Image and Vision Computing 102 (2020) 103977

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Multimodal facial biometrics recognition: Dual-stream convolutional neural networks with multi-feature fusion layers

Leslie Ching Ow Tiong [a], Seong Tae Kim [b], Yong Man Ro [c,*]

[a] Computational Science Research Center, Korea Institute of Science and Technology (KIST), 5 Hwarang-ro, 14-gil Seongbuk-gu, Seoul 02792, Republic of Korea
[b] Computer Aided Medical Procedures, Technical University of Munich, Boltzmanstr 3, Garching 85748, Germany
[c] Image and Video System Lab, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

ARTICLE INFO

Article history:
Received 14 October 2019
Accepted 29 June 2020
Available online 6 July 2020

Keywords:
Multimodal facial biometrics recognition
Deep multimodal learning
Dual-stream convolutional neural network
Network fusion layers

ABSTRACT

Facial recognition for surveillance applications still remains challenging in uncontrolled environments, especially with the appearances of masks/veils and different ethnicities effects. Multimodal facial biometrics recognition becomes one of the major studies to overcome such scenarios. However, to cooperate with multimodal facial biometrics, many existing deep learning networks rely on feature concatenation or weight combination to construct a representation layer to perform its desired recognition task. This concatenation is often inefficient, as it does not effectively cooperate with the multimodal data to improve on recognition performance. Therefore, this paper proposes using multi-feature fusion layers for multimodal facial biometrics, thereby leading to significant and informative data learning in dual-stream convolutional neural networks. Specifically, this network consists of two progressive parts with distinct fusion strategies to aggregate RGB data and texture descriptors for multimodal facial biometrics. We demonstrate that the proposed network offers a discriminative feature representation and benefits from the multi-feature fusion layers for an accuracy-performance gain. We also introduce and share a new dataset for multimodal facial biometric data, namely the Ethnic-facial dataset for benchmarking. In addition, four publicly accessible datasets, namely AR, FaceScrub, IMDB_WIKI, and YouTube Face datasets are used to evaluate the proposed network. Through our experimental analysis, the proposed network outperformed several competing networks on these datasets for both recognition and verification tasks.

1. Introduction

In recent years, many studies have applied multimodal biometric recognition to produce much-improved recognition performance using advanced algorithms for a surveillance camera, identity authentication, border control, etc. [1–4]. In a multimodal setting, using multiple modalities of the biometric data to build various representation creates more challenges. To overcome such scenarios, deep learning has introduced an impressive ability to learn high-dimensional and complex data in the visual domain [5]. This yields a more vibrant feature representation result that could be used to enhance the performance of recognition.

This paper studies the limitations of face recognition in surveillance through multimodal facial biometrics recognition by focusing on the face and periocular modalities. The first question is raised here, "Why do we choose the face and periocular modalities?" Note that most surveillances are image-based applications. Several studies have proven that the easiest and fastest ways to extract the biometric modalities from the camera are either through the face, periocular or gait [6–8]. Furthermore, most of the surveillance cameras fail to identify the criminal suspects due to appearance occlusions such as wearing masks, covering their faces with scarfs, ethnic groups effects, cosmetic products, etc. [9–11]. For all these reasons, this motivates us to add the periocular as an additional feature in enhancing the performance of recognition.

The next concern is, "How can we present multimodal biometric data into a deep learning network?" A good deep multimodal learning network must satisfy certain properties such that it should be easy to obtain even in the presence of the hidden information in several modalities. In other words, useful representations can be learned through such data by fusing the modalities into a joint feature representation, which captures the correlations of the data that it corresponds to. This perhaps increases the interest of deep multimodal learning in the biometrics and computer vision communities [12,13].

In this paper, we attempt to address the challenges of surveillance in uncontrolled environments, especially for the ethnicities effect, which remains not well-addressed by the current works [10,14–16]. Thus, we study this problem by means of implementing multi-feature fusion layers in two independent dual-stream Convolutional Neural Networks (CNNs), which accepts the multimodal biometric data (face and periocular regions). Both dual-stream CNNs aggregate the RGB data and texture descriptors to support efficient feature learning. The

* Corresponding author.
E-mail address: ymro@kaist.ac.kr (Y.M. Ro).

https://doi.org/10.1016/j.imavis.2020.103977
0262-8856/

proposed fusion layers are designed to incorporate the multimodal inputs by strengthening the feature representations in the networks and improving the recognition performance.

1.1. Related works

For deep multimodal learning, several studies have been considered in the literature [12,17], whereby it highlights the importance of a fusion algorithm. Early studies conducted by Srivastava & Salakhutdinov [18] and Zagoruyko & Komodakis [19] proposed several representations that fused across fusion layers during training. Other similar studies done by Kahou et al. [20], Liu et al. [21], Simonovsky et al. [22], and Zhang et al. [23] have demonstrated deep multimodal learning networks with a simple fusion approach, where the prior knowledge is exploited to merge discriminant representations from multimodal data. However, these networks are less robust under "in-the-wild" variations such as low-resolution or occlusions. This is because all the networks are only applied feature concatenation to represent biometrics features, which could not perform robust learning hierarchical representations across late-fusion layers and were unable to be abstracted into discriminating features at various levels.

Another study related to image correlation was presented by Feichtenhofer et al. [24] and Hu et al. [25] proposed a fusion strategy at different stages within a deep learning architecture. Hu et al. [25] suggested that fusion approaches at the feature stage that can improve performance, while Feichtenhofer et al. [24] identified that implemented fusion approaches at the convolutional (*conv*) layers that can extract more discriminative information for complex data learning. However, these networks were found to underperform when temporal images contained too much noise or low-resolution images that could cause misalignments in extracting the features.

Presently, Soleymani et al. [26] introduced a multi-level abstraction fusion CNN, where the face, fingerprint, and iris features are fused at the fully-connected (*fc*) layer. Its fusion layer is designed to concatenate or merge at different levels of the *fc* layers as a multi-feature representation with RGB data. This approach leveraged several biometrics modalities whereby all of them may not always be available such as having a mask-wearing face or an iris far from camera distance. In addition, such multimodal biometrics applications may jeopardize the usability of the system such as fingerprint and iris modalities that are required for stable cooperation from the individuals.

In the previous works of deep learning networks that consume text descriptor, Levi et al. [27] and Anwer et al. [28] established the use of a text descriptor as an input to their networks for classification. The authors demonstrated that texture descriptors are beneficial to train the network. The previous works motivate us to investigate and analyze the impact of RGB data and texture descriptors with network fusion layers for presenting multimodal data within a deep multimodal learning network.

1.2. Motivation and contributions

Currently, the challenges of face recognition technology for surveillance are still concerned about the "in-the-wild" environments and ethnic groups effects, such as differences in appearances, cameras location, level of illuminations, plastic surgery, and others [7,29]. Especially after the "Boston Marathon bombings", migration issues, and the recent terror attacks in Paris and Brussels, the security experts and the police departments in Europe and the U.S. have agreed that face recognition technology is still possesses remaining challenges, such as the appearances of subjects wearing masks, covered by scarf/veil, and low-resolution camera [10,30,31].

This paper offers a solution for the mentioned challenges in face recognition by investigating several fusion strategies in the proposed networks. We firstly propose two independent networks with several network fusion layers to exploit the different features among the RGB data and texture descriptors in order to represent the multimodal data for recognition. Specifically, the proposed network for face modality, named as the Multi-Fusion Layers Network (MFLN), performs early fusion layers at the first block of *conv* layers by correlating the RGB data and texture descriptors, offering robust activation vector for complex data learning. On the other hand, another network for periocular modality, named as the Multi-feature Deep Layer Network (MDLN), performs late-feature fusion layers at the *fc* layers to offer better latent space features and exploit discriminatory information for robust feature learning.

To validate our network, a new multimodal dataset that contains face and periocular modalities is introduced, namely the Ethnic-facial dataset. Our dataset is designed based on five ethnic groups: *African*, *Asian*, *Latin American*, *Middle Eastern*, and *Caucasian*. The dataset is created such a way to avoid any unbalanced selections and there are huge differences between the shape and skin texture of the periocular region for each ethnic group [32].

Thus, the contributions of this paper are summarized as follows:

- Various multi-feature fusion layers across two independent dual-stream CNNs are introduced in this paper. The role of the fusion layers is to aggregate the RGB data and texture descriptors for complex data learning. Hence, both networks are benefiting from these features to deliver better accuracy performance.
- A weighted rank strategy is introduced to handle the multimodal biometrics features from two independent networks for better recognition performance. This approach incorporates the rank-*K* scores fusion effectively with the multimodal biometrics to formulate better decisions.
- A new multimodal facial biometrics dataset with face and periocular modalities, namely the Ethnic-facial dataset, is created and shared in [33]. The images were collected across large-scaled variations such as different ethnicities, appearances, locations, uncontrolled subject-camera distances, etc. The dataset includes training and testing schemes for the performance analysis and evaluation of recognition and verification tasks.

This paper is organized as follows: Section 2 describes the proposed networks with different distinct fusion strategies. The detailed dataset information is presented in Section 3. Section 4 discusses the experimental results and analysis. A conclusion is summarized in Section 5.

2. Proposed network

We propose two independent dual-stream CNNs with multimodal facial biometrics using the face and periocular modalities with multi-feature fusion layers. Both networks conceive the RGB data and texture descriptors as first and second streams, respectively. The networks are explained in detailed in the subsections.

2.1. Texture descriptors

RGB data along with texture descriptor are deployed as dual-stream inputs to form better feature representations to train the proposed networks. We utilize spatial information to capture edge appearance information by using RGB data and texture information to capture geometry surface information using the descriptors. This eliminates confounding factors and emphasizes the network's efforts on variations, such as illumination and occlusions. Thus, the proposed networks can compensate for hidden information in the multi-feature fusion layers using the observed data to represent them efficiently during training.

In our experiments, we only study on the Entropy texture and the Histogram of Oriented Gradients (HOG), which are well-known to reduce the sensitivity of image noise and levels of illumination as both descriptors create better texture information in representing objects. The descriptors are summarized as follows.

2.1.1. Entropy

The Entropy texture was designed as a statistical measurement in information theory [34]. This technique can be used to characterize uncertainty factors across the complementary information of an image. A descriptor is applied to target the difference between the neighboring pixel regions from a given image by maximizing the local context of the various images, including low-resolution and illumination.

2.1.2. HOG

HOG is a well-known descriptor to represent gradient orientations in a regular area of an image, which was introduced by Dalal and Triggs [35] for object detection. The appearance of a given image can be characterized by the distribution of intensity gradients and edge directions. We follow the implementation of Dalal and Triggs [35] to construct the HOG descriptor by using a cell size of 5×5 with 9-bin histograms, and a block is configured by grouping the 2×2 cells.

2.2. Multi-feature fusion layers and dual-stream CNNs

We propose two independent networks, known as MDLN and MFLN for handling different biometric features. Furthermore, both networks are built upon dual-stream CNNs with different multifeature fusion layers. Fig. 1 illustrates the overall architecture of our networks, which are explained in detailed as follows.

2.2.1. Architecture of MDLN

MDLN is designed to extract feature representations of a periocular modality. This is because the periocular modality itself contains complex information. To be more explicit, MDLN is focusing on late-feature fusion representations, where it takes place at the *fc* layers. The advantage is that the proposed feature fusion layers are devised to strengthen the feature activations of the network.

As shown in Fig. 1, the architecture of MDLN consists of 14 *conv* layers and 8 max-pooling (*maxpool*) layers. The *conv* layers are designed to learn the correspondence between the RGB data and texture descriptors of the periocular region and discriminate between themselves with the shared weights. Table 1 tabulates the architecture of the network.

Table 1
Configuration of each layer for MDLN. The *fc* layers are added with the ReLu layers.

Layer	Configuration
$conv_1$, $conv_2$	*f.m.*[a]: $64 \times 50 \times 150$; *k*[b]: 3×3; *maxpool*: 2×2
$conv_3$, $conv_4$	*f.m.*: $128 \times 25 \times 75$; *k*: 3×3
$conv_5$, $conv_6$	*f.m.*: $128 \times 25 \times 75$; *k*: 3×3; *maxpool*: 2×2
$conv_7$, $conv_8$	*f.m.*: $256 \times 12 \times 37$; *k*: 3×3
$conv_9$, $conv_{10}$	*f.m.*: $256 \times 12 \times 37$; *k*: 3×3; *maxpool*: 2×2
$conv_{11}$, $conv_{12}$	*f.m.*: $512 \times 6 \times 18$; *k*: 3×3
$conv_{13}$, $conv_{14}$	*f.m.*: $512 \times 6 \times 18$; *k*: 3×3; *maxpool*: 2×2
FC_1, FC_2	4096
$Fuse_{avg}$, $Fuse_{max}$	4096
$\phi_{p,1}$, $\phi_{p,2}$	*C*

[a] *f.m.* is defined as the dimension of feature maps.
[b] *k* is defined as filter size.

We proposed two fusion layers, namely $Fuse_{avg}$ and $Fuse_{max}$, to aggregate the periocular information from RGB data and texture descriptor (D), as shown in Fig. 1. The $Fuse_{avg}$ takes an average of activation from the FC_1 and FC_2 with N nodes ($N = 4096$). The layer is defined as follows:

$$Fuse_{avg} = [FC_1 + FC_2]^{1/2}, \tag{1}$$

where $FC_* = \mathbf{w}^T \cdot F_* + \mathbf{b}$ and $* \in \{1,2\}$. $\mathbf{w}^T$ is defined as weight matrix, $\mathbf{b}$ is defined as the bias matrix, and F_* is defined as the activation vectors from different input streams. On the other hand, $Fuse_{max}$ layer takes a larger activation from the FC_1 and FC_2 with N nodes ($N = 4096$). The layer can be represented as:

$$Fuse_{max} = \max[FC_1(n) + FC_2(n)], \tag{2}$$

where n is defined as the index of N nodes.

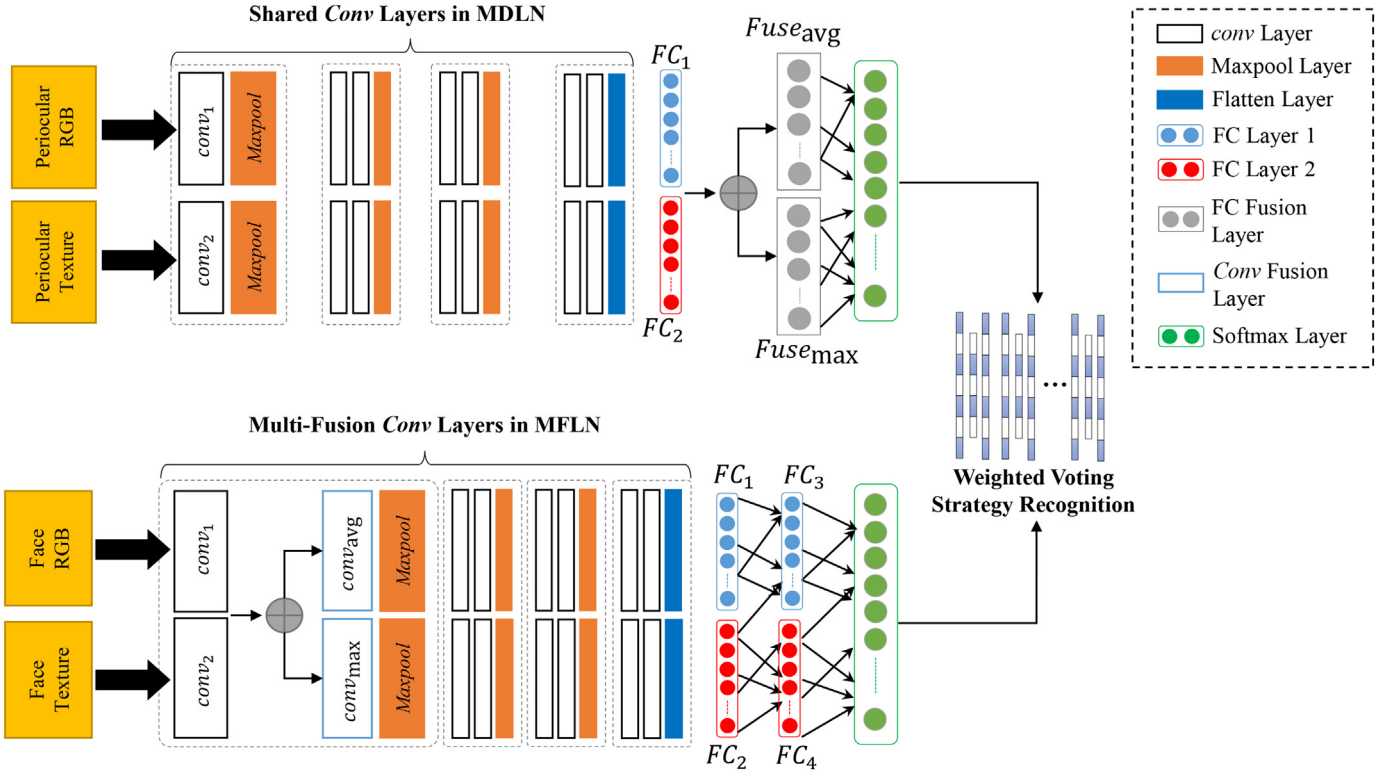

Fig. 1. The architecture of the proposed network.

A total loss function is implemented for training, which is composed of a summation of cross-entropy of logit vector of FC_{avg} and $FC_{\max}$. The encoded labels for the loss function are utilized such that:

$$total_{loss} = \mathcal{L}(FC_{\text{avg}}) + \mathcal{L}(FC_{\max}), \tag{3}$$

$$\mathscr{S}(FC_*) = -\sum_{i}^{A}\sum_{j}^{C} L_{ij}\,\log\left(\mathrm{S}(FC_*)_{ij}\right), \tag{4}$$

where $* \in \{\text{avg}, \max\}$. L, A, and C denote class labels, the number of training sample, and the number of classes, respectively. $S(\cdot)$ is defined as softmax function.

2.2.2. Architecture of MFLN

MFLN focuses on *conv* fusion to extract the hidden information between the feature maps in the earlier block of *conv* layers. To be specific, MFLN performs early fusion at the $conv_1$ and $conv_2$ layers to aggregate the RGB data and texture descriptors from the face images in order to offer robust feature activations and to complement information for complex data learning.

The architecture of MFLN consists of 14 *conv* layers, 2 *conv* fusion layers, and 8 *maxpool* layers. The *conv* fusion layers are designed to extract better feature representations across several *conv* layers so that the network learns the correspondence between the inputs (RGB data and texture descriptors of the face) robustly and discriminate between themselves with the shared weights. Table 2 summarizes the entire architecture and configurations of MFLN.

Two *conv* fusion layers, namely $conv_{\text{avg}}$ and $conv_{\max}$, are proposed to aggregate the face features from the RGB images and texture descriptor, as shown in Fig. 1. The $conv_{\text{avg}}$ layer computes the average of activation of the feature maps from the $conv_1$ and $conv_2$ to employ the arbitrary correspondence to its best effect. Let us denote the layer as follows:

$$conv_{\text{avg}} = \mathbf{w}^{\mathrm{T}}\left[(conv_1 + conv_2)^{1/2} \cdot \mathbf{K}\right] + \mathbf{b}, \tag{5}$$

where $\mathbf{K}$ is defined as the filter matrix. The rest of the variables' description is similar in Eq. (1). On the other hands, the $conv_{\max}$ layer takes the largest activations in the feature maps of the earlier *conv* layers to employ the arbitrary correspondence by extracting each feature activation at the prior layer. The layer can be represented as:

$$conv_{\max} = \mathbf{w}^{\mathrm{T}}[H \cdot \mathbf{K}] + \mathbf{b}, \tag{6}$$

where $H = \max_{\mathrm{e}}[conv_1(m), conv_2(m)]$ and $\max_{\mathrm{e}}[\cdot]$ denotes a function that finds the maximum element-wise values between the feature maps of $conv_1$ and $conv_2$. The rest of the variables' description is similar in Eq. (5). For training, we also implement a total loss function that composed of summation of cross-entropy of logit vector of FC_3 and FC_4 based on Eqs. (3) and (4).

Table 2
Configuration of the MFLN. The *fc* layers are added with the ReLu layers.

Layer	Configuration
$conv_1$, $conv_2$	*f.m.*: 64 × 128 × 128; *k*: 3 × 3
$conv_{\text{avg}}$, $conv_{\max}$	*f.m.*: 64 × 128 × 128; *k*: 3 × 3; *maxpool*: 2 × 2
$conv_3$, $conv_4$	*f.m.*: 128 × 64 × 64; *k*: 3 × 3
$conv_5$, $conv_6$	*f.m.*: 128 × 64 × 64; *k*: 3 × 3; *maxpool*: 2 × 2
$conv_7$, $conv_8$	*f.m.*: 256 × 32 × 32; *k*: 3 × 3
$conv_9$, $conv_{10}$	*f.m.*: 256 × 32 × 32; *k*: 3 × 3; *maxpool*: 2 × 2
$conv_{11}$, $conv_{12}$	*f.m.*: 512 × 16 × 16; *k*: 3 × 3
$conv_{13}$, $conv_{14}$	*f.m.*: 512 × 16 × 16; *k*: 3 × 3; *maxpool*: 2 × 2
FC_1, FC_2	4096
FC_3, FC_4	4096
$\phi_{\mathrm{f},1}$, $\phi_{\mathrm{f},2}$	*C*

2.2.3. Weighted voting layer

We propose a weighted voting layer to merge the distance scores from the softmax vectors for decision-making. Let $\phi_{*,1} = \text{softmax}(FC)$ and $\phi_{*,2} = \text{softmax}(FC)$ be the softmax vectors of the last latent features. Since our network is trained with face and periocular modalities, we differentiate the softmax vector ϕ_{f} as face and ϕ_{p} as periocular. Each individual ϕ_* contains ϕ_{f} and ϕ_{p} as the sum of its corresponding $\phi_* = \phi_{*,1} + \phi_{*,2}$.

We evaluate the proposed network through two common tasks: recognition and verification. To recognize an unknown identity, the testing data are divided into a gallery and probe set. The gallery set of each individual is composed of his/her softmax vectors as $\phi_j^G = \{\phi_{j,\mathrm{f}}^G, \phi_{j,\mathrm{p}}^G\}$, where $j = 1, 2, 3, \ldots, C$ and the probe set is represented as $\phi^U = \{\phi_{\mathrm{f}}^U, \phi_{\mathrm{p}}^U\}$. Then, we compute a weighted voting strategy $\boldsymbol{\omega}$ with sum rule and rank-K function as follows:

$$\boldsymbol{\omega}\left(\phi^U, \phi_j^G\right) = \mathrm{rank}_K\left(\phi_{\mathrm{f}}^U, \phi_{j,\mathrm{f}}^G\right) + \mathrm{rank}_K\left(\phi_{\mathrm{p}}^U, \phi_{j,\mathrm{p}}^G\right) \tag{7}$$

where $\mathrm{rank}_K(\cdot)$ denotes a function that sums up the top-K highest cosine similarity distance between the probe set and each subject of the gallery set. Finally, the recognition δ is defined as follows:

$$\delta = \max_j\left[\boldsymbol{\omega}\left(\phi^U, \phi_j^G\right)\right]. \tag{8}$$

A verification protocol refers to the process of verifying an individual's identity that is claimed as either a genuine or an impostor. Let $\phi^U = \{\phi_{\mathrm{f}}^R, \phi_{\mathrm{p}}^R\}$ as the reference set and $\phi^Q = \{\phi_{\mathrm{f}}^Q, \phi_{\mathrm{p}}^Q\}$ as the query set, to verify ϕ^Q is genuine or impostor, ν is decided by using Eq. (7) as follows:

$$\nu = \begin{cases} 1, & \boldsymbol{\omega}\left(\phi^U, \phi_j^G\right) \le t \\ 0, & \boldsymbol{\omega}\left(\phi^U, \phi_j^G\right) < t \end{cases}, \tag{9}$$

where t is defined as the dependence threshold value.

3. Ethnic-facial dataset

We propose this new dataset to support a balanced collection of multimodal facial biometrics images among different ethnicities. In addition, we also ensured that all the images are collected in common and everyday settings, such as appearances with and without make-up, locations, level of illuminations, poses, and uncontrolled subject-camera distances.

3.1. Collection setup

To design our dataset, we followed the example of the VGG Face [36] dataset collection. We then randomly selected 1062 subjects' names from BBC News [37], CNN News [38], Fox News [39], Naver News [40], Phoenix [41], and Sin Chew Daily [42], in order to search for the images of these subjects across Google's image search engine. In the search, the top 400 images for each subject were downloaded. The views of the facial region in these images were between −60° and 60°. Then, the images were manually verified to ensure that the images are correctly labeled by the subjects. This dataset contains 188,756 images across 1062 subjects.

Next, to extract the face and periocular regions from each image, we first aligned all the images by fixing the coordinates of facial feature points based on the Viola-Jones face detector bounding box. Then, the images were cropped into the face by using the technique from [43]. The results were resized to 128 × 128 individually as shown in Fig. 2. For the periocular regions, we also implemented the same technique from [43] by fixing the coordinates of periocular feature points based on the Viola-Jones face detector bounding box. Then, the images were

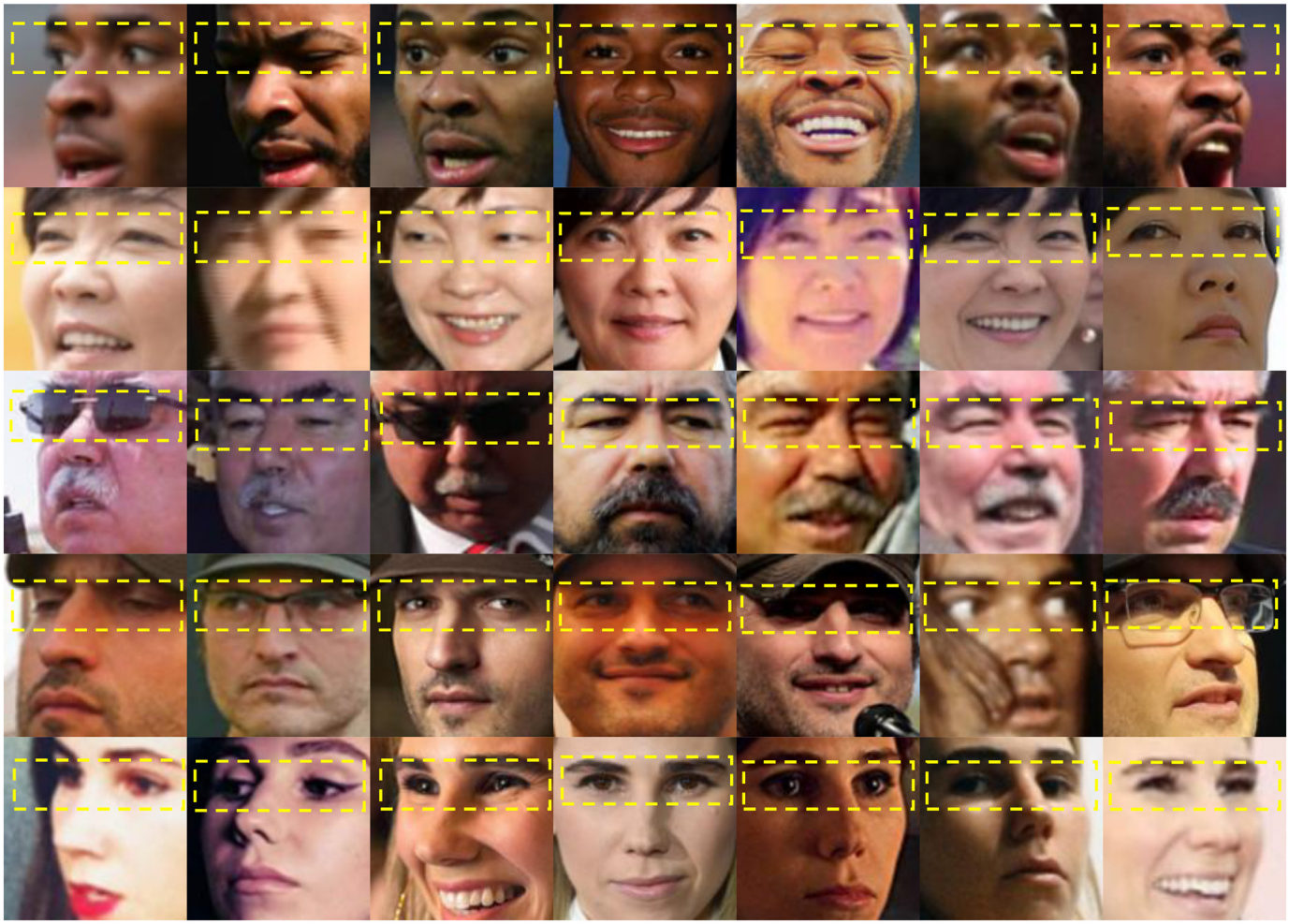

Fig. 2. The sample images of our dataset. Each row represents an individual with different ethnic groups, such as *African*, *Asian*, *Middle Eastern*, *Latin American*, and *Caucasian*. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

cropped into the periocular region (yellow dotted box) and the results were resized to 50 × 150 individually as shown in Fig. 2.

3.2. Dataset protocol

The dataset provided training and benchmark protocols; 733 subjects were randomly selected as training and the rest of the subjects were used as benchmarks. Note that no subjects for training overlapped with the benchmarking set. To develop our own networks, we designed the protocol by dividing the images for each subject with the ratio of training and validation at 70:30.

In the benchmarking scheme, we designed the recognition and verification tasks. For the recognition task, the task was to decide which of the identifies was represented by the probe set. In the experiments, we divided the images per subject by selecting five images as gallery set and the remaining images as a probe set. This selection process was repeated three times. For the verification task, the goal was to verify two sets of biometric images and decide whether the claim was represented as genuine or impostor. In the experiments, we randomly selected 500 reference-query pairs as 'same' labels and another 500 pairs as 'not same'. The selection process was also repeated three times.

4. Experiments

We selected our dataset - Ethnic-facial and four public datasets, namely AR [44], FaceScrub [45], IMDB_WIKI [46], and YouTube Face (YTF) [47] as the target datasets to evaluate the performance comparison of recognition and verification tasks between our network and other benchmark networks. All the configurations of networks are described next.

4.1. Experimental setup

4.1.1. Configuration of proposed networks

Our network was implemented using TensorFlow [48]. For the configurations of MDLN and MFLN, the learning rate was defined as 1.0×10^{-4} Adam Optimizer was applied to both MDLN and MFLN, where the weight decay and momentum were set to 1.0×10^{-4} and 0.9, respectively.

In our experiments, the batch size was set to 64 and the training was carried out for 1000 epochs. The training was done by using 577 subjects from the VGG face dataset and 733 subjects from Ethnic-facial by following the protocols that were mentioned in Section 3.2; thereby encompassing 1310 subjects across 192,478 images were used. Note that both MDLN and MFLN were trained independently; it was performed by using Nvidia Titan Xp GPUs.

4.1.2. Configuration of benchmark networks

Several popular benchmark networks in face recognition were selected, namely AlexNet [49], FaceNet [6], LCNN [50], multi-level abstraction fusion CNN [26], ResNet [51], and VGG Face [36]. These networks have been proven to be successful in very large recognition and

verification tasks. In our experiments, we utilized all the pre-trained models that were provided by the respective authors, except for AlexNet, Multi-level abstraction fusion CNN, and ResNet. In the case of Multi-level abstraction fusion CNN, the network is not publicly available. We therefore did our best effort to implement the network from scratch by following [26]. Likewise, for AlexNet and ResNet, we performed fine-tuning to improve the networks themselves by training with our dataset and VGG Face dataset as mentioned in Section 4.1.1.

4.1.3. Configuration of benchmark datasets

Four public datasets, namely AR [44], FaceScrub [45], IMDB_WIKI [46], and YouTube Face (YTF) [47], which are selected in our studies. These datasets fulfill the scenarios of controlled and uncontrolled environments. However, all the datasets were not designed for periocular recognition; we therefore implemented the technique from [43] to crop the periocular regions by using the given coordinates of facial feature points.

4.2. Experimental results

4.2.1. Performance analysis on proposed network

This section analyzes the robustness and performance of the proposed network using the AR dataset [44]. The dataset consists of 100 subjects with neutral, expression (*Exp*), illumination (*Illum*), and scarf conditions that were captured across controlled environments with two sessions. For evaluation, we designed the experimental protocols as the following cases:

- '*Exp* + *Illum*': 7 non-occluded images for each subject from Session 1 were used as gallery set, and another 7 non-occluded images for each subject from Session 2 were used as probe set;
- '*Scarf*': 7 non-occluded images per subject from Session 1 were used as gallery set, 12 scarf occlusion images per subject from both sessions were used as probe set;
- '*Blur*': applied a Gaussian blur to all the images from Session 2 as a probe set with blurring effects (see Fig. 3(a)) by increasing the σ values from 1 to 5;
- '*Occlusion*': created random 'occlusion box' to all images (see Fig. 3(b)) from Session 2 as a probe set by increasing its size.

Table 3 shows that the proposed network using RGB + Entropy achieved the highest Rank-1 recognition accuracies across the '*Exp* + *Illum*' and '*Scarf*' cases with 98.57% and 94.33%, respectively.

Table 3
Performance evaluation of the recognition task on the AR dataset. The highest accuracy is written in bold.

Network	*Exp* + *Illum* (%)	*Scarf* (%)
Proposed network (using RGB + Entropy)	**98.57**	**94.33**
Proposed network (using RGB + Entropy and score fusion)	98.00	92.86
Proposed network (using RGB + HOG)	97.43	92.86
Tiong et al. [4]	98.00	93.00
Multimodal CNN[a] with RGB data	94.00	90.14
Multimodal CNN with Entropy	83.57	68.57
Multimodal CNN with HOG	82.00	67.86
CNN (Face) with RGB data	92.43	76.00
CNN (Periocular) with RGB data	80.29	79.14

[a] Multimodal CNN refers to two CNNs that accept face and periocular modalities, respectively. The networks used score fusion approach to formulate final decision-making.

Besides, the proposed network using RGB + HOG achieved Rank-1 recognition accuracies across '*Exp* + *Illum*' and '*Scarf*' cases with 97.43% and 92.86%, respectively. In addition, we also evaluated our network using score fusion approach and Tiong et al. [4]; both networks achieved 98.00% for the '*Exp* + *Illum*' cases. For the case of '*Scarf*', Tiong et al. [4] achieved 93.00% as the second-best Rank-1 recognition accuracy and the proposed network using score fusion approach only achieved 92.86%.

As compared to a multimodal CNN, the network using RGB data only achieved the Rank-1 accuracies across '*Exp* + *Illum*' and '*Scarf*' cases with 94% and 90.14%, respectively. Furthermore, multimodal CNN using Entropy and HOG descriptors only achieved 83.43% and 82%, respectively for the '*Exp* + *Illum*' case. Both multimodal CNN using Entropy and HOG descriptors only achieved 68.57% and 67.86% accuracies, respectively. Besides, we also implemented CNN with face and periocular modalities for comparison, respectively. For the '*Exp* + *Illum*' case, CNN with face modality achieved 92.43%. Interestingly, CNN with periocular modality attained to achieve 79.14% for '*Scarf*' case.

As can be seen in Table 3, the proposed network using RGB + Entropy achieved the highest recognition accuracies across '*Exp* + *Illum*' and '*Scarf*' cases. These results indicate that our network provides more complementary information than other networks. Furthermore, we also evaluate the importance of a periocular modality for the challenge of appearances with scarf/masks.

Fig. 4 visualizes the robustness performance between the proposed network and others for '*blur*' and '*occlusion*' cases. Through the analysis,

Fig. 3. A sample of probe images for the (a) '*blur*' and (b) '*occlusion*' cases. The yellow boxes are defined as the periocular modality.

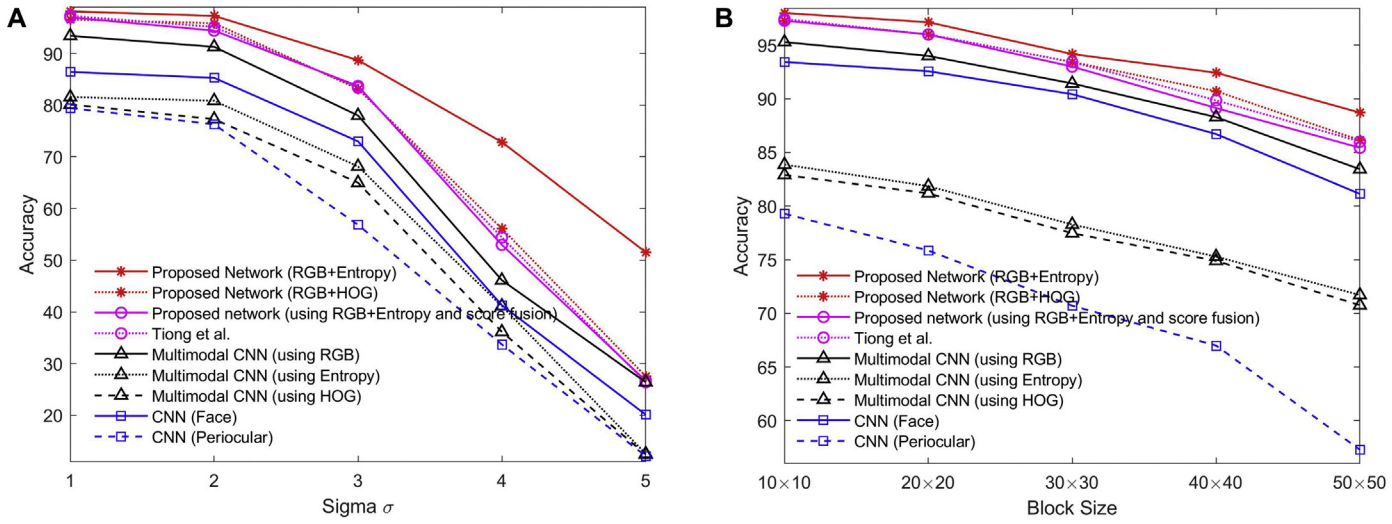

Fig. 4. The performance of recognition on AR database with (a) '*blur*' and (b) '*occlusion*' cases.

our network using RGB + Entropy achieved at least 89.5% accuracy for the '*occlusion*' case, and also achieved at least 88% accuracy for the 'blurring' case when $\sigma \leq 3$. As compared to multimodal CNN, the network using RGB data could only achieve 88% accuracy with 10×10 'occlusion box' and less than 80% accuracy with other sizes for the '*occlusion*' case. Furthermore, the network only achieved 75% accuracy for the '*blur*' case when $\sigma = 1$ and did not perform well when $\sigma \geq 2$.

As can be observed, the proposed multi-feature fusion layers were fully utilized in our network to aggregate the RGB data and texture descriptors, overcoming the limitations of RGB data. Thus, our network successfully transformed new knowledge representations to perform better recognition accuracy by discovering rich features information. In addition, to overcome the complexities of multimodal data and formulate a decision, the weighted voting strategy preserves the robustness of our network in contributing towards better decision-making from multimodal biometrics to achieve better recognition performance. Besides, texture descriptors are only beneficial to support the complex learning instead of using as standalone input. This is because the descriptors could not represent precisely the observed high-dimensional features.

4.2.2. Performance evaluation on recognition tasks

This section presents the experimental results on the recognition task by conducting several public datasets "in-the-wild" environments. We used our dataset – Ethnic-Facial and two public datasets, the FaceScrub and IMDB_WIKI, to evaluate the performance of the proposed network and other benchmark networks. we evaluated the performance by using a Cumulative Matching Characteristic (CMC) curve with a 95% confidence interval (CI). All the experimental results are outlined in the following sub-section.

Evaluation on FaceScrub dataset

To evaluate whether our network performs well on standard datasets, we tested its performance on a more subjective experiment with FaceScrub dataset. This dataset is a real-life dataset that contains 530 subjects. The images were collected from the Internet under uncontrolled environmental conditions, which contained different appearances, poses, illuminations, expressions and time. The details of the dataset were described in Ng and Winkler [45]. As a performance comparison with the benchmark networks, the experimental protocol for the recognition task was designed by dividing the images of each subject into three groups. We selected one of them as a gallery set, the remaining two groups as probe sets. The division process was repeated three times.

According to Table 4, the proposed network achieved the highest average accuracies for Rank-1 and Rank-5 recognition with $93.86 \pm 1.3\%$ and $97.5 \pm 0.7\%$ accuracies. Besides, the multi-level abstraction fusion CNN achieved the second-best performance with $90.52 \pm 1.1\%$ and $96.45 \pm 0.6\%$ for Rank-1 and Rank-5 recognition accuracies. We also present the Rank-1 to Rank-10 recognition results in Fig. 5. As can be seen in the figure, our network achieved the best result among the benchmark networks. The result indicates that the proposed network is capable of learning the features of the RGB data and texture descriptor decently for improving the performance of recognition.

Evaluation on IMDB_WIKI dataset

We also conducted another more challenging experiment with the IMDB_WIKI dataset to verify the robustness of the proposed network. This dataset consists of 100,000 subjects whereby the images are assigned based on the age and timestamp information related to individuals [46]. Since the data itself was not well-organized for facial recognition, we performed data re-arrangement by means of removing the

Table 4
Performance evaluation of the recognition task on the FaceScrub, IMDB_WIKI, and Ethnic-facial datasets. The highest accuracy is written in bold.

Networks	FaceScrub (%)		IMDB_WIKI (%)		Ethnic-Facial (%)	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
AlexNet '12	58.88 ± 7.1	80.19 ± 1.5	21.60 ± 6.5	35.90 ± 2.8	48.58 ± 3.4	62.14 ± 0.6
FaceNet '15	89.78 ± 2.5	96.01 ± 0.5	63.41 ± 4.3	79.62 ± 1.9	82.66 ± 1.3	88.67 ± 0.5
VGG Face '15	86.67 ± 3.2	95.18 ± 0.9	57.98 ± 5.7	77.32 ± 2.3	80.48 ± 1.6	87.96 ± 0.7
ResNet '16	89.49 ± 2.9	95.42 ± 0.7	60.49 ± 3.5	78.33 ± 1.9	80.57 ± 1.4	89.04 ± 0.5
LCNN '18	84.68 ± 3.4	89.32 ± 0.8	61.47 ± 3.1	79.01 ± 1.6	81.58 ± 1.8	89.35 ± 0.6
Multi-level abstraction fusion CNN '18	90.52 ± 1.1	96.45 ± 0.6	67.22 ± 3.4	80.61 ± 1.8	83.19 ± 1.7	90.90 ± 0.4
Our network	**93.86 ± 1.3**	**97.50 ± 0.7**	**73.11 ± 4.2**	**86.47 ± 2.1**	**89.03 ± 1.7**	**96.60 ± 0.5**

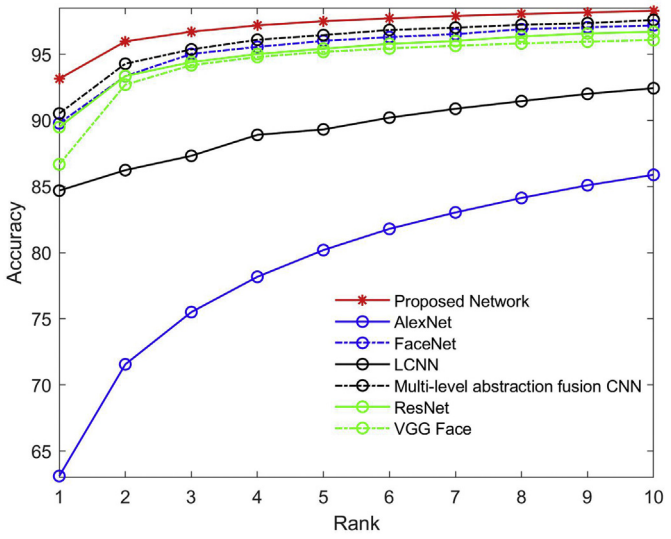

Fig. 5. Performance comparison for the recognition task on FaceScrub dataset.

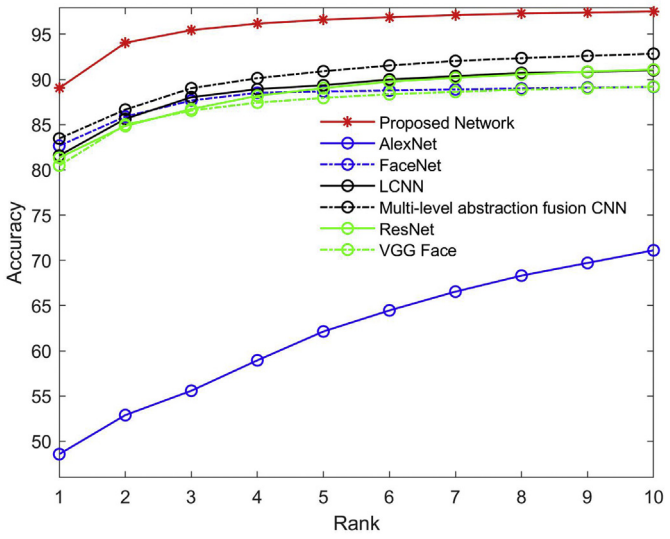

Fig. 7. Performance comparison for the recognition task on Ethnic-facial dataset.

images that are not representative of the genuine. In this experiment, we have only selected 2129 subjects, and each subject contained at least 15 images; the total number of images was 89,424. For the evaluation protocols, we divided the images such that the ratio between the gallery sets and probe sets is 40:60. This division process was repeated three times.

Table 4 presents that our network achieved the highest average Rank-1 and Rank-5 recognition accuracies with 73.11 ± 4.2% and 86.47 ± 2.1%, respectively. The second best was achieved by the Multi-level abstraction fusion CNN with 67.22 ± 3.38% and 80.61 ± 1.8% as Rank-1 and Rank-5 accuracies, respectively. Fig. 6 illustrates the CMC curve of the proposed network, showing that the network outperformed other benchmark networks from Rank-1 to Rank-10 recognition accuracies. The results indicated that the deterministic fusion layers are capable of correlating the RGB data and texture descriptors.

Evaluation on Ethnic-Facial dataset

We present the experimental results in Table 4 by following the recognition protocol, which is mentioned in Section 3.1. When compared against several benchmark networks (see Table 4), our network achieved the highest Rank-1 and Rank-5 recognition accuracies with 89.03 ± 1.7% and 96.6 ± 0.4%. Fig. 7 illustrates the CMC curve of the proposed network, which shows that the network outperformed other benchmark networks from Rank-1 to Rank-10 recognition accuracies. The results prove that our network can learn new features from the multi-feature fusion layers in order to transfer new knowledge between the networks to perform better recognition performance.

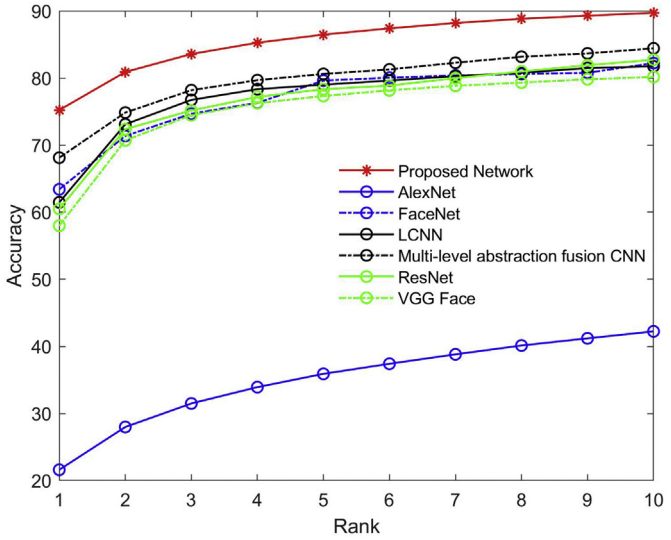

Fig. 6. Performance comparison for the recognition task on IMDB_WIKI dataset.

4.2.3. Performance evaluation on verification tasks

This section conducted the performance comparison for verification. We selected our dataset and YTF datasets for the evaluation. We reported the performance Receiver Operating Characteristic (ROC) curve and area under the ROC curve (AUC) for each dataset's evaluation.

Evaluation on YTF dataset

YTF is a real-life video dataset that consists of 1595 subjects from the YouTube [47]. The videos have been acquired with a wide range of appearance variations, poses, and 'super' low-resolution. To evaluate the robustness of the proposed network for verification task, we followed the protocols reported by [52] by selecting the subjects that at least have four or more videos. We then randomly selected 500 reference-query pairs as 'same' labels and another 500 pairs as 'not same'. The selection process was repeated three times. According to Table 5, the proposed network achieved the lowest EER as 16.47 ± 1.48% and AUC as 0.9084. Multi-level abstraction fusion CNN attained second-lowest performance with 17.74 ± 1.53% for EER and 0.8946 as AUC. Fig. 8 illustrates the ROC curve, which demonstrates that our network obtained the best performance of AUC and the lowest EER.

Evaluation on Ethnic-Facial dataset

We presented the experimental results in Table 5 by following the verification protocol as mentioned in Section 3.2. For the results of verification, our proposed network achieved the lowest EER 5.76 ± 0.43% and 0.9933 AUC, respectively. Except for AlexNet, all the benchmark networks achieved the EER between 6.8% and 10%. Fig. 9 illustrates the ROC curve of our proposed network, which outperformed other networks with respect to all the benchmarks for the best performance of AUC. This is evidence to demonstrate our network outperformed most of the benchmark networks and achieved the highest recall rate against other networks. Besides, the results proved that our network can learn new feature representations from the fusion layers for better verification.

Table 5
Performance evaluation of the verification task on the YTF and Ethnic-facial datasets. The lowest EER is written in bold.

Networks	YTF		Ethnic-Facial	
	EER (%)	AUC	EER (%)	AUC
AlexNet '12	49.69 ± 3.24	0.5899	40.82 ± 1.87	0.6799
FaceNet '15	18.37 ± 1.95	0.8776	8.21 ± 0.75	0.9733
VGG Face '15	19.26 ± 2.42	0.8686	10.03 ± 0.83	0.9625
ResNet '16	18.77 ± 1.67	0.8636	9.74 ± 1.13	0.9703
LCNN '18	18.08 ± 1.77	0.8812	7.59 ± 0.25	0.9772
Multi-level abstraction fusion CNN '18	17.74 ± 1.53	0.8946	6.87 ± 0.69	0.9811
Our network	**16.47 ± 1.48**	**0.9084**	**5.76 ± 0.61**	**0.9933**

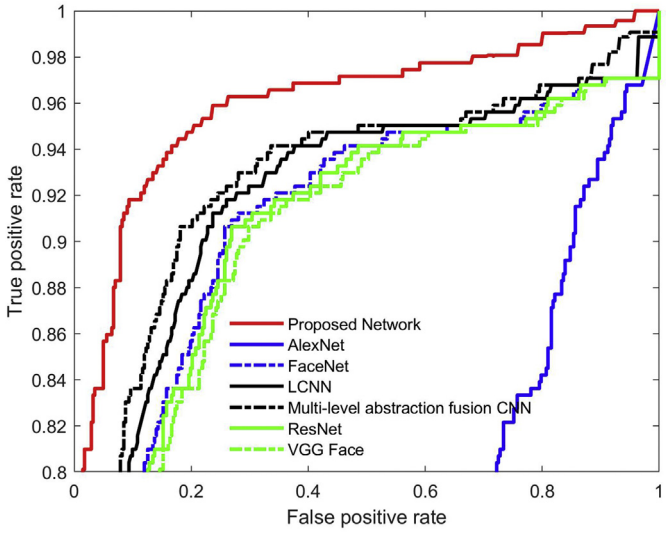

Fig. 8. Performance comparison for the verification task on YTF dataset.

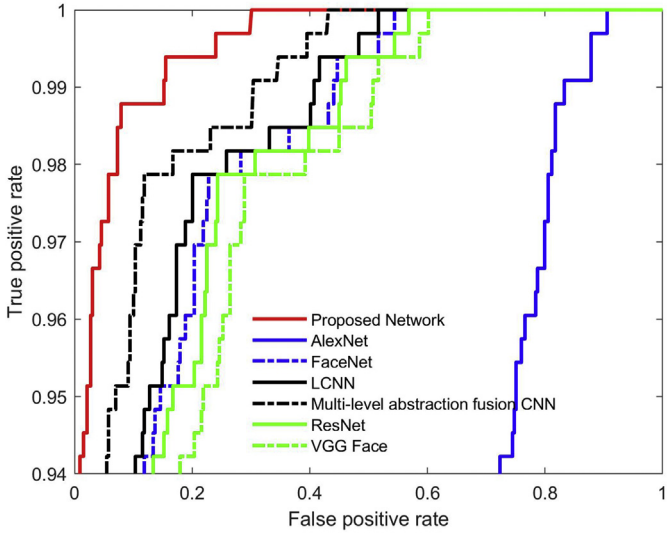

Fig. 9. Performance comparison for the verification task on Ethnic-facial dataset.

4.3. Discussion

Several observations can be listed through experimental analysis and results. First, the evaluation results in Section 4.2 reported that only using RGB as input does not provide the best performance for recognition and verification tasks. Most of the existing deep learning networks are focusing on filtering out confounding factors such as illumination and occlusion. Hence, using our network can exploit the discriminatory features through the RGB data and texture descriptors for better recognition. The proposed network utilizes texture information to enrich latent and complement information for complex data learning, which contributes to a more robust representation for the challenges of surveillance.

Our experimental results also proved that the proposed network is able to achieve better performance due to its ability to learn the complexities of multimodal data by using the proposed multi-feature fusion layers. Specifically, with the real scenarios of uncontrolled environments and ethnic group effects, the weighted voting strategy preserves the robustness of our networks and formulate better decision-making in various datasets. The effectiveness of the proposed fusion layers and weight voting strategy provides strong support towards our assumption confidently such that multi-feature learning can achieve better results than using raw data or unimodal biometric data.

5. Conclusion

This paper proposed the design of multi-feature fusion layers that contributed to offering a more robust feature representation in multimodal facial biometrics recognition. By aggregating the dual inputs (RGB data and texture descriptors) into the network fusion layers, the proposed network achieved better accuracy performance by learning new features. We also collected a new Ethnic-facial dataset, which consisted of a large collection of multimodal biometrics images based on different ethnicities and uncontrolled environments. Through the extensive experiments by comparing with several networks on our dataset and other available datasets, the proposed network achieved better performance in both recognition and verification tasks under controlled and uncontrolled environments.

However, this work is still limited to some extreme cases of individuals who are wearing “large and wider” sunglass. In the future, we plan to study generative models to recover and predict such cases for identifying criminal suspects. In addition, we shall incorporate the gait analysis, which is useful to identify the terror behaviors of the suspects in real-time.

Declaration of Competing Interest

We would like to declare that the authors do not have any potential conflicts.

Acknowledgements

This work was supported by the Brain Korea 21 Plus Project.

References

[1] A.K. Jain, K. Nandakumar, A. Ross, 50 years of biometric research: accomplishments, challenges, and opportunities, Pattern Recogn. Lett. 79 (2016) 80–105, https://doi.org/10.1016/j.patrec.2015.12.013.

[2] Y. Chen, J. Yang, C. Wang, N. Liu, Multimodal biometrics recognition based on local fusion visual features and variational Bayesian extreme learning machine, Expert Syst. Appl. 64 (2016) 93–103, https://doi.org/10.1016/j.eswa.2016.07.009.

[3] G. Goswami, P. Mittal, A. Majumdar, M. Vatsa, R. Singh, Group sparse representation based classification for multi-feature multimodal biometrics, Inf. Fusion 32 (2016) 3–12, https://doi.org/10.1016/j.inffus.2015.06.007.

[4] L.C.O. Tiong, S.T. Kim, Y.M. Ro, Implementation of multimodal biometric recognition via multi feature deep learning networks and featur fusion, Multimedia Tools Appl. 78 (16) (2019) 22743–22772, https://doi.org/10.1007/s11042-019-7618-0.

[5] S. Poria, E. Cambria, A. Gelbukh, Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, Proceeding of Conference on Empirical Methods on Natural Language Process, Lisbon, Portugal 2015, pp. 2539–2544.

[6] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, International Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA 2015, pp. 815–823.

[7] Y. Wang, T. Bao, C. Ding, M. Zhu, Face recognition in real-world surveillance videos with deep learning method, International Conference on Image, Vision and Computing, IEEE 2017, pp. 239–243, https://doi.org/10.1109/ICIVC.2017.7984553.

[8] J.H. Koo, N.R. Baek, M.C. Kim, K.R. Park, CNN-based multimodal human recognition in surveillance environments, Sensors 18 (2018) 1–34, https://doi.org/10.3390/s18093040.

[9] D. White, J.D. Dunn, A.C. Schmid, R.I. Kemp, Error rates in users of automatic face recognition software, PLoS One 10 (10) (2015) 1–14, https://doi.org/10.1371/journal.pone.0139827.

[10] Case study: the impetus for changing border security paradigms to stem trans-regional migration crisis and global open border terrorism through an innovative, integrated approach, Tech. Rep. Securiport, 2018.

[11] E. Mckone, L. Wan, M. Pidcock, K. Crookes, K. Reynolds, A. Dawel, E. Kidd, C. Fiorentini, A critical period for faces: other-race face recognition is improved by childhood but not adult social contact, Sci. Rep. 9 (2019) 1–13, https://doi.org/10.1038/s41598-019-49202-0.

[12] D. Ramachandram, G.W. Taylor, Deep multimodal learning: a survey on recent advances and trends, IEEE Signal Proc. Mag. (November) (2017) 96–108, https://doi.org/10.1109/MSP.2017.2738401.

[13] Z. Cheng, X. Zhu, S. Gong, Surveillance Face Recognition Challenge, arXiv Preprint, 2018 1–30 (arXiv:arXiv:1804.09691v6).

[14] P. Grother, M. Ngan, K. Nanaoka, Ongoing face recognition vendor test (FRVT) part 2: identification, Tech. Rep. NIST, 2018https://doi.org/10.6028/NIST>IR.8238.

[15] Z. Wang, K. He, Y. Fu, R. Feng, Y.-G. Jiang, X. Xue, Multi-task deep neural network for joint face recognition and facial attribute prediction, International Conference on Multimedia Retrieval, ACM 2017, pp. 365–374.

[16] M. Zabłocki, K. Go, D. Frejlichowski, R. Hofman, Intelligent video surveillance systems for public spaces – a survey, J. Theor. Appl. Comput. Sci. 8 (4) (2014) 13–27.

[17] Y. Xu, Y. Lu, Adaptive weighted fusion: a novel fusion approach for image classification, Neurocomputing 168 (2015) 566–574, https://doi.org/10.1016/j.neucom.2015.05.070.

[18] N. Srivastava, R. Salakhutdinov, Learning representations for multimodal data with deep belief nets, International Conference on Machine Learning Workshop, Edinburgh, Scotland, UK 2012, pp. 1–8.

[19] S. Zagoruyko, N. Komodakis, Learning to compare image patches via convolutional neural networks, International Conference on Computer Vision and Pattern Recognition (CVPR), IEEE 2015, pp. 4353–4361.

[20] S.E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-lewandowski, R.C. Ferrari, M. Mirza, D. Warde-farley, C. Aaron, P. Vincent, R. Memisevic, C. Pal, Y. Bengio, D. Warde-farley, EmoNets: multimodal deep learning approaches for emotion recognition in video, J. Multimodal User Interfaces 10 (2) (2016) 99–111, https://doi.org/10.1007/s12193-015-0195-2.

[21] Y. Liu, Y. Guo, T. Georgiou, M.S. Lew, Fusion that matters: convolutional fusion networks for visual recognition, Multimed. Tools Appl. 77 (2018) 29407–29434, https://doi.org/10.1007/s11042-018-5691-4.

[22] M. Simonovsky, B. Gutierrez-Becker, D. Mateus, N. Navab, N. Komodakis, A deep metric for multimodal registration, Lecture Notes in Computer Science, vol. 9902, Springer, Cham 2016, pp. 10–18, https://doi.org/10.1007/978-3-319-46726-9.

[23] Q. Zhang, H. Li, Z. Sun, T. Tan, Deep feature fusion for iris and periocular biometrics on mobile devices, IEEE Trans. Inf. Forensics Secur. 13 (11) (2018) 2897–2912, https://doi.org/10.1109/TIFS.2018.2833033.

[24] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, International Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, Nevada, USA 2016, pp. 1933–1941.

[25] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S.Z. Li, T. Hospedales, When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition, Internationa Conference on Computer Vision Workshop (ICCVW), IEEE, Santiago, Chile 2015, pp. 4321–4329.

[26] S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson, N.M. Nasrabadi, Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification, International Conference on Pattern Recognition (ICPR), IEEE, Beijing, China 2018, pp. 3469–3476.

[27] G. Levi, T. Hassner, Emotion recognition in the wild via convolutional neural networks and mapped binary patterns, International Conference on Multimodal Interaction, ACM, Seattle, Washington, USA 2015, pp. 503–510.

[28] R.M. Anwer, F.S. Khan, J. van de Weijer, M. Molinier, J. Laaksonen, Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification, ISPRS J. Photogramm. Remote Sens. 138 (2018) 74–85, https://doi.org/10.1016/j.isprsjprs.2018.01.023.

[29] M. Latonero, K. Hiatt, A. Napolitano, G. Clericetti, M. Penagos, Digital Identity in the Migration & Refugee Context: Italy Case Study, 2019.

[30] J.S.d. Rio, D. Moctezuma, C. Conde, I.M. de Diego, E. Cabello, Automated border control e-gates and facial recognition systems, Computers & Security 62 (2016) 49–72, https://doi.org/10.1016/j.cose.2016.07.001.

[31] J.C. Klontz, A.K. Jain, A case study of automated face recognition: the Boston marathon bombings suspects, Computer 46 (2013) 91–94.

[32] S.C. Rhee, K.-s. Woo, B. Kwon, Biometric study of eyelid shape and dimensions of different races with references to beauty, Aesthet. Plast. Surg. 36 (2012) 1236–1245, https://doi.org/10.1007/s00266-012-9937-7.

[33] Ethnic-Facial Dataset, URL https://drive.google.com/drive/folders/11c3m_TXxZWo9KJcxphWOidvy4nC1q9kl?usp=sharing 2018.

[34] M.H. Bharati, J.J. Liu, J.F. MacGregor, Image texture analysis: methods and comparisons, Chemom. Intell. Lab. Syst. 72 (1) (2004) 57–71, https://doi.org/10.1016/j.chemolab.2004.02.005.

[35] N. Dalal, W. Triggs, Histograms of oriented gradients for human detection, International Conference on Computer Vision Pattern Recognition (CVPR), IEEE, San Diego 2005, pp. 886–893, https://doi.org/10.1109/CVPR.2005.177.

[36] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, British Machine Vision Conference 2015, pp. 1–12.

[37] BBC News, , URL https://www.bbc.com/news.

[38] CNN News, , URL https://edition.cnn.com.

[39] Fox News, , URL https://www.foxnews.com.

[40] Naver News, , URL http://news.naver.com/.

[41] Phoenix CNE, , URL http://www.pcne.tv.

[42] Sin Chew Daily, , URL https://www.sinchew.com.my.

[43] V. Štruc, N. Pavešić, The complete Gabor-Fisher classifier for robust face recognition, EURASIP J. Adv. Signal Proc. 2010 (2010) 1–26, https://doi.org/10.1155/2010/847680.

[44] A. Martinez, R. Benavente, The AR face database, Tech. Rep. The Ohio State University, Barcelona, 1998.

[45] H.W. Ng, S. Winkler, A data-driven approach to cleaning large face datasets, International Conference on Image Processing (ICIP), IEEE 2014, pp. 343–347, https://doi.org/10.1109/ICIP.2014.7025068.

[46] R. Rothe, R. Timofte, L. Van Gool, Deep expectation of real and apparent age from a single image without facial landmarks, Int. J. Comput. Vis. 126 (2–4) (2016) 144–157, https://doi.org/10.1007/s11263-016-0940-3.

[47] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, International Conference on Computer Vision Pattern Recognition (CVPR), IEEE, Colorado Springs, CO, USA 2011, pp. 529–534.

[48] TensorFlow, , URL https://tensorflow.org.

[49] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA 2012, pp. 1097–1105.

[50] X. Wu, R. He, Z. Sun, T. Tan, A light CNN for deep face representation with noisy labels, IEEE Trans. Inf. Forensics Secur. 13 (11) (2018) 2884–2896, https://doi.org/10.1109/TIFS.2018.2833032.

[51] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, International Conference on Computer Vision Pattern Recognition (CVPR) 2016, pp. 770–778.

[52] M. Hayat, M. Bennamoun, S. An, Deep reconstruction models for image set classification, IEEE Trans. Pattern Anal. Mach. Intell. 37 (4) (2015) 713–727, https://doi.org/10.1109/TPAMI.2014.2353635.