithm is evaluated on BioCop database [40]. This database is one of the few databases that allows disjoint training and testing of multimodal fusion at feature level. The BioCop dataset is collected under four disjoint years; 2008, 2009, 2012 and

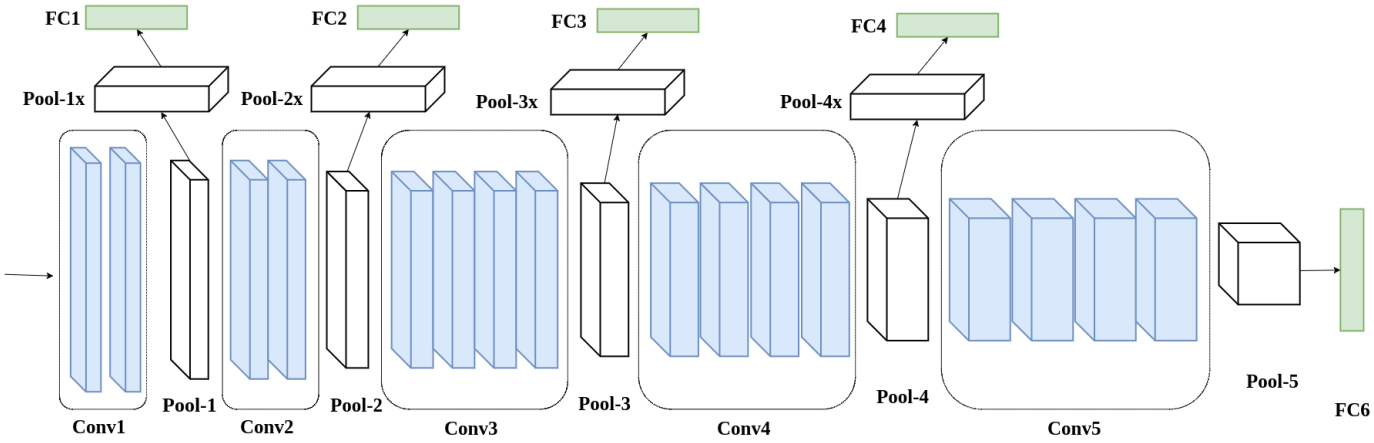


Fig. 2: Our modality-dedicated network is a CNN which consists of the first 16 layers of a VGG19 network and fully connected layers of size 1024. In our fusion algorithm, only $pool_{3x}$ and FC3 of size 1024 are considered for the proposed multi-abstract architecture.

Input	$FC6_{m1}$, $FC6_{m2}$, $FC6_{m3}$
fusion layer	1024
classification layer	294
softmax	

(a) Weighted feature fusion for BioCop database

Input	$FC3_{m1}$, $FC3_{m2}$, $FC3_{m3}$, $FC6_{m1}$, $FC6_{m2}$, $FC6_{m3}$
fusion layer	1024
classification layer	294
softmax	

(b) Multi-abstract fusion for BioCop database

Input	$FC6_{n1}$, $FC6_{n2}$, $FC6_{n3}$, $FC6_{n4}$	$FC6_{n5}$, $FC6_{n6}$
fusion layers	1024	1024
classification layer	219	
softmax		

(c) Bi-level weighted feature fusion for BIOMDATA database

Input	$FC3_{n1}$, $FC3_{n2}$, $FC3_{n3}$, $FC3_{n4}$, $FC3_{n1}$, $FC6_{n2}$, $FC6_{n2}$, $FC6_{n3}$, $FC6_{n4}$	$FC3_{n5}$, $FC3_{n6}$, $FC6_{n5}$, $FC6_{n6}$
fusion layers	1024	1024
classification layer	219	
softmax		

(d) Bi-level multi-abstract fusion for BIOMDATA database

TABLE III: Joint representation architectures and modality-dedicated embedding layers for BioCop (a and b) and BIOMDATA (c and d) databases. m_1, m_2, and m_3 represent three BioCop database modalities. n_1, n_2, n_3, and n_4 represent four fingerprint modalities, and n_5 and n_6 represent two iris modalities for BIOMDATA database.

2013. Each label consists of different biometric modalities for each subject; face, iris, fingerprint, palm print, hand geometry, voice and soft biometrics. To evaluate the performance of the proposed architectures, the following three biometric modalities are considered in this experiment: face, iris, and right index fingerprint.

Under each label, the biometrics are acquired during either one or two separate sessions. The 2012 and 2013 databases contain 1,200 and 1,077 subjects, respectively. To make the training-test splits mutually exclusive, the 294 subjects that are common in labels 2012 and 2013 are considered. The proposed algorithms are trained on 294 mutual subjects in year 2013 dataset, and are tested on the same subjects in year 2012 dataset. It worths mentioning that although the databases are labeled as 2012 and 2013, the date of data acquisition for common subjects in the datasets can vary between one to three years, which also adds the advantage of investigating the age-progression effect. We have also considered the left and right irises as a single class, which results in heterogeneous classes for the iris modality.

In both 2012 and 2013 databases, for each individual, the number of samples per modality may vary. Therefore, in each database, for each individual, 250 triplet of modalities are randomly chosen. Each triplet includes preprocessed face, iris, and right index fingerprint images. The number of image triplets in both training and test sets is the same, and equal to $73,500$.

BIOMDATA multimodal database: BIOMDATA database [37] is a challenging database, since many of the image samples are damaged with blur, occlusion, sensor noise and shadows [16]. This database is a collection of biometric modalities: iris, face, voice, fingerprint, hand geometry, and palm print, from subjects of different ethnicity, gender, and age. Due to privacy issues, face data is not available in combination with other modalities. To evaluate the performance of the proposed architectures, the following six biometric modalities are considered in this experiment: left and right iris, and thumb and index fingerprints from both hands.

The experiments are conducted on 219 subjects that have samples in all six modalities. For each modality, four randomly

		Train set	Test set
BioCop	Face	6833	6960
	Iris	36636	39725
	Fingerprint	1822	991
BIOMDATA	Left iris	874	584
	Right iris	871	581
	Left thumb	875	644
	Left index	872	632
	Right thumb	871	647
	Right Index	870	624

TABLE IV: The size of the training and test sets for each modality in BioCop and BIOMDATA databases.

chosen samples are considered for the training phase and the remaining samples are considered for the test phase. For any modality in which the number of the samples is less than five, one sample is considered for the test phase and the remaining samples are considered for training. The summary of the considered databases is presented in Table IV. For both the training and test sets, for each individual, 250 set of samples are randomly chosen, where each set includes normalized left and right irises, and enhanced left index, right index, left thumb and right thumb fingerprint images. The number of samples in both training and test sets is the same, and equal to $54,750$.

B. Preprocessing

For the face modality, the frontal images are considered. The face images are cropped, aligned to a template [44], and resized to 224×224 images. Fingerprint images are enhanced using the method described in [45]. The core point is detected from the enhanced images [46]. Finally, the 224×224 region centered by the core point is cropped.

Iris images are segmented and normalized using OSIRIS [47]. Although OSIRIS software does not mask eyelids and eyelashes, the segmented images do not contain much occlusion due to eyelids [48]. OSIRIS algorithm finds the iris inner and outer contours. This two contours are used to transform the iris area into a 64×512 strip.

VI. Joint Optimization of Networks

In this section, the training of the multimodal CNN architecture is discussed. Here, we explain the implementation of each modality-dedicated network, the joint fusion representation layer, and the concurrent optimization of the multimodal CNNs and the fusion layer. The fusion layer can be either the weighted fusion layer or multi-abstract fusion layer.

A. Modality-dedicated networks

Initially, the modality-dedicated CNNs are trained independently and each CNN is optimized for a single modality. As explained in section III, each of these CNN networks consists of the first 16 convolutional layers of VGG19 network with an added fully-connected feature reduction FC6 layer of size 1024. The extra layer is dedicated to make the feature level fusion tractable. For each modality, the conventional VGG19 network is trained as explained below. For all the modalities, the networks are initialized by VGG19 pre-trained on Imagenet [38].

CNN-Face: To optimize the weights for extracting the face features, the pre-trained network is fine-tuned on the CASIA-Webface [39] and the BioCop [40] face 2013. The network is then trained on 294 subjects in the dataset 2013 that are also present in the 2012 dataset. Finally, previously trained weights for the first 16 layers of the network, along with FC6 layer of size 1024 and the softmax layer, are fine-tuned on the 294 subjects in 2013 dataset.

The face image inputs are 224×224 RGB images. The preprocessing algorithm consists of the channel-wise mean subtraction on RGB values, where channel means are calculated on the whole training set. The training algorithm is deployed by minimizing the softmax cross-entropy loss using mini-batch stochastic gradient descent with momentum. The training is regularized by weight decay and 50% dropout for the fully-connected layers except for the last layer. The batch size, momentum and L_2 penalty multiplier are set to 32, 0.9, and 0.0005, respectively. The initial learning rate is set to 0.1. The learning rate is decreased exponentially by a factor of 0.1 for every 2 epochs of training. In this network, batch normalization [41] is applied. The moving average decay is set to 0.99.

CNN-Iris: Similar to the face network, the training is performed over the Imagenet pre-trained VGG19. To specify the kernels to extract the iris-specific features, the pre-trained network is finetuned on the CASIA-Iris-Thousand [42] and Notre Dame-IRIS 04-05 [43].

For the BioCop [40] database, the network is then tuned on BioCop iris 2013. The network is then fine-tuned on 294 subjects in the dataset 2013 which are also present in the 2012 dataset. After dropping the last two layers of VGG19 and adding FC6 and the softmax layers, the network is once again trained on these 294 subjects in 2013 dataset. The iris image inputs are 64×512 grayscale images. The optimization parameters are the same as face architecture. In this network, batch normalization is also applied. The moving average decay is set to 0.9. The learning rate decrease exponentially by a factor of 0.1 every 5 epochs.

For the BIOMDATA database [37], for each of the two modalities, the pre-tuned network is tuned on all subjects that have samples in that modality, and then, on 219 subjects that have samples in all six modalities. Since the number of samples in this dataset is much smaller than the number of samples in BioCop database, the learning rate decay is set to 0.99. Similar to CNN-Face, each tuned networks is fine-tuned after dropping the fully connected layers and adding FC6.

CNN-Fingerprint: Fingerprint networks are initiated with Imagenet pre-trained VGG19 weights. For BioCop database, it is then trained on BioCop 2013 fingerprint dataset. Then, the network is fine-tuned on 294 subjects in the dataset 2013 which are also present in the 2012 dataset. For the BIOMDATA database, for each of the four modalities, the pre-tuned network is tuned on all subjects that have samples in

that modality, and then, on 219 subjects that have samples in all six modalities. The inputs are 224×224 grayscale images. The optimization parameters are the same as CNN-Face architecture. The learning rate decreases exponentially by a factor of 0.1 every 10 epochs.

B. Joint optimization of networks

Initially, to train the joint representation, the modality-dedicated networks' weights are frozen, and the joint representation layer is optimized greedily upon the extracted features from the modality-dedicated networks. The optimization parameters are the same as the CNN-Fingerprint network. Finally all the networks are jointly optimized. Here, the batch size is further reduced, and the initial learning rate is reduced to the the smallest final learning rate among modality-dedicated networks. In all the mentioned steps, Rectified Linear Unit (ReLU) activation function is utilized for all the layers except the classification layer.

C. Hyperparameter optimization

The hyperparameters in our experiments are : λ the regularization parameter, α_0 initial learning rate, n number of epochs per decay for the learning rate, d moving average decay, and the m as the momentum. For each optimization, the five-fold cross-validation method is considered to estimate the best hyperparameters.

VII. Experiments and disscussions

A. Evaluation metrics

The performance of different experiments are reported and compared using two classification metrics: classification accuracy and *Recall@K*. The classification accuracy is the fraction of correctly classified samples regarding their classes. The *Recall@K* metric is the probability that a subject class is correctly classified at least at rank-k, while the candidate classes are sorted by their similarity score to the query samples. The calculation of *Recall@K* is done per class, and is averaged over all available classes.

The reported values are the average values for five randomly generated training and test sets. As explained in section V. *A*, training and test sets consist of $73,500$ triplet images for BioCop database and $54,750$ sets of six images for BIOMDATA database.

B. Data augmentation

For both databases, data augmentation is performed on the fingerprint images. 20 augmented samples of each fingerprint image is generated by translating the core point both vertically and horizontally using Gaussian distribution. For each fingerprint image, ten augmented images are generated using Gaussian distribution with parameters $\mu = 0$ and $\sigma = 2.5$. The remaining ten augmented images are generated with $\mu = 0$ and $\sigma = 5$ being considered. In Table V studies the effect of data augmentation on the rank-one recognition rate for the modality-dedicated CNN-Fingerprint for BioCop database. This table also includes the recognition rate for NBIS software [49].

	NBIS	CNN w/o	CNN
Right index	95.67	96.08	97.28

TABLE V: Rank-one recognition rate for BioCop database utilizing NBIS software, CNN without data augmentation and CNN with data augmentation.

		KNN	SVM	CNN
BioCop	Face	89.68	88.76	98.14
	Iris	70.52	79.26	99.05
	Right index	91.22	90.61	97.28
BIOMDATA	Left iris	66.61	71.92	99.35
	Right iris	64.89	71.08	98.95
	Left thumb	61.23	63.96	80.15
	Left index	82.91	84.70	93.43
	Right thumb	62.11	63.52	82.63
	Right Index	82.05	84.46	93.12

TABLE VI: Rank-one recognition rate for single modalities.

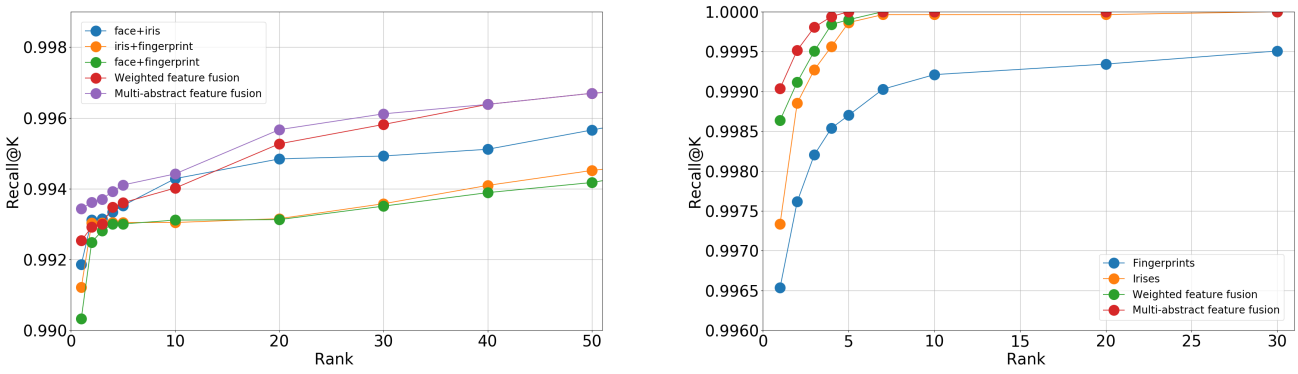
C. Results

To compare the results for the proposed algorithms with the state-of-the-art algorithms, Gabor features in five scales and eight orientations are extracted from all modalities. For the face images, $31,360$ features are extracted from 224×224 aligned images. While, for the iris images, $36,630$ features are extracted from 64×512 segmented and normalized image. In the case of fingerprint images, $31,360$ features are extracted from the enhanced 224×224 images, as described in Section V-B, around the core point. These features are used for all the state-of-the-art algorithms except CNN-Sum, CNN-Major, and two proposed algorithms.

Table VI presents single modality rank-one recognition rate for both the databases. In this table the Gabor features are used for SVM and KNN algorithms. The performance of the proposed weighted feature level fusion and multi-abstract fusion algorithms are compared with that of several state-of-the-art feature, score and decision level fusion algorithms in Tables VII and VIII. SVM-Sum and CNN-Sum use the probability outputs for the test sample of each modality, added together to produce the final score vector. SVM-Major and CNN-Major chose the maximum number of modalities taken to be from the correct class.

The feature level fusion techniques include the serial feature fusion [17], the parallel feature fusion [18], the CCA-based feature fusion [19], JSRC [14], SMDL [20] and DCA/MDCA [16] algorithms. Note that in case of more than two modalities, the parallel feature fusion method cannot be applied. Tables VII and VIII present the results for BioCop and BIOMDATA databases, respectively. As presented in this table, both fusion algorithms outperform single-modality and two-modality architectures for BioCop database. Similarly, both the fusion algorithms outperform single-modality, two irises, and four fingerprints architectures. The proposed algorithms excel the score-level and the decision-level fusion algorithms for the independently-optimized CNNs as well.

Figure 3 presents Cumulative Match Curves (CMCs) for



(a) (b)

Fig. 3: CMC curves for (a) BioCop, and (b) BIOMDATA databases.

Modality	{1,2}	{1,3}	{2,3}	{1,2,3}
SVM-Major	79.22	89.27	80.47	90.32
Serial + PCA + KNN	71.12	86.28	75.69	76.18
Serial + LDA + KNN	80.12	91.28	79.69	82.18
Parallel + PCA + KNN	74.69	88.12	77.58	-
Parallel + LDA + KNN	82.53	93.21	82.56	-
CCA + PCA + KNN	87.21	95.27	86.44	95.33
CCA + LDA + KNN	89.12	95.41	86.11	95.58
DCA/MDCA + KNN	83.02	96.36	83.44	86.49
CNN-Sum	99.10	98.85	98.92	99.14
CNN-Major	98.51	97.70	98.31	99.03
Weighted feature fusion	99.18	99.03	99.12	99.25
Multi-abstract fusion	99.31	99.16	99.20	99.34

TABLE VII: Rank-one accuracy evaluation on BioCop database, for different fusion settings. 1, 2, and 3 represent face, iris, and fingerprint, respectively.

Modality	2 irises	4 fingerprints	6 modalities
SVM-Major	78.12	88.34	93.31
SVM-Sum	81.23	94.13	96.85
Serial + PCA+ KNN	72.31	90.71	89.11
Serial + LDA+ KNN	79.82	92.62	92.81
Parallel + PCA+ KNN	76.45	-	-
Parallel + LDA+ KNN	83.17	-	-
CCA + PCA + KNN	88.47	94.72	94.81
CCA + LDA + KNN	90.96	94.13	95.12
JSRC	78.20	97.60	98.60
SMDL	83.77	97.56	99.10
DCA/MDCA + KNN	83.77	98.1	99.60
CNN-Sum	99.54	99.46	99.82
CNN-Major	99.31	99.42	99.48
Weighted feature fusion	99.73	99.65	99.86
Multi-abstract fusion	99.81	99.72	99.91

TABLE VIII: Rank-one accuracy evaluation on BIOMDATA database, for different fusion settings.

both databases. Figure 3 (a) compares *Recall@K* for two and three modality weighted feature fusion algorithms with the three modality multi-abstract feature fusion algorithm. Similarly, Figure 3 (b) compares three weighted feature fusion scenarios (two irises, four fingerprints, and all six modalities) with the six modality multi-abstract feature fusion algorithm. For both the studied databases, the multi-abstract fusion algorithm excels the weighted fusion algorithm both in terms of rank-one recognition rate and CMC curve.

VIII. Conclusion

In this paper, we proposed a joint CNN architecture with feature level fusion for multimodal recognition using multiple modalities of face, iris, and fingerprint. We proposed a multi-abstract network to handle the spatial mismatch problem and yet having no loss in performance with significant reduction in network parameters. We demonstrated that the proposed multi-stream CNNs with multimodal fusion at different feature level abstraction and jointly optimization of modality-dedicated networks, joint representation, and classifier, significantly improve unimodal representation accuracy by incorporating the captured multiplicative interactions of the low-dimensional modality-dedicated feature representations. Two fusion methods at the fully-connected layer are studied, and it is concluded that the multi-abstract fusion outperforms the weighted feature fusion algorithm.

Acknowledgement

This work is based upon a work supported by the Center for Identification Technology Research and the National Science Foundation under Grant #1650474.

References

[1] H. Jaafar and D. A. Ramli, "A review of multibiometric system with fusion strategies and weighting factor," *International Journal of Computer Science Engineering (IJCSE)*, vol. 2, no. 4, pp. 158–165, 2013.

[2] C.-A. Toli and B. Preneel, "A survey on multimodal biometrics and the protection of their templates," in *IFIP International Summer School on Privacy and Identity Management*, 2014, pp. 169–184.

[3] 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, 2012.

[4] R. Connaughton, K. W. Bowyer, and P. J. Flynn, "Fusion of face and iris biometrics," in *Handbook of Iris Recognition*, 2013, pp. 219–237.

[5] S. Singh, A. Gyaourova, G. Bebis, and I. Pavlidis, "Infrared and visible image fusion for face recognition," in *Proceedings of SPIE*, vol. 5404, 2004, pp. 585–596.

[6] M. F. Nadheen and S. Poornima, "Feature level fusion in multimodal biometric authentication system," *International Journal of Computer Applications*, vol. 69, no. 18, 2013.

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

[8] G. Goswami, P. Mittal, A. Majumdar, M. Vatsa, and R. Singh, "Group sparse representation based classification for multi-feature multimodal biometrics," *Information Fusion*, vol. 32, pp. 3–12, 2016.

[9] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.

[10] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller, "One-to-many face recognition with bilinear cnns," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[11] S. Soleymani, J. Dawson, and N. M. Nasrabadi, "Generalized bilinear deep convolutional neural networks for multimodal biometric identification," in *IEEE International Conference on Image Processing*, 2018.

[12] M. Faundez-Zanuy, "Data fusion in biometrics," *IEEE Aerospace and Electronic Systems Magazine*, vol. 20, no. 1, pp. 34–38, 2005.

[13] A. A. Ross and R. Govindarajan, "Feature level fusion of hand and face biometrics," in *Defense and Security*. International Society for Optics and Photonics, 2005, pp. 196–204.

[14] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 113–126, 2014.

[15] M. Eshwarappa and M. V. Latte, "Multimodal biometric person authentication using speech, signature and handwriting features," *International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*, 2011.

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

[17] C. Liu and H. Wechsler, "A shape-and texture-based enhanced fisher classifier for face recognition," *IEEE transactions on image processing*, vol. 10, no. 4, pp. 598–608, 2001.

[18] J. Yang, J.-y. Yang, D. Zhang, and J.-f. Lu, "Feature fusion: parallel strategy vs. serial strategy," *Pattern recognition*, vol. 36, no. 6, pp. 1369–1381, 2003.

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

[20] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE transactions on Image Processing*, vol. 25, no. 1, pp. 24–38, 2016.

[21] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Fusion strategies in multimodal biometric verification," in *Proceedings, International Conference on Multimedia and Expo (ICME)*, vol. 3, July 2003, pp. III–5–8 vol.3.

[22] A. Lumini and L. Nanni, "Overview of the combination of biometric matchers," *Information Fusion*, vol. 33, pp. 71–85, 2017.

[23] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *arXiv preprint arXiv:1603.01249*, 2016.

[24] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: adaptive multi-modal gesture recognition," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2016.

[25] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *Workshop at the European conference on computer vision*, 2014, pp. 474–490.

[26] H. Kazemi, S. Soleymani, A. Dabouei, M. Iranmanesh, and N. M. Nasrabadi, "Attribute-centered loss for soft-biometrics guided face sketch-photo recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2018.

[27] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 98–113, 1997.

[28] S. M. Iranmanesh, A. Dabouei, H. Kazemi, and N. M. Nasrabadi, "Deep cross polarimetric thermal-to-visible face recognition," in *International Conference on Biometrics*, 2018.

[29] H. Kazemi, M. Iranmanesh, A. Dabouei, S. Soleymani, and N. M. Nasrabadi, "Facial attributes guided deep sketch-to-photo synthesis," in *IEEE Winter Applications of Computer Vision Workshops*, 2018, pp. 1–8.

[30] A. Gangwar and A. Joshi, "Deepirisnet: Deep iris representation with applications in iris recognition and cross-sensor iris recognition," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 2301–2305.

[31] R. F. Nogueira, R. de Alencar Lotufo, and R. C. Machado, "Fingerprint liveness detection using convolutional neural networks," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 6, pp. 1206–1213, 2016.

[32] A. Dabouei, H. Kazemi, S. M. Iranmanesh, J. Dawson, and N. M. Nasrabadi, "Fingerprint distortion rectification using deep convolutional neural networks," in *International Conference on Biometrics*, 2018.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[34] A. Broumand, M. S. Esfahani, B.-J. Yoon, and E. R. Dougherty, "Discrete optimal bayesian classification with error-conditioned sequential sampling," *Pattern Recognition*, vol. 48, no. 11, pp. 3766–3782, 2015.

[35] A. Broumand and T. Hu, "A length bias corrected likelihood ratio test for the detection of differentially expressed pathways in rna-seq data," in *Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on*, 2015, pp. 1145–1149.

[36] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto, "Few-shot adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, 2017, pp. 6673–6683.

[37] S. Crihalmeanu, A. Ross, S. Schuckers, and L. Hornak, "A protocol for multibiometric data acquisition, storage and dissemination," *Technical Report, WVU, Lane Department of Computer Science and Electrical Engineering*, 2007.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

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

[40] "Biocop database, http://biic.wvu.edu/." [Online]. Available: http://biic.wvu.edu/

[41] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

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

[43] K. W. Bowyer and P. J. Flynn, "The ND-IRIS-0405 iris image dataset," *arXiv preprint arXiv:1606.04853*, 2016.

[44] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[45] S. Chikkerur, C. Wu, and V. Govindaraju, "A systematic approach for feature extraction in fingerprint images," *Biometric Authentication*, pp. 1–23, 2004.

[46] A. K. Jain, S. Prabhakar, L. Hong, and S. Pankanti, "Filterbank-based fingerprint matching," *IEEE transactions on Image Processing*, vol. 9, no. 5, pp. 846–859, 2000.

[47] E. Krichen, A. Mellakh, S. Salicetti, and B. Dorizzi, "Osiris (open source for iris) reference system," *BioSecure Project*, 2008.

[48] K. P. Hollingsworth, K. W. Bowyer, and P. J. Flynn, "The best bits in an iris code," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 964–973, 2009.

[49] K. Ko, "Users guide to export controlled distribution of nist biometric image software (nbis-ec)," *NIST Interagency/Internal Report (NISTIR)-7391*, 2007.