Multimedia Tools and Applications (2019) 78:22743–22772
https://doi.org/10.1007/s11042-019-7618-0

Implementation of multimodal biometric recognition via multi-feature deep learning networks and feature fusion

Leslie Ching Ow Tiong[1] · Seong Tae Kim[1] · Yong Man Ro[1]

Received: 9 March 2018 / Revised: 14 January 2019 / Accepted: 8 April 2019 /
Published online: 24 April 2019

Abstract

Although there is an abundance of current research on facial recognition, it still faces significant challenges that are related to variations in factors such as aging, poses, occlusions, resolution, and appearances. In this paper, we propose a Multi-feature Deep Learning Network (MDLN) architecture that uses modalities from the facial and periocular regions, with the addition of texture descriptors to improve recognition performance. Specifically, MDLN is designed as a feature-level fusion approach that correlates between the multimodal biometrics data and texture descriptor, which creates a new feature representation. Therefore, the proposed MLDN model provides more information via the feature representation to achieve better performance, while overcoming the limitations that persist in existing unimodal deep learning approaches. The proposed model has been evaluated on several public datasets and through our experiments, we proved that our proposed MDLN has improved biometric recognition performances under challenging conditions, including variations in illumination, appearances, and pose misalignments.

Keywords Deep multimodal learning · Multimodal biometric recognition · Multi-feature fusion layers · Texture descriptor representations

1 Introduction

Over the past two decades, the multimedia and security communities have commonly implemented biometric systems to perform tasks such as facial recognition, social network identity tagging, physical access controls, and surveillance cameras. At present, the performances and capabilities of these biometric systems are still far from achieving their requirements due to the existing limitations of unimodal biometrics [23] that is commonly utilized by these systems.

✉ Yong Man Ro
ymro@kaist.ac.kr

[1] Image and Video System Lab, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

Besides, most biometric applications are still unimodal by means of them relying on a single feature to achieve recognition. To overcome these limitations, multimodal biometric recognition has become a subject of recent research [8, 16].

This paper proposes a Multi-feature Deep Learning Network (MDLN) architecture to perform a multi-feature learning approach by using multimodal biometrics with texture descriptors. Note that the primary input source for the facial recognition system is an image-based data [11]. This underlies our motivation to use different biometric modalities could be extracted to form the multimodal biometric model. Several studies have been developed to extract several modalities across different face regions such as the periocular [40] and facial area [5, 12]. Indeed, several studies have proven that the periocular area is one of the most discriminating facial features [38, 41, 60] to recognize. This motivates us to utilize the periocular as an additional feature in enhancing facial recognition performance. Periocular, as defined in this paper, refers to the region surrounding the eyes and the eyebrows. As compare to facial modality, periocular modality is less complex and standalone information [38, 41]. For instance, facial recognition can be affected by occlusions such as wearing a mask, and medical alteration such as plastic surgery and cosmetic products. The skin textural changes and shape variations to face could be changed. The periocular modality will be sufficiently stable and contains different periorbital region (region contained around the eye and orbit) information as compared to the face. This yields a more vibrant feature representation that may be used to enhance the performance of recognition as compared to using a single biometric modality.

In recent years, deep learning is often applied to very favorable outcomes in the fields of image processing and computer vision [30, 42]. There has also been a number of deep learning applications for facial recognition [24, 27, 47]. In the most of these works, the input to the network depends on the raw pixels' RGB data, so the model is implicitly learning the features inside the network layers, which can be a difficult task. In other words, most of the existing deep learning models only focus on filtering out the confounding factors such as illuminations and occlusions. As a result, recognizing an individual is still a challenging task due to certain intra-class variabilities such as aging, illuminations, occlusions, poses, and lack of distinctiveness.

This motivates us to design a Convolutional Neural Network (CNN) by using two-stream inputs, yielding better representations as compared to using only a single modality in the deep learning architecture. The input of our proposed model forms through two-streams of data, namely RGB data and texture descriptors. The purpose of using texture descriptors is to create discriminatory features by transforming the color-based images in a different manner to reduce any variabilities due to illumination changes. We then assemble the proposed model using the two-stream CNN network by integrating the RGB data and the handcrafted texture descriptors within the network to be robust enough to compensate for any hidden information during inference. These inputs explicitly describe the different hidden information in the network, which makes the recognition easier, as the network does not need to estimate biometric features implicitly. This prevents the architectures from relying on singular sources to extract insufficient features for facial recognition.

Besides, one of the significant challenges in multimodal learning models is the implementation of a fusion algorithm. This fusion algorithm can be applied to the data-level, the intermediate (feature-level), and the decision-level [48]. Data-level fusion focuses on the combined data from multiple sources by removing correlations between the modalities or by representing the fused data in a lower-dimensional subspace. Intermediate fusion offers a flexible strategy for multimodal data [5], where the multimodal data can be exploited as a useful feature representation, and the features can, therefore, be learned by fusing the modalities into a joint representation [63]. Decision-level fusion strategy refers to the decision from multiple classifiers that are trained on separate modalities; this fusion strategy is often used for classification errors that are caused by different classifiers due to uncorrelated features, which was popular in multimodal biometrics and machine-learning communities [62]. Several methods have been considered in the literature [22, 31], and [32], where it is reported that managing the fused feature representations highlights the importance of the fusion algorithm. However, most of the fusion layer in deep learning uses feature concatenation, which becomes very inefficient as the feature space dimensionality is increased [1, 15, 46, 58]. Also, this approach does not explore features across different levels of extraction and representation. This means that the effects of fusion at these levels are not investigated for deep learning based multimodal biometric systems.

We consider several research questions that relate to in this area – What are the advantages of using several different texture descriptors as opposed to using raw RGB data? However, can these descriptors improve texture analysis in CNN? Furthermore, how should we design the robust fusion layer using dual descriptors for extracting additional information that operates in two-stream CNN? To answer those questions, we have proposed our new texture descriptor layer and tested several existing descriptors that utilizes feature fusion layers in the preprocessing pipelines to support efficient features learning. Moreover, with the MDLN structure, it is essential to determine the training weights for the different features. Thus, we propose a few feature fusion layers and a score fusion layer to fuse all the weights to recognize individuals. Feature fusion layers are designed to join information from each of the dual-inputs in order to transfer knowledge representation in between the MDLN networks to perform facial recognition. The score fusion layer is employed to cast weighted scores to all labels; it is hence allocated the highest weight in the recognition.

Thus, the contributions of this paper can be summarized as follows:

- A large-scale analysis of texture descriptor has been conducted for facial recognition via a deep learning approach. Motivated by the challenges in facial recognition, we devise a new texture descriptor layer as the second stream in MDLN to extract additional or more explanatory factors. This paper discovers that texture descriptors could be served as a complementary role for recognition to provide an invariance of complex transformation inputs.
- A feature fusion strategy is developed by utilizing a multi-feature learning approach with dual-stream inputs. The idea is to regard the convolutional layers of MDLN as a filter bank and create a new feature representation using feature fusion layers, to compensate for hidden information during inference.
- A score fusion layer is applied to construct a joint feature similarity measurement vector across all the biometric modalities to formulate a decision. This layer enhances our recognition model by casting multiple measurement functions from two independent MDLNs (one for facial and the other for periocular), and then sampling the highest weighted measurement for recognition. Extensive experiments and comparisons show that the proposed model improves the recognition accuracy over unimodal deep learning approaches.

This paper is organized in the following way; Section 2 reviews the related work in texture descriptors and multimodal biometrics recognition. Section 3 provides the presentation of our proposed model using multimodal biometrics with MDLN architectures. Section 4 describes the presentation of experimental results and analysis. A conclusion of our proposed model is summarized in Section 5.

2 Background and related works

Several works related to the texture descriptors of human biometric traits has been studied in [43, 64], to help with the early stage of texture analysis in gender recognition [7], or emotional recognition [29]. This is because texture descriptors are ubiquitous enough to provide useful properties of objects, especially when the shape is not useful. However, not many studies can be found where a texture descriptor is analyzed in deep learning architecture from a pure biometrical perspective. As a result, this motivates us to investigate the impact of texture descriptors using deep learning.

Traditionally, multimodal data has widely been used by the multimedia and biometric communities [46] for facial recognition. In the deep learning fields, these research communities have investigated several fusion techniques using multimodal data, covering several fusion perspectives. The advantages of using deep learning are that it can automatically be learned from a given input, instead of manually designing a feature extraction and representation to feed a machine-learning algorithm. Recently, deep learning studies are converging to face modality for facial analysis [47], mobile-based authentication [2, 45], etc. These active research areas interest the academia and industry, and it has state-of-the-art performances [20, 27, 54]. This paper focuses on the intermediate fusion layer as this fusion method results in a better discriminative classifier as it integrates different feature extracts from different biometric modalities into a feature representation.

In a multimodal context, several deep learning research works adopt an intermediate fusion, by constructing a shared representation layer that merges these multimodal data, thereby forcing the network to learn its inputs [6, 52]. This shared representation layer can be either a single shared layer that fuses multiple channels at a certain depth or through all of the network's layers. However, many research works apply concatenation for the features or weights in the shared representation layer. For instance, early works conducted by [54] and [44] involves using images and text as the multimodal data. Their works demonstrated a simple concatenation of inputs and the utility of fusing representations of different modalities in a deep learning framework.

Moreover, Kahous et al. [25] introduced a hierarchical classifier for combing audio and video features, to which they fused the modality features using a multiple kernel learning method. Feichtenhofer et al. [14] and Karpathy et al. [26], conducted experiments whereby their networks learned several representations of video streams that fused across multiple fusion layers during training. Both experiments consistently produced better results for the video classification problem, as compared to utilizing a single modality or data-level fusion technique. Another work related to image correlation was presented by Liu et al. [31], where they proposed a fusion strategy for a deep learning architecture to discover a correlation between color and temporal images.

Although the fusion strategy is flexible, many recent works adopted a single fusion layer, which requires careful implementation regarding how and what can be fused in the networks.

None of the previous works discussed for constructing multi-feature representations based on the usage of different biometric characteristics in a deep learning application. Additionally, the selection process of whichever biometric modalities that are to be fused is usually based on experimental perceptions. In this paper, we devise a new multi-feature learning method using the combination of two inputs for each biometric modality, followed by the construction of a shared representation layer. This shared representation layer is employed across several feature fusion strategies to merge multiple modalities into a single feature representation vector.

3 Proposed framework

This section describes our proposed model, which performs multiple features learning by using multimodal biometrics with several descriptors. This model is separated into three parts: (1) feature extraction and its descriptors, (2) a multi-feature learning network with feature fusion layers, and (3) a score fusion layer.

3.1 Feature extraction and descriptors

In this paper, we focus on multimodal biometrics for the face and periocular with data that are obtained from the same source. First, let us denote the face as $f \in \mathbb{R}^{(m_1 \times n_1)}$ and the periocular as $p \in \mathbb{R}^{(m_2 \times n_2)}$ as the input features, where m_1 and n_1 are the height and width of the facial modality, and m_2 and n_2 are the height and width of the periocular modality, respectively. Both f and p are represented in the RGB color space.

Next, we are interested in using several feature descriptors to compare their recognition performance, as the initial performance of deep learning is dependent on the raw data fed to it. Thus, we hypothesize that using a combination of feature descriptors can extract one or more explanatory factors, aside from the RGB data. For texture descriptors, we focus on several methods that are well known in literature, namely Local Binary Pattern (LBP), Local Ternary Pattern (LTP), Entropy texture, Histogram of Oriented Gradients (HOG), alongside our proposed descriptor.

In face recognition, the shapes of face modality are important as it provides richer information to represent the features [35]. We deploy a preprocessing pipeline with feature transformations based on [10] and [50], resulting in textures extracted from f and p to develop a frequency domain procedure. The preprocessing pipeline can improve the appearance of an image by maximizing the intensity range and enhancing contrast to support efficient learning in our proposed model. In this process, we applied a face detection and an image cropping technique to extract the region of interest (ROI) of f and p based on [55].

The apparent changes in the images are related to the variations in lighting conditions. The reason behind using the method is that we need to separate the illumination components from the given I. To enhance the reflectance while reducing illumination for texture descriptors, the illumination component from the given f and p has to be separated first, before filtering the resulting image in the frequency domain. We initially represent f and p as an image in a two-dimension matrix $I(x, y)$. First, we transform I into Fourier domain as matrix Z:

$$Z(u, v) = FT[I(x, y)], \tag{1}$$

where FT denotes as Fourier transform function. Furthermore, by applying a low-pass filter function, namely Butterworth filter B, to Z in the frequency domain, we then obtain a filtered matrix H:

$$H(u,v) = Z(u,v)\cdot B(u,v). \tag{2}$$

We selected the Butterworth filter as it has a maximally flat magnitude in a generally linear phase response. In addition, the filter results in no gain ripples. The Butterworth filter function is written as [10]:

$$\begin{aligned} B(u,v) &= \frac{1}{1+\left[\frac{D_o}{D(u,v)}\right]^{2n}}, \\ D(u,v) &= \left[(u-h/2)^2 + (v-w/2)^2\right]^{1/2}, \end{aligned} \tag{3}$$

where n is defined as the order of filter in the Butterworth filter and D_0 is defined as the cutoff distance between $D(u, v)$ and the center of surrounding coordinates. $D(u, v)$ is a passband frequency between the order and cutoff frequencies of the band pass filter. The parameters for n and D_0 depends on the experimental perspective (for further reading, readers can refer to [10, 13]). In our experiments, we define $n = 1$ and $D_0 = 0.5$.

Now, we can obtain a filtered version of $I'(x, y)$ through the filter $B(u, v)$ in the frequency domain. To receive the desired filtered image $I'(x, y)$, taking an inverse of Fourier transform of (1):

$$\begin{aligned} Z'(x,y) &= FT^{-1}[H(u,v)], \\ I'(x,y) &= e^{Z'(x,y)}, \end{aligned} \tag{4}$$

where e denotes an exponential operation. Finally, we can apply the filtered images $I'(x, y)$ along with several texture descriptors to represent f and p as texture descriptors, respectively. Figure 1 summarizes and illustrates the image preprocessing pipeline.

As mentioned earlier, we study several texture descriptors of the face and periocular regions. Each descriptor's performance is individually evaluated to determine the best configuration for texture representations. The detailed analysis is performed in Sect. 4.2.1. From the perspective of several texture descriptors' studies, texture can be served as a complimentary role to provide an invariance of complex transformation inputs. Accordingly, each of the descriptors is described in the detail next subsection.

3.1.1 Entropy

Entropy texture is a statistical measurement of information theory, which is used to characterize the uncertainty factors across the information content of an input image [4]. This paper applies an Entropy function based on [4] with the given I', in order to maximize the mutual

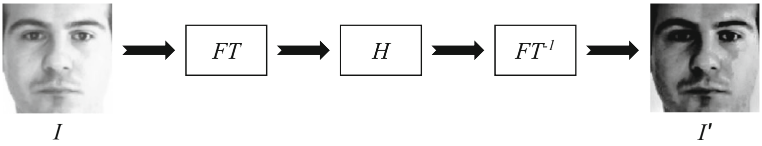


Fig. 1 Illustration of image preprocessing pipeline using Butterworth filter

information across the various images used in our experiments, including variations of pose alignment and illumination. The Entropy texture F_E is computed based on a single pixel with the given x and y coordinates of I'. Then, the entropy function will define a neighborhood around the pixel of interest, and then calculate the statistic for the neighborhood to determine the pixel value to assign the value as the output pixel. Note that, by default, the neighborhood around the pixel of interest is defined as a 3 × 3 neighborhood around the pixel of interest. The equation is written as follows:

$$F_E(x,y) = -\sum_{i=y-1}^{y+1}\sum_{j=x-1}^{x+1} P(i,j)\log\{P(i,j)\} \tag{5}$$

where P is defined as the histogram distribution of the filtered image I', x and y denote as the coordinates of F_E with the 3 × 3 neighborhood around the pixel of interest, and log function is defined as a logarithmic function.

3.1.2 Histogram of oriented gradients

HOG was introduced by [9] to represent the gradient orientations in a regular area of an image. The local appearance and shape of a given image can be characterized by its distribution of intensity gradients and edge directions. This descriptor removes much of the non-essential information to highlight the outlines. Given an I', it is divided into several cells to compute a magnitude of gradients g, which is then calculated for the histogram of gradient directions θ. In this paper, we follow the implementation of [9] to construct a cell size of 5 × 5 with a 9 bin histogram, and a block that is configured by grouping the 2 × 2 cells in I'.

3.1.3 Local binary pattern

LBP is introduced by [39] for texture analysis. This descriptor works by thresholding a 3 × 3 neighborhood using the gray level of the central pixel in the binary code. This code is multiplied by the threshold values, using the power of two and adding the results in a clockwise manner. In our experiments to incorporate I' into LBP texture, we apply an LBP function from [39] to form labels for each image pixel by thresholding a 3 × 3 neighborhood matrix. This process avoids the complete loss of the features' hidden information.

3.1.4 Local ternary pattern

LTP is an extension of the primary LBP function with three-value codes [56]. This descriptor is more discriminant and less sensitive to noise texture. Also, it reduces the sensitivity of noise in uniform areas, to present the facial image with smooth illumination gradients. To do so, we follow the implementation from [56] to construct the LTP descriptors. Typically, this function generates two features by considering the codes in two halves, positive and negative codes. In our experiments, the constant threshold τ is used to reduce the influence of noise, which is set as $\tau = 7$.

3.1.5 Orthogonal combination of LBP and LTP

The images are complex due to the unconstrained environments. The existing texture descriptors still have limitation and not well-performed. Thus, we devise a new texture descriptor F_{oc}

by using an orthogonal combination of LBP and LTP. The main difference is that F_{oc} computes an orthogonal combination of binary codes from LBP and LTP to reduce the sensitivity of image noise and enhance illumination transformation by maximizing the texture areas (see Fig. 2). The process of orthogonal combination drastically retains the invariance with respect to illumination transformation by combining the different four-orthogonal neighbor operators (see Fig. 2). Therefore, it can reduce the sensitivity of illumination transformation by characterizing the edges and texture of the image with a better result.

Given an I', this algorithm computes the LBP and LTP features inside the binary code, which is illustrated in Fig. 2 and described in Algorithm 1. The LBP and LTP operators are considered whereby the LBP consists of the horizontal and vertical neighbors, and the LTP consists of the positive and negative matrices each with horizontal and vertical neighbors. By concatenating these two features, we compute the orthogonal combination between the LTP positive matrix with LBP (t_1 and t_2 as illustrated in Fig. 2), along with the LTP negative matrix with the LBP (t_3 and t_4 as illustrated in Fig. 2).

This orthogonal combination is designed by splitting the LBP and LTP into four non-overlapping orthogonal groups, where each LTP binary code is computed separately for each orthogonal group while concatenating the LBP codes which is used as the image description. For instance, given a pixel, this approach will generate a LBP matrix and LTP positive matrix as shown in Fig. 3. Then, we selected the values from red dotted line boxes in LTP positive (see Fig. 3) and the values from green dotted line boxes in LBP (see Fig. 3), and concatenated all of them into t_1 as the orthogonal combination. The similar process is done for t_2, t_3, and t_4 as shown in Fig. 2. Those combinations are performed to retain the invariance against illumination transformation by removing any outlying disturbances. The final descriptor F_{oc} is then constructed by maximizing all of the histograms from the cells, where we adopt a uniform strategy for pixel weighting, which is computed by combining the detected feature in each batch (t):

$$OC(i,j) = \max[t_1(i,j), t_2(i,j), t_3(i,j), t_4(i,j)], \tag{6}$$

where i and j are defined as the coordinates of the t matrix, and t represents each subset of the orthogonal combination of OC. Image intensity is between 0 and 255, where zero represents black and 255 represents white, thus using the max[.] function can easily remove the noise.

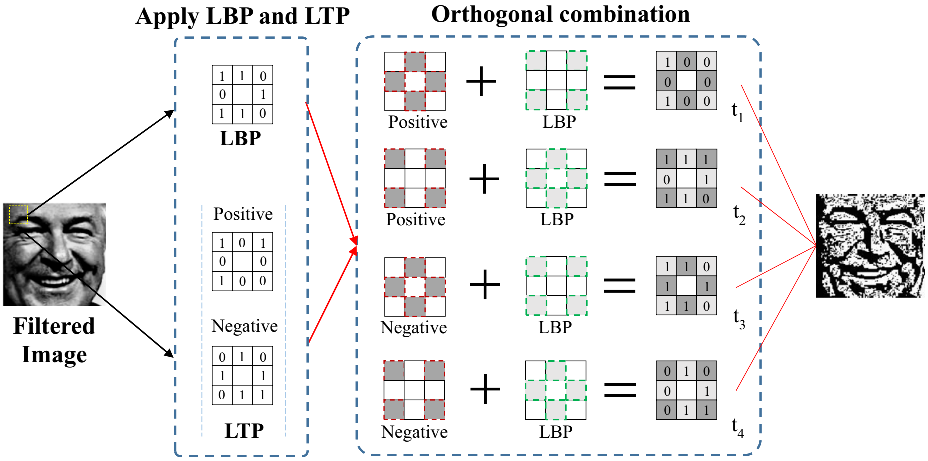

Apply LBP and LTP
Orthogonal combination
Filtered Image
LBP
1 1 0
0 1
1 1 0
Positive
1 0 1
0 0
1 0 0
Negative
0 1 0
1 1
0 1 1
LTP
Positive + LBP = 1 0 0 / 0 0 / 1 0 0 t1
Positive + LBP = 1 1 1 / 0 1 / 1 1 0 t2
Negative + LBP = 1 1 0 / 1 1 / 1 1 0 t3
Negative + LBP = 0 1 0 / 0 1 / 0 1 1 t4

Fig. 2 Illustration of the proposed descriptor by an orthogonal combination of LBP and LTP

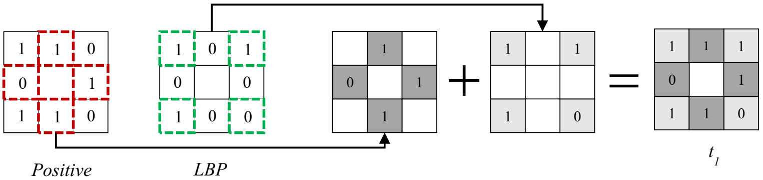


Fig. 3 Demonstration of t_1 orthogonal combination using LBP and LTP (positive)

To obtain texture that encoded the gradient image information by describing the appearance of a subject, we utilize continuous non-linear sigmoid function [19] as follows:

$$F_{oc}(i,j) = \frac{1}{1+\exp^{c[th-OC(i,j)]}}, \tag{7}$$

where c is defined as the slope of the sigmoid function and th is the threshold value.

Algorithm 1 Proposed descriptor by an orthogonal combination of LBP and LTP

```
Input: Low-pass filtered image $I'(x,y)$ from the Butterworth filter according to (4).
Output: Proposed Texture Descriptor $F_{oc}$
// Initialization
Initialize the size of $F_{oc}$
// Apply LBP and LTP to $F_{oc}$
for $i$=1 to $m$ do
    for $j$=1 to $n$ do
        compute $t_1$ and $t_2$ using LTP positive matrix with LBP function with horizontal and vertical matrices
        compute $t_3$ and $t_4$ using LTP negative matrix with LBP function with horizontal and vertical matrices
        $OC(i,j)$ ← calculate $OC(i,j)$ using (6) by given $t_1$, $t_2$, $t_3$, and $t_4$
        $F_{oc}(i,j)$ ← apply (7) by given $OC(i,j)$
    end for
end for
```

3.2 Multi-feature learning network architecture with feature fusion layer

This section describes our proposed model dubbed the MDLN that is inspired by [52]. An MDLN is built upon a two-stream shared convolutional (*conv*) layer architecture, where one layer is applied to process the RGB stream and the other is used to handle the texture stream. The parameters of this shared *conv* layers are simulated across the networks by sharing the same parameterization to form a feature map. More explicitly, we intend to correspond both networks to learn the modalities at the same spatial location using several different texture descriptors. Thus, the networks can then concurrently discriminate itself from these different descriptors.

Figure 4 illustrates the network structure of our proposed model using two independent MDLNs for each biometric modality. Each of the MDLN's parameter initialization consists of two *conv* networks, which accept different inputs but are joined by the shared layers. This shared network consists of 18 *conv* layers and 6 max pooling (*maxpool*) layers. During training, we selected four input descriptors (two for *f*, and another two for *p*). However, the correlation between every two sets of descriptors may not follow a consistent pattern, and thus the understanding of the correlation between these feature vectors can be difficult. Therefore, utilizing shared weights in MDLN helps our proposed model to converge better. In other words, the networks can learn the correspondence between these relevant features to discriminate between themselves.

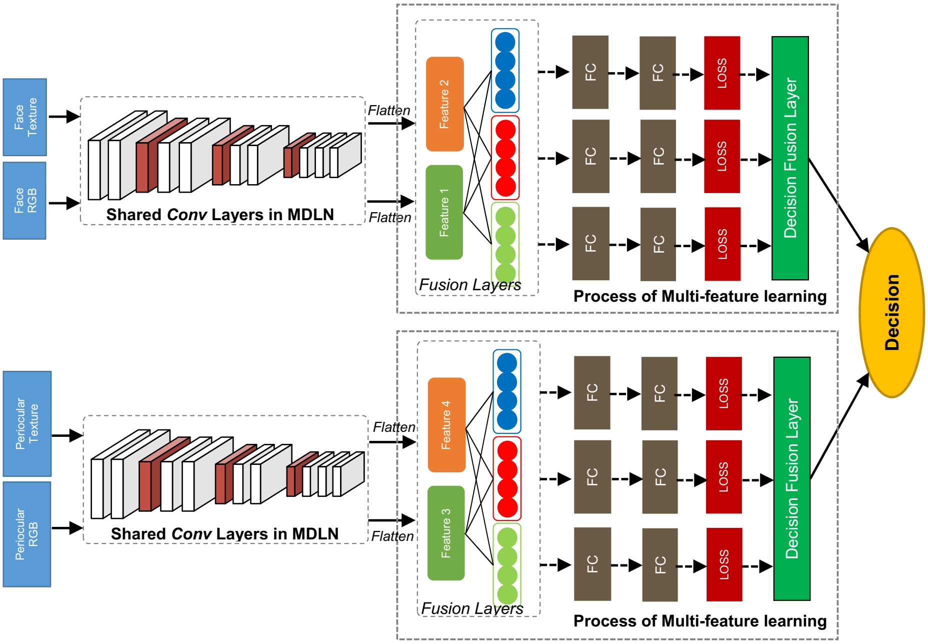


Fig. 4 Illustration of our proposed model with multimodal biometrics and score fusion layers. The proposed model contains two independent MDLN networks for each biometric modality. MDLN focuses on learning several features with shared layers

To capture the different hidden information properties from the RGB data and texture descriptors, we adopt several feature fusion strategies for the MDLN. The main goal for the fusion layer is to train the MDLN such that the ultimate joint feature representation outperforms single modality representations. The advantage of our feature fusion layers is that our feature fusion method extracts the network's fully-connected (*fc*) layers and then fuses its *fc* vector across different streams as feature representations to train the network. In this paper, we intend to fuse the *conv* layers. First, we extract features from the last *conv* layer of the network, where we then compute a flattened operation in the last *conv* layer based on the network's architecture type. Furthermore, we perform several fusion strategies to strengthen feature activations and fine-tune the rest of the *fc* layers as shown in Fig. 4. The advantage of using those fusion strategies is that they can exploit the activations' correspondence across the different feature representations. By using these strategies, the model can therefore fuse the feature maps accordingly, thereby achieving remarkable improvements.

To make the ideas of MDLN more concrete, we will now discuss several ways of fusing feature layers in a completely fused network. A fusion function fuses two features $\overrightarrow{A} = [a_1, a_2, \cdots, a_i]$ and $\overrightarrow{B} = [b_1, b_2, \cdots, b_i]$ to produce an output Y_{fusion}. In this study, we apply several fusion layers to feedforward MDLN architecture, consisting of *fc* layers that Y_{fusion} can be applied at different points in the network to implement intermediate fusion. We consider several fusion layers, which are discussed each in detail next.

3.2.1 Average fusion layer

The *avg* layer Y_{avg} computes the sum of two one-dimensional feature vectors (a 1 × N matrix) in the *fc* layers, then compute the average on the feature vectors to employ an arbitrary correspondence of 1 × N matrix to its best effect by optimizing over the filters of each network (refers to Fig. 5):

$$Y_{avg} \leftarrow \left[\left(\overrightarrow{A} + \overrightarrow{B} \right) / 2 \right] \cdot \mathrm{W}^{\mathrm{T}} + \mathrm{b}, \tag{8}$$

where W^{T} is defined as a weighted matrix and b denotes the bias matrix, $\overrightarrow{A}$ and $\overrightarrow{B}$ are defined as the given feature vectors of the flattened layers.

3.2.2 Concatenation fusion layer

The *concat* layer Y_{concat} stacks two one-dimensional feature vectors (refers to Fig. 5), which provides:

$$Y_{concat} \leftarrow \left[\overrightarrow{A} \| \overrightarrow{B} \right] \cdot \mathrm{W}^{\mathrm{T}} + \mathrm{b}, \tag{9}$$

where || is defined as a concatenation function. The definition of other variables is the same as those that were previously defined. This concatenation function defines a correspondence, by combining the arbitrary information without any optimization. This means that after the feature vectors are stacked, the subsequent layers can then learn these correspondences to discriminate between these features.

3.2.3 Max fusion layer

The *max* layer $Y_{\max}$ takes the maximum of two feature vectors, to employ the arbitrary correspondence (see Fig. 5):

$$Y_{\max} \leftarrow \overrightarrow{C} \cdot \mathrm{W}^{\mathrm{T}} + \mathrm{b}, where\ \overrightarrow{C} = [c_1, c_2, \cdots, c_i], c_i = \max(a_i, b_i). \tag{10}$$

$\overrightarrow{C}$ denotes as a 1 × N matrix, which is defined as the operation of finding the absolute highest values in the given inputs, W^{T} is defined as a weighted matrix, and b denotes a bias matrix. Like the *avg* layer, the correspondence between the networks is arbitrary.

We devise this model for both *f* and *p* to learn through the MDLNs using multiple descriptors, which is as illustrated in Fig. 5. The *avg*, *concat*, and *max* fusion layers at the

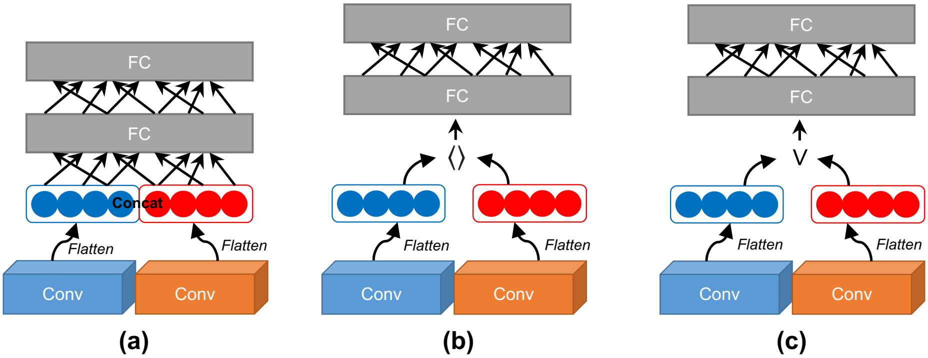


Fig. 5 Illustration of each feature fusion layer: **a** *concat* layer, **b** *avg* layer, and **c** *max* layer

last flattened convolutional layer directly enhance the network with less complexity. In addition, the fusion layer combines the outputs of the internal fully-connected layers, which provides an additive property between the different feature activations. Each fusion layer performs a loss function as classification labels. In the experimental section (see Sec. 4.2.1), we evaluate and compare the performance of each of these possible fusion methods with respect to their classification accuracy.

3.3 Score fusion layer

Since the goal of our proposed multimodal biometrics recognition approach is to determine whether the inputs of biometric modalities belong to the subject of the biometric modalities inputs, score fusion layer is designed to measure the similarity between the presented multiple biometrics and the storage feature vectors. Then, we combine the measurements and generate a joint feature similarity measurement vector that can be used by the identification module for decision-making.

Our score fusion layer is built upon the weighted fusion of the final embedded layers of the MDLN. To benefit from the feature fusion layers, the score fusion layer combines each similarity measurement originating from different feature layers in order to generate a joint feature similarity measurement to weight the scores before a recognition decision is made. For training, a softmax function is utilized to calculate the distances. Let us denote these measurements as U to calculate the joint feature similarity measurement vector based on the multi-feature fusion vectors for the identification module:

$$\begin{aligned} U(X) &= \mathrm{argmax}\left[\mathrm{softmax}\left(X_{avg}\right) + \mathrm{softmax}\left(X_{concat}\right) + \mathrm{softmax}\left(X_{\max}\right)\right], \\ \mathrm{softmax}(x) &= \frac{\exp(x)}{\sum_{j=1}^{\theta} \exp\left(x_j\right)}, \end{aligned} \tag{11}$$

where $X \in \{X_{avg}, X_{concat}, X_{max}\}$ is defined as the test feature representations matrix, X_{avg} is defined as the feature representation vector based on the *avg* fusion layer using Eq. 8, X_{concat} is defined as the feature representation vector based on the *concat* fusion layer using Eq. 9, and X_{max} is defined as the feature representation vector based on the *max* fusion layer using Eq. 10. softmax(·) is defined as the probability distribution over X different for each possible target class over all possible target classes $\theta = \{1, 2, 3, \cdots, N)$, exp(·) is defined as an exponential function, and arg max function is used to identify the class that X belongs to. N is defined as the size of classes θ. For testing, we applied a cosine similarity measurement $\varDelta(\cdot)$ to each network to compile the distance score vector $dist[\cdot]$:

$$\begin{aligned} dist\left[x, X_{\theta}^{'}\right] &= \left[\varDelta\left(x, x_1^{'}\right), \varDelta\left(x, x_2^{'}\right), \cdots, \varDelta\left(x, x_N^{'}\right)\right], \\ \varDelta\left(x, x^{'}\right) &= 1 - \cos\left(x, x^{'}\right) \\ &= 1 - \frac{x \cdot x^{'}}{\|x\| \cdot \|x^{'}\|}, \end{aligned} \tag{12}$$

where x denotes the test feature representations matrix, $X_{\theta}^{'}$ is defined as the feature matrix of a particular θ class, $\theta = \{1, 2, 3, \cdots, N)$ and N is defined as the number of classes. To combine the measurements into a single scalar:

$$\varPsi(X) = dist\left[X_{avg}, X_{\theta}^{'}\right] + dist\left[X_{concat}, X_{\theta}^{'}\right] + dist\left[X_{max}, X_{\theta}^{'}\right], \tag{13}$$

where Ψ denotes the sum of the joint feature similarity measurement vector based on the multi-feature fusion layers, and $X \in \{X_{avg}, X_{concat}, X_{max}\}$ denotes the test feature representations which are extracted by using feature fusion layers. Each feature representation vector presents different distances; we therefore have to sum up all the measurements to one for measurement comparisons. Then, Ψ compares the distance between each inter-class and assists in the selection of robust features towards the decision-making process. Note that, this paper is working on multimodal biometrics recognition for the identification module, the calculation of Ψ is computed with all the classes to identify the class that XX belongs to. The recognition δ is hence:

$$\begin{aligned} \omega(XX) &= \Psi\left(XX_f\right) + \Psi\left(XX_p\right) \\ \delta(XX) &= \operatorname{argmax}[\omega(XX)] \end{aligned} \tag{14}$$

where XX is defined as multimodal biometric features, XX_f denotes facial features and XX_p denotes periocular features. $\omega(\cdot)$ is defined as the sum of Ψ for both face and periocular modalities. $\delta(\cdot)$ denotes the argmax function for returning the class which has maximum score from $\omega(\cdot)$. Algorithm 2 summarizes the entire recognition process of our proposed model by using a score fusion layer to enhance the performance in facial recognition.

Algorithm 2 Recognition process using proposed score fusion layer

```
Input: multimodal biometric features as $XX = \{XX_f, XX_p\}$ computed through MDLN network.
Output: Recognition label result
// Apply score fusion layers after multi-feature fusion layers
$\Psi(XX_f)$ ← apply (13)
$\Psi(XX_p)$ ← apply (13)
$\omega(XX)$ ← calculate the weights based on the given $\Psi(XX_f)$ and $\Psi(XX_p)$
Recognition: $\delta(XX)$ ← apply (14)
```

4 Experiments

We conducted several experiments to evaluate and compare the performance of the proposed method with other existing methods for the tasks of facial recognition. These experiments are conducted under constrained and unconstrained environments. Five existing RGB public datasets and our dataset are selected, namely the AR dataset [33], EURECOM KinectFaceDB dataset [34], PubFig dataset [28], FaceScrub dataset [37], and YouTube Face dataset [59]. All the datasets are designed for the task of facial recognition either with constrained or with unconstrained environments. The detailed descriptions of experimental settings and the performance of our proposed method, along with other methods are presented in Sect. 4.1 and 4.2, respectively.

4.1 Experimental setups

In these experiments, we tested several deep learning networks, namely ConvNet that uses unimodal biometrics, MDLN that uses unimodal biometrics with dual descriptors, and our proposed network (see Fig. 4). ConvNet consists of eight *conv* layers, three *maxpool* layers, and two fully-connected layers. For ConvNet's parameter initialization, we followed the guidelines in [53] to design it. For instance, the *Conv1*, which consisted of 64 feature maps, was defined with size 128 × 128, which were denoted as 64@128 × 128. The structure of MDLN will not be described in this section as it was covered in Sect. 3.2 and 3.3. Note that, both ConvNet and MDLN were trained using different unimodal biometrics independently. Table 1 describes the details of ConvNet and MDLN for using face modalities. Each parameter configuration of MDLN is described in Table 1. For the case of periocular modality, Table 2 describes the details of ConvNet and MDLN for using periocular modality.

The training process was performed using an open-source deep learning toolkit TensorFlow [57] with relative modifications. We implemented an annealed learning rate, which started from 1.0×10^{-3} and subsequently divided by 10 for every 10 epochs (the minimum rate was defined as 1.0×10^{-5}). The batch size was set to 32 and the training was carried out for 1000 epochs. For optimizations, we used the Adam Optimizer with the weight decay and momentum that were set to 0.0005 and 0.9, respectively.

Furthermore, we also applied a number of popular benchmark models in facial recognition, namely FaceNet [49], LCNN-29 [61], and VGG Face [42]. All the models are provided by the authors. The experimental protocols of the benchmark models are designed the same as the proposed model, in order to evaluate the unbiased protocols of recognition.

4.2 Experimental results

4.2.1 Performance analysis of texture descriptors and fusion strategies

This section evaluated both unimodal and multimodal biometrics by using a single texture descriptor, a combination of texture descriptors, and different fusion strategies via the proposed model. To verify the multimodal effect, we built a new database that contains the face and

Table 1 The configuration of ConvNet and MDLN models using face modality *f*

ConvNet model configuration			MDLN model configuration			
Layer	Filter	Output	Layer		Filter	Output
Conv1	5 × 5	64@128 × 128	*Conv1 (Shared)*		5 × 5	64@128 × 128
Conv2	5 × 5	64@128 × 128	*Conv2 (Shared)*		5 × 5	64@128 × 128
Max Pool	2 × 2 (Stride)	128@64 × 64	*Max Pool*		2 × 2 (Stride)	128@64 × 64
Conv3	3 × 3	128@64 × 64	*Conv3 (Shared)*		3 × 3	128@64 × 64
Conv4	3 × 3	128@64 × 64	*Conv4 (Shared)*		3 × 3	128@64 × 64
Max Pool	2 × 2 (Stride)	256@32 × 32	*Max Pool*		2 × 2 (Stride)	256@32 × 32
Conv5	3 × 3	256@32 × 32	*Conv5 (Shared)*		3 × 3	256@32 × 32
Conv6	3 × 3	256@32 × 32	*Conv6 (Shared)*		3 × 3	256@32 × 32
Max Pool	2 × 2 (Stride)	512@16 × 16	*Max Pool*		2 × 2 (Stride)	512@16 × 16
Conv7	3 × 3	512@16 × 16	*Conv7 (Shared)*		3 × 3	512@16 × 16
Conv8	3 × 3	512@16 × 16	*Conv8 (Shared)*		3 × 3	512@16 × 16
Flatten	–	1 × 1 × 512 × 3	*Conv9 (Shared)*		3 × 3	512@16 × 16
FC 1	–	1 × 1 × 4096	*Flatten*	*F1*	–	1 × 1 × 512 × 3
FC 2	–	1 × 1 × 4096		*F2*	–	1 × 1 × 512 × 3
			Fusion Layer	Y_{concat}	–	1 × 1 × 8192
				Y_{avg}	–	1 × 1 × 4096
				Y_{max}	–	1 × 1 × 4096
			FC Layer	Y_{concat}	–	1 × 1 × 8192
				Y_{avg}	–	1 × 1 × 4096
				Y_{max}	–	1 × 1 × 4096

Table 2 The configuration of ConvNet and MDLN models using periocular modality *p*

ConvNet model configuration			MDLN model configuration			
Layer	Filter	Output	Layer		Filter	Output
Conv1	5 × 5	64@50 × 150	*Conv1 (Shared)*		5 × 5	64@50 × 150
Conv2	5 × 5	64@50 × 150	*Conv2 (Shared)*		5 × 5	64@50 × 150
Max Pool	2 × 2 (Stride)	128@25 × 75	*Max Pool*		2 × 2 (Stride)	128@25 × 75
Conv3	3 × 3	128@25 × 75	*Conv3 (Shared)*		3 × 3	128@25 × 75
Conv4	3 × 3	128@25 × 75	*Conv4 (Shared)*		3 × 3	128@25 × 75
Max Pool	2 × 2 (Stride)	256@12 × 37	*Max Pool*		2 × 2 (Stride)	256@12 × 37
Conv5	3 × 3	256@12 × 37	*Conv5 (Shared)*		3 × 3	256@12 × 37
Conv6	3 × 3	256@12 × 37	*Conv6 (Shared)*		3 × 3	256@12 × 37
Max Pool	2 × 2 (Stride)	512@6 × 18	*Max Pool*		2 × 2 (Stride)	512@6 × 18
Conv7	3 × 3	512@6 × 18	*Conv7 (Shared)*		3 × 3	512@6 × 18
Conv8	3 × 3	512@6 × 18	*Conv8 (Shared)*		3 × 3	512@6 × 18
Flatten	–	1 × 1 × 512 × 3	*Conv9 (Shared)*		3 × 3	512@6 × 18
FC 1	–	1 × 1 × 4096	*Flatten*	*F1*	–	1 × 1 × 512 × 3
FC 2	–	1 × 1 × 4096		*F2*	–	1 × 1 × 512 × 3
			Fusion Layer	Y_{concat}	–	1 × 1 × 8192
				Y_{avg}	–	1 × 1 × 4096
				Y_{max}	–	1 × 1 × 4096
			FC Layer	Y_{concat}	–	1 × 1 × 8192
				Y_{avg}	–	1 × 1 × 4096
				Y_{max}	–	1 × 1 × 4096

periocular regions. Our database is a real-life dataset that contains 806 subjects with unconstrained environments, which is designed based on five ethnic groups: *African*, *White*, *Asian*, *Latin American*, and *Middle Eastern* (see Table 3 for the brief ethnic groups). The subjects are celebrities, politicians, sportsmen, etc., and are randomly selected based on BBC News [3], Naver News [36], and IMDB [21]. To build this dataset, we downloaded the images from the Google image search engine by using a Python script. The sample images are shown in Fig. 6 with large variations, such as poses, illuminations, appearances, and occlusions.

These experiments randomly selected 577 subjects from the VGG face dataset and 477 subjects randomly from our dataset. Thus, 1054 subjects that contained 166,468 images were used for training; 24,301 images were used as testing in phase 1, and 23,370 images were used for testing in phase 2. In these experiments, we defined the size of *f* as 128 × 128 and the size of *p* as 50 × 150.

Performance analysis of texture descriptor with and without image processing In Table 4, we analyze the performance of the MDLN model using with and without image preprocessing

Table 3 Descriptions of five ethnic group in our dataset

Ethnic group	Descriptions
African	The subjects are from Europe, America, and Africa regions.
White	The subjects are from America, Europe, and Oceania.
Asian	The subjects are from East Asia, South-east Asia, and South-west Asia regions.
Latin American	The subjects are from Middle and South America regions.
Middle Eastern	The subjects are from Northern Africa, Southern Europe, and West Asia.

We categorize the ethnic group based on the major five regions

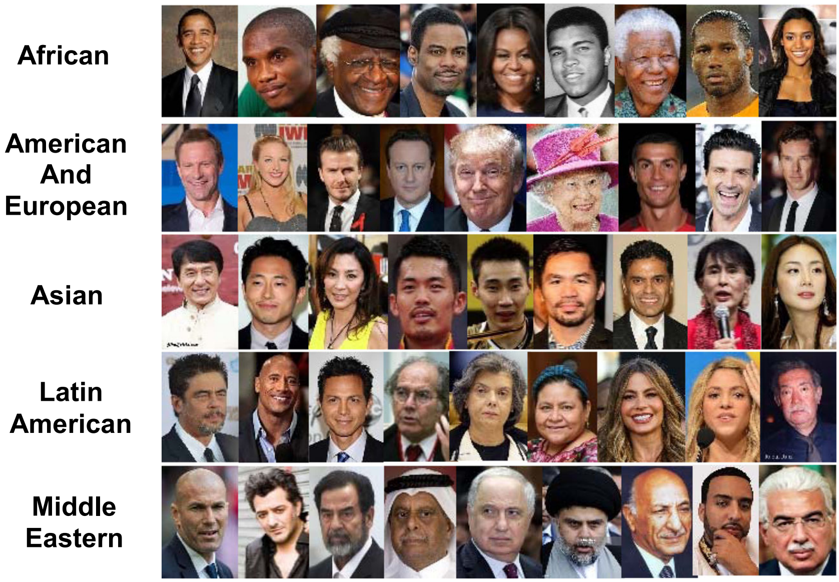


Fig. 6 Sample images of five major ethnic groups from our dataset. Each row shows several subjects from different ethnic groups

for the feature descriptors on the facial, periocular regions, and mulitmodal. All the images are used from Testing phase 2. As can be seen in Table 4, the experimental results that did not use image preprocessing step, which is not performed well for the recognition performance. The proposed model that used image processing step (see Sec. 3.1) achieved at least 2% improvement for recognition accuracy. This is because the texture descriptor (without image preprocessing step) contains too much noise information and it does not provide richer information to represent the biometrics modalities clearly to train the proposed model. Thus, we concluded that the image preprocessing step is very important to represent the texture descriptor. In the following experiments, the descriptors were created using the filtered image.

Performance analysis of fusion strategies We present a qualitative study to analyze the performance of the MDLN model using different fusion strategies with the combination of feature descriptors on the facial and periocular regions in Table 5. All the images are used from Testing phase 1.

For all fusion strategies shown in Table 5, we observed that the capability of the MDLN model utilized feature fusion layers to learn the correlation between texture descriptors in the Type 3 strategy effectively. This comparison indicates that the Type 3 fusion strategy outperformed the other fusion strategies by achieving accuracies 87.83% (RGB + F_E),

Table 4 Performance analysis results on the proposed model using RGB + F_{oc} texture descriptor with and without image preprocessing

Methods	Face	Periocular	Multimodal
Proposed model (with preprocessing)	87.99%	83.48%	90.04%
Proposed model (without preprocessing)	85.14%	80.32%	87.93%

The experiments are conducted with a face, periocular and multimodal from Testing phase 2

Table 5 Analysis results on MDLN by using different fusion strategies and descriptors

Type	Fusion	MDLN-Face					MDLN-Periocular				
		$RGB + F_E$	$RGB + F_{HOG}$	$RGB + F_{LBP}$	F_{LTP}	$RGB + F_{oc}$	$RGB + F_E$	$RGB + F_{HOG}$	RGB + F_{LBP}	F_{LTP}	$RGB + F_{oc}$
1	Y_{concat}	87.15%	87.02%	87.13%	81.58%	87.80%	83.08%	84.27%	83.37%	74.62%	82.17%
	Y_{avg}	86.30%	85.98%	86.71%	80.33%	87.07%	82.80%	84.05%	82.44%	73.67%	81.73%
	Y_{max}	86.52%	86.30%	86.19%	80.98%	87.38%	82.78%	84.18%	82.51%	74.35%	81.64%
2	$Y_{concat} + Y_{avg}$	87.64%	87.33%	87.70%	81.89%	87.90%	83.59%	84.44%	83.41%	74.97%	82.44%
	$Y_{concat} + Y_{max}$	87.55%	87.12%	87.48%	81.33%	87.87%	83.65%	84.38%	83.49%	74.78%	82.44%
	$Y_{max} + Y_{avg}$	86.70%	86.88%	87.37%	81.42%	87.57%	82.89%	84.21%	82.82%	74.50%	81.97%
3	**All**	**87.83%**	**87.62%**	**87.73%**	**82.95%**	**87.99%**	**83.67%**	**84.73%**	**83.53%**	**75.22%**	**82.48%**
4	Score fusion	84.34%	84.28%	84.10%	79.60%	84.41%	80.12%	78.53%	80.10%	77.70%	80.47%

The experiments are conducted with face and periocular from Testing phase 1. Type 1 uses individual feature fusion layers, Type 2 applies two feature fusion strategies, and Type 3 combines three layers. Type 4 uses a score fusion strategy. The highest accuracy is written in bold

Table 6 Analysis results on ConvNet, MDLN and proposed model using different feature descriptors

Method Features	ConvNet-Face	ConvNet-Periocular	MDLN-Face	MDLN-Periocular	Proposed
RGB	86.23%	81.67%	N/A[a]	N/A	N/A
F_E	73.19%	66.12%	N/A	N/A	N/A
F_{HOG}	72.22%	66.53%	N/A	N/A	N/A
F_{LBP}	70.54%	65.90%	N/A	N/A	N/A
F_{oc}	73.33%	66.01%	N/A	N/A	N/A
F_{LTP}	N/A	N/A	85.98%	80.32%	86.98%
RGB + F_E	N/A	N/A	87.83%	83.67%	89.23%
RGB + F_{HOG}	N/A	N/A	87.62%	84.73%	89.32%
RGB + F_{LBP}	N/A	N/A	87.73%	83.53%	89.13%
RGB + F_{oc}	N/A	N/A	87.99%	83.48%	89.42%
Mix 1[b] (Proposed)	N/A	N/A	N/A	N/A	**90.04%**
Mix 2[c]	N/A	N/A	N/A	N/A	89.79%
Mix 3[d]	N/A	N/A	N/A	N/A	89.91%
Mix 4[e]	N/A	N/A	N/A	N/A	89.26%
Mix 5[f]	N/A	N/A	N/A	N/A	89.71%

The experiments are conducted with a face, periocular and multimodal from Testing phase 2. ConvNet and MDLN used unimodal biometrics. The highest accuracy is written in bold

[a] N/A = Not applicable

[b] *Mix 1* = using RGB + F_{HOG} for periocular and RGB + F_{oc} for face

[c] *Mix 2* = using RGB + F_E for periocular and RGB + F_{oc} for face

[d] *Mix 3* = using RGB + F_{HOG} for periocular and RGB+ F_E for face

[e] *Mix 4* = using RGB + F_{LBP} for periocular and RGB + F_E for face

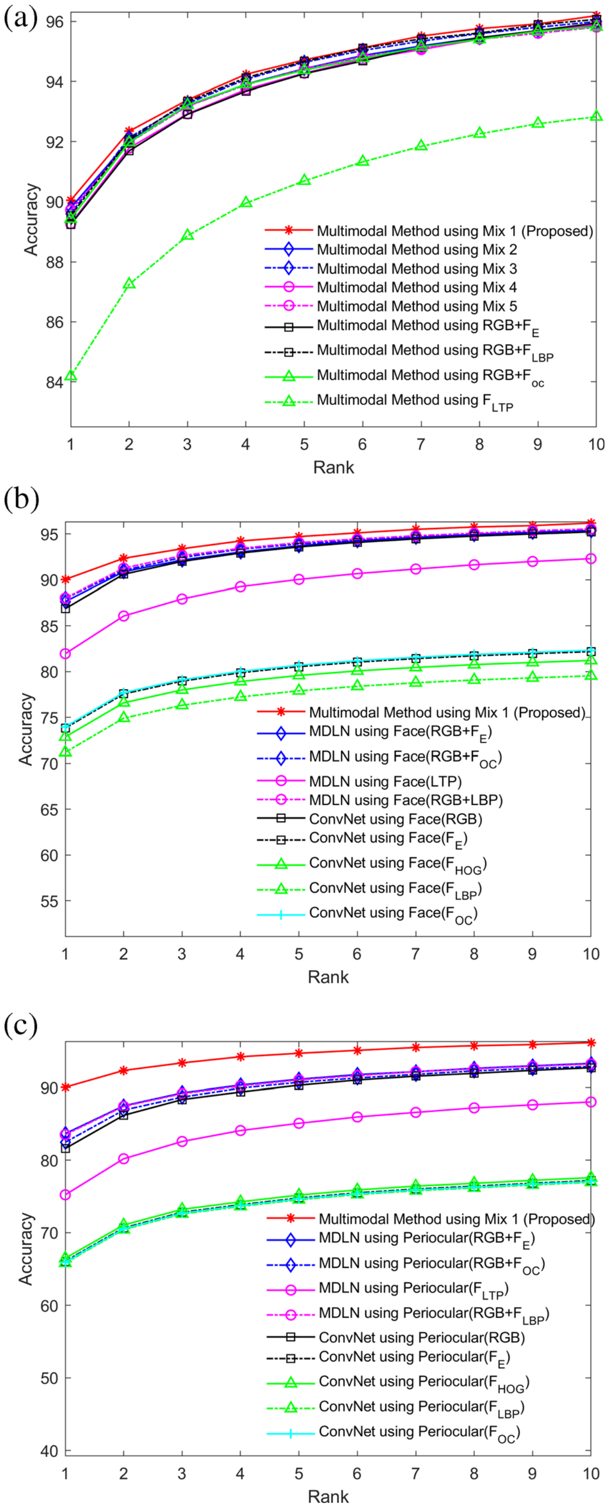
[f] *Mix 5* = using RGB + F_{LBP} for periocular and RGB + F_{oc} for face

87.62% (RGB + F_{HOG}), 87.73% (RGB + F_{LBP}), 82.95% (F_{LTP}), and 87.99% (RGB + F_{oc}) for using face, and also achieving 83.67% (RGB + F_E), 84.73% (RGB + F_{HOG}), 83.53% (RGB + F_{LBP}), 75.22% (F_{LTP}) and 82.48% (RGB + F_{oc}) for periocular regions. However, using a score fusion layer is not performing well as compared to feature fusion layers. This is due to the correlation between the *fc* layers are not interrelated in the proposed model, and only merging the loss-function measurement for recognition. Through the experimental results, we also demonstrated that utilizing all feature fusion layers:Y_{avg}, Y_{concat}, and $Y_{max}$with Type 3 fusion strategy, which provided better results than alternatives. This is because the features are already highly informative while collapsing and maximizing all the information from different fusion layers.

Performance analysis of texture descriptors From our previous experimental analysis, using Type 3 fusion strategy, MDLN has led to better performance in the recognition task. Following that, we further conducted another experiment to analyze the combination of different texture descriptors by using our pre-trained model from the previous section and the images are used from Testing phase 2. As mentioned in Sect. 1, the illumination and pose can affect the appearance of an individual, and in this case, using one single texture descriptor will fail to extract the details of

▶

Fig. 7 CMC curves for the effects of our proposed model and other approaches using different biometric traits; **a** our proposed model with different texture descriptors; **b** our proposed model vs. other approaches using face with different texture descriptors; **c** our proposed model vs. other approaches using periocular region with different texture descriptors


(a)
Accuracy
96
94
92
90
88
86
84
1
2
3
4
5
6
7
8
9
10
Rank
Multimodal Method using Mix 1 (Proposed)
Multimodal Method using Mix 2
Multimodal Method using Mix 3
Multimodal Method using Mix 4
Multimodal Method using Mix 5
Multimodal Method using RGB+F_E
Multimodal Method using RGB+F_{LBP}
Multimodal Method using RGB+F_{oc}
Multimodal Method using F_{LTP}
(b)
Accuracy
95
90
85
80
75
70
65
60
55
1
2
3
4
5
6
7
8
9
10
Rank
Multimodal Method using Mix 1 (Proposed)
MDLN using Face(RGB+F_E)
MDLN using Face(RGB+F_{OC})
MDLN using Face(LTP)
MDLN using Face(RGB+LBP)
ConvNet using Face(RGB)
ConvNet using Face(F_E)
ConvNet using Face(F_{HOG})
ConvNet using Face(F_{LBP})
ConvNet using Face(F_{OC})
(c)
Accuracy
90
80
70
60
50
40
1
2
3
4
5
6
7
8
9
10
Rank
Multimodal Method using Mix 1 (Proposed)
MDLN using Periocular(RGB+F_E)
MDLN using Periocular(RGB+F_{OC})
MDLN using Periocular(F_{LTP})
MDLN using Periocular(RGB+F_{LBP})
ConvNet using Periocular(RGB)
ConvNet using Periocular(F_E)
ConvNet using Periocular(F_{HOG})
ConvNet using Periocular(F_{LBP})
ConvNet using Periocular(F_{OC})

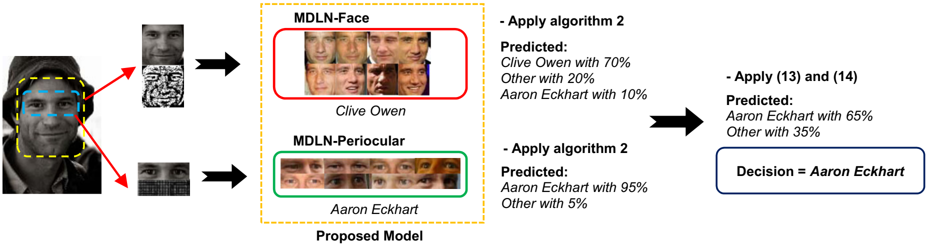


Fig. 8 Demonstration of our proposed model to recognize an individual by using multimodal biometrics and the proposed *Mix 1* texture descriptors

factors in learning performance. This is because singular descriptors do not precisely represent the observed high-dimensional data. Hence, this experiment investigates our proposed model that utilizes multi-feature learning approaches with the feature fusion layers' effect on the performance, as compared to the unimodal biometric deep learning approaches.

Table 6 reported that our proposed model has outperformed the other models. We observed that the proposed model with multimodal biometrics (using *Mix 1* descriptors) boosted performance significantly by 4% to 5%. This is evidently confirmed by Cumulative Match Characteristic (CMC) curves in Fig. 7. By observing the performance of our proposed model with *Mix 1* descriptor, Fig. 7 visualizes the performance of unimodal biometrics with all kinds of feature descriptors. The extracted features for unimodal biometrics seemed to provide lower complementary accuracy.

Through our analysis, we discovered that using RGB + F_{oc} results in the best achievement in representing the face compared to using other feature descriptors. This is because F_{oc} employs complementary information from image data by extracting desirable distortions and artifacts from facial with less noise, thus, collapsing its spatial information to prevent the adverse effect on illumination and appearance. In addition, our study also discovered that using HOG on the periocular provides more complementary information than using other descriptors to identify an individual. This is because HOG can avoid the complete loss of feature spatial information, to compose the periocular representation from a novel modality. Therefore, this is reasoned by the fact that, each MDLN model is trained to focus on different properties of the multi-feature learning approaches from multimodal biometrics in the fusion layers.

To detail our proposed model's achievement, Fig. 8 demonstrates an example of our evaluation to recognize an individual, namely Aaron Eckhart. We utilized the RGB + F_{oc} for f and the RGB + F_{HOG} for p. The individual was recognized as Clive Owen based on face and was correctly recognized as himself using the periocular. Evidently, the appearances of the

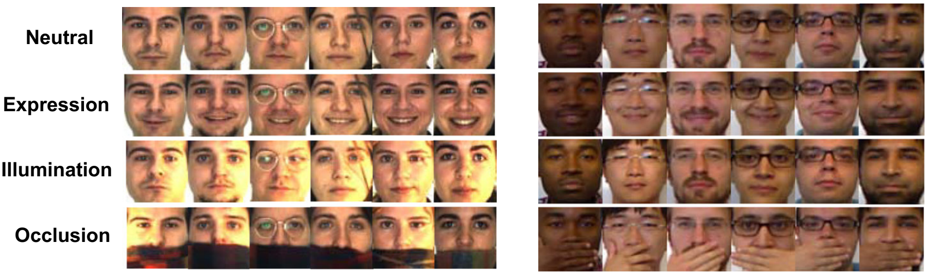


Fig. 9 Sample images from AR dataset (left) and EURECOM KinectFaceDB dataset (right) under several controlled environment conditions

examples were affected by illuminations and cosmetic products. This confirmed the validity of the feature fusion layers to provide reliable performance due to its uniqueness whereby it contained hidden texture information. In other words, using the fusion layers with texture descriptors convinced our proposed model to alter its decision. In summary, our proposed model addressed that the combination of different texture descriptors, which can, therefore, be used to overcome the challenging scenarios by enhancing recognition performance.

4.2.2 Performance comparison between proposed model and other benchmark models under constrained environment datasets

To better understand the feature performance, we reported the experimental results for benchmark evaluation on our proposed model under constrained environment datasets, such as AR dataset and KinectFaceDB. These datasets are popular choices for multimodal biometric recognition, which provides a good baseline for proof-of-concept. The constrained environment is defined as all images that are collected based on highly controlled environments, such as using artificial light for recording and the fixed distance between the subjects and the camera. Besides, both datasets have another characteristic – presence of a scarf or 'hand on face' occlusions.

Evaluation on AR dataset The AR face dataset consists of faces with varying illumination, expression, and occlusion conditions that were captured in two sessions. The illumination conditions and the distance from the subject to the camera were strictly controlled during the capturing sessions (see Fig. 9). The detail of this dataset structure and acquisition environment are carefully described in [33]. For comparison, we conducted two experiments in this dataset by following the protocol from [51] in our evaluation:

- First, we evaluated our proposed model on a set of 100 subjects with different illumination and expression conditions. Thus, 7 images for each subject from Session 1 were used as gallery images, and another 7 images for each subject from Session 2 were used as probe images. For testing the algorithms, the experimental setup was similar to the previous section, we utilized [55] to crop out the ROI of the face and periocular regions.
- Furthermore, we also studied the robustness of recognition performance in scarf occlusion condition. Likewise, the first experiment, 7 non-occluded images per subject from Session

Table 7 Comparison results of Rank-1 recognition on AR dataset with other methods under Expression (Exp), illumination (Illum) and Scarf occlusion conditions

Methods	Exp + Illum (%)	Scarf (%)
SMBR-WE [51] reported	96.90	86.20
SMBR-E [51] reported	96.00	80.00
FaceNet [49]	94.00	79.77
LCNN-29 [61]	90.00	77.58
VGG Face [42]	93.50	78.17
Multimodal method using Mix 1 (Proposed)	**97.71**	**93.57**
Multimodal method using RGB + F_{oc} for f and p	97.43	93.00
Multimodal method using RGB + F_E for f and p	97.43	93.17

The highest accuracy is written in bold

Table 8 Comparison results on Rank-1 recognition on EURECOM KinectFaceDB dataset with other methods under Expression, illumination and "Hand on Face" occlusions

Methods	Neutral (%)	Illum (%)	Smile (%)	Hand on Face (%)	Overall (%)
LGBP using RGB + Depth [34] reported	100.00	100.00	100.00	98.08	99.52
LBP using RGB + Depth [34] reported	100.00	99.04	100.00	96.16	98.80
PCA using RGB + Depth [34] reported	82.69	95.19	89.42	68.67	84.00
FaceNet [49]	100.00	99.04	99.04	98.08	99.04
LCNN-29 [61]	100.00	98.08	100.00	98.08	99.04
VGG Face [42]	100.00	98.08	99.04	80.77	94.47
Multimodal method using Mix 1 (Proposed)	**100.00**	**100.00**	**100.00**	**99.04**	**99.76**
Multimodal method using RGB + F_{oc} for f and p	100.00	99.04	100.00	96.16	98.80
Multimodal method using RGB + F_E for f and p	100.00	98.08	100.00	98.08	99.04

The highest accuracy is written in bold

1 were used as gallery images, 12 scarf occlusion images per subject from both sessions were used as probe images.

Table 7 shows the performance of rank-1 recognition with different multimodal biometrics approaches on the AR dataset. The performance of different biometric modalities achieved by [51], and benchmark models such as FaceNet, LCNN-29, and VGG Face are presented. It is shown that using our proposed model, the recognition performance achieved 97.71% accuracy for the first experiment and achieved 93.57% accuracy for the scarf occlusion. The proposed model clearly achieved about 1% improvement over other algorithms for the first experiment and achieved about 7% improvement for the second experiment. Thus, robust recognition using multimodal modalities is achieved with a significant improvement over the other benchmarks.

Furthermore, a comparison with FaceNet, LCNN-29, and d VGG Face shows that the approaches performed poorly as compared to our proposed method. This is related to the robustness of texture descriptors with the feature fusion layers, which created correlated feature representations in different complementary information. In addition, periocular modality assisted the performance of recognition in the occlusion and illumination scenarios. Hence, the fusion strategy boosted the performance of recognition.

Evaluation on KinectFaceDB dataset We conducted another experiment on EURECOM KinectFaceDB dataset in the benchmark evaluation by following the protocol from [34].

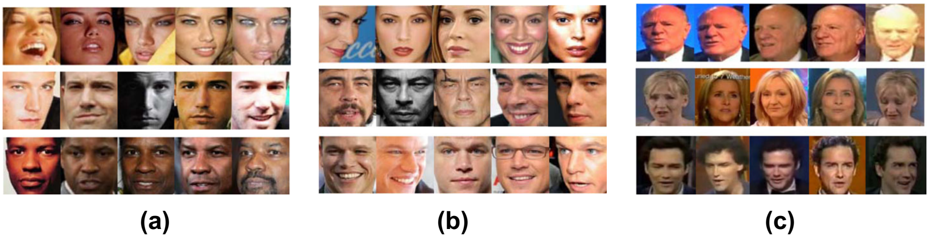


Fig. 10 Sample images from **a** PubFig, **b** FaceScrub, and **c** YTF datasets. The images from all the datasets contained uncontrolled environmental conditions

The EURECOM KinectFaceDB dataset is a multimodal face database, which is collected based on the Kinect sensor (see Fig. 9). There are 52 subjects were captured in two sessions and consists of RGB-based and depth-based images with varying illumination, expression ('Neutral' and 'Smile'), and 'Hand on Face' occlusion condition, as described in [34]. In this experiment, we do not apply depth-based images. First, we followed the protocol from [34], by using the neutral faces from Session 1 as gallery images. Then, the other images such as 'Illum', 'Smile', and 'Hand on Face' (including the neutral images from Session 2) for each subject from both were used as probe images. For performance tests and comparison with the benchmark approaches, we also implemented FaceNet, LCNN-29, and VGG Face in this experiment.

In our evaluation, the rank-1 identification rate was reported in Table 8 for all tested variations in both sessions. As can be observed in this table, our proposed method achieved 100% accuracy for the cases of 'Neutral', 'Illum', and 'Smile'. [34] reported that the fusion of RGB and depth images, extended to use multiple modalities, which achieved the same results for the cases of 'Neutral', 'Illum', and 'Smile'. In the case of 'Hand on Face' occlusion condition, our proposed model achieved 99.04% accuracy that outperformed other methods (see Table 8). The overall result in Table 8 shows that our proposed model achieved the highest accuracy with 99.76%. This implies that the straightforward way of *Mix 1* descriptors improves the performance of our proposed method, by combining the various handcrafted texture with *conv* layers and fusion layers. Besides, compared to other approaches, *Mix 1* descriptors are not profoundly affected by noise such as low resolution, and therefore, it yields higher performance.

4.2.3 Performance comparison between proposed model and other benchmark models under unconstrained environment datasets

Our interest is in large-scale and realistic face identification scenarios where the diversity is naturally captured. To evaluate the holistic performance of the proposed model against the current benchmark models on a large scale, three public face datasets, namely PubFig, FaceScrub, YouTube Face (YTF), and our dataset are selected to verify the generalization ability of our proposed model. All the datasets are real-life datasets with unconstrained environments. In our experiments, those datasets are selected as they can be used to extract periocular modality for multimodal biometrics' configuration. The sample images of a few subjects are shown in Fig. 10. In these experiments, all the images of the facial and periocular regions were extracted using [53]. The database descriptions and their evaluation results are presented next. In addition, FaceNet, LCNN-29, and VGG Face are implemented as benchmark methods to evaluate the proposed model with the same protocol.

Evaluation on our dataset We conducted another experiment by using the remaining 329 subjects from our dataset as testing phrase. All the subjects were not used in the training process. We a new protocol by randomly selecting five images as gallery images for each subject and the remaining images were used as probe images. The experimental results have justified the advantage of the proposed model in Table 9, shows that our proposed model achieved 87.41% accuracy, which is competitive when compared to benchmark approaches by achieving 2% to 5% improvements in recognition performance. Moreover, Fig. 11 illustrated several examples of our evaluation to recognize individuals. We proved that these two

Table 9 Comparison results on Rank-1 recognition, using proposed model with different descriptors and benchmark approaches

Methods	Accuracy (%)
FaceNet [49]	83.26
LCNN-29 [61]	79.58
VGG Face [42]	80.48
Multimodal using *Mix 1* (proposed)	**87.41**
Multimodal using *Mix 3*	86.39
Multimodal using RGB + F_{oc} for f and p	85.81
Multimodal using RGB + F_{oc} for f and p	85.47
Multimodal using RGB + F_{HOG} for f and p	85.39

The highest accuracy is written in bold

combinations of descriptors have complementary effects in the recognition task. The recognition accuracy is also higher than other methods in Table 9, even though we also used several combinations of texture descriptors, which implies that combining *Mix 1* descriptors in our proposed method improves the accuracy.

Evaluation on PubFig dataset PubFig [28] consists of 200 subjects with real-life images that collected from the internet. The images have been acquired in unconstrained situations without any user cooperation (see Fig. 10). As the experimental setting of [18], we split the dataset into three groups, with the first group as gallery images and the remaining two groups as probe images. Each experiment was tested five times and the average results are reported.

The performance of our proposed model is compared to the performance of benchmark models, which are reported in Table 10. At the same time, it shows that the proposed model achieved the best performance than all the competing methods in this experiment. The proposed model achieved 98.92 ± 0.40% accuracy. Furthermore, our proposed model using F_{oc} for f and p shows that they also achieve better performance with 98.21 ± 0.61% accuracy compared to the other approaches (excluding [18]). By appropriately weighing different biometric modalities, our proposed model using *Mix 1* descriptors achieved better results. Furthermore, the robustness recognition of our proposed model, using multiple texture descriptors with fusion layers results in a significant improvement over the current benchmark,

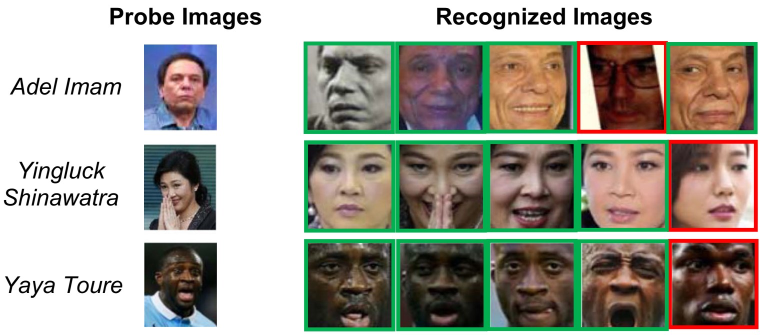


Fig. 11 Sample results of Rank-5 recognition for three people in our dataset: Adel Imam (top), Yingluck Shinawatra (middle), and Yaya Toure (bottom). Predicted inliers are marked with green rectangles, outliers with red rectangles. The recognition is done by using our proposed model with *Mix 1* descriptors

Table 10 Comparison results on Rank-1 recognition on PubFig, FaceScrub, and YTF datasets with other methods

Methods	PubFig (%)	FaceScrub (%)	YTF (%)
[18] reported	98.60 ± 0.30	91.50 ± 0.5	N/A
DRM-PWV [17] reported	89.90 ± 0.86	N/A	51.45 ± 3.06
FaceNet [49]	98.37 ± 0.31	90.10 ± 1.50	48.76 ± 2.16
LCNN-29 [61]	97.55 ± 1.89	89.77 ± 2.12	45.28 ± 2.47
VGG Face [42]	95.03 ± 1.57	87.78 ± 3.18	43.24 ± 2.23
Multimodal method using *Mix 1* (Proposed)	**98.92 ± 0.51**	**92.38 ± 1.21**	**58.11 ± 1.33**
Multimodal method using RGB + F_{oc} for f and p	98.21 ± 0.61	91.30 ± 1.44	57.36 ± 1.61
Multimodal method using RGB + F_E for f and p	97.88 ± 1.06	91.11 ± 1.66	55.29 ± 1.82

The highest accuracy is written in bold

which is competitive when compared to others by achieving 0.3% to 6% improvements in recognition performance.

Evaluation on FaceScrub dataset The FaceScrub dataset consists of 530 subjects whereby the images are assigned based unconstrained environments [37]. The images contain a different appearance, poses, illuminations, expressions and time. The details of the dataset were described in [37]. For evaluation, we followed the protocols from [17] and [18], by dividing the images into three groups, with the first group as gallery images and the remaining two groups as probe images. Experiments were done five different times for a different random selection of images into each evaluation. Following the evaluation configurations, we also performed experiments and compare our method with the benchmark methods. A detailed analysis, extensive performance evaluations and comparisons are presented next.

The experimental results are presented in Table 10, which shows that the proposed model achieves an average rank-1 recognition accuracy 92.38 ± 1.21%, which has outperformed all the benchmark models. Besides, the proposed method with RGB + F_{oc} achieved 91.30 ± 1.44% accuracy. FaceNet and LCNN-29 achieved 90.10 ± 1.50% accuracy and 89.77% ± 2.12% accuracy, respectively. By considering the deterministic texture descriptors and feature fusion layers, we proved that utilizing our proposed approach are able to extract the complementary information from the face and periocular modalities to improve the recognition via deep learning architecture.

Evaluation on YouTube face dataset The YFT dataset is a real-life video dataset that contains 1595 subjects that were downloaded from YouTube. The videos of the dataset have been acquired in a wide range of appearance variations and low-resolution images. Several sample images of a sample are shown in Fig. 10. Since the dataset itself and its protocol was originally developed for face verification. To evaluate our method and the benchmark methods for the task of face identification, we followed the protocols reported by [17], by selecting a subset of the dataset with four or more videos per person. Considering each video as different sessions, we randomly choose three videos for gallery sets and the remaining videos are used for probe sets. Experiments are repeated at least three times for different random selections of the training and testing videos.

This experiment evaluated the robustness of the proposed method under the low-resolution condition and a wide range of appearance variations. The evaluation results present in Table 10 shows that the proposed method achieves an average rank-1 recognition with $58.11 \pm 1.13\%$ accuracy and outperforms all the benchmark models. The capability of the feature fusion layers has extracted desirable distortions and artifacts from texture descriptors to prevent the adverse effect on appearance. Besides, the proposed method with $\text{RGB} + F_{oc}$ is the second-best performance, which achieved $57.36 \pm 1.61\%$ accuracy. FaceNet, LCNN-29, and VGG Face only achieved $48.76 \pm 2.16\%$, $45.58 \pm 2.47\%$, and $43.24 \pm 2.23\%$, respectively.

4.2.4 Discussion

We study the performance of feature fusion layers extracted from *fc* layers of a two-stream CNN in the proposed model. Results on various datasets, obtained as in Sect. 4.2, showed that the performance using feature fusion layers with texture descriptors beyond significantly outperform the other methods. Through the experimental results, several observations are addressed in this analysis. First, the experimental results showed that deep learning models with unimodal biometrics are still unreliable due to inconsistencies in environmental conditions such as illumination, human behavior, and variation in appearances. Most of the existing deep learning models only focus on filtering out the confounding factors such as illuminations and occlusions.

By considering the deterministic feature descriptors, maximizing the mutual information from multimodal biometrics such that all the questions that were mentioned before, this has been answered by our experimental results. We answer these questions in the affirmative in a series of experiments on the several challenging datasets. We demonstrate the proposed model that used a multi-feature learning approach can boost recognition accuracy. This implies a clear explanation that the combination of texture descriptors captures a wide variety of visual properties of textures as complementary information to overcome its drawbacks such as appearances and illuminations.

5 Conclusion

Facial recognition is a challenging task due to the frequent appearance changes caused by pose, expression, and illumination variations. Motivated by the needs of feature representations to achieve better performance of recognition, we have devised a MDLN architecture that employs complementary information from multimodal biometrics. Also, we have proved that unimodal biometrics has its limitations when applied to biometric recognition systems due to the existence of unreliable features that could be caused by a single source in deep learning architecture.

From a fusion perspective, we proved that the feature fusion layer in deep multimodal learning could simplify a flexible intermediate-fusion approach, which allows learning a joint feature representation efficiently at various depths in our proposed model. Our results suggest the importance of learning correspondences in MDLN that highly extract the features; one intriguing finding is that there is such an improvement by combining MDLN with dual-stream inputs. Comparative experimentations have proved that our proposed model is robust and significantly improves the overall recognition accuracy. These results have highlighted the superiority of our proposed model due to its robustness to solve real-time scenarios for identifying individuals.

References

1. Ahmad MI, Woo WL, Dlay S (2016) Non-stationary feature fusion of face and palmprint multimodal biometrics. Neurocomputing 177:49–61. https://doi.org/10.1016/j.neucom.2015.11.003
2. Ahuja K, Islam R, Barbhuiya FA, Dey K (2017) Convolutional neural networks for ocular smartphone-based biometrics. Pattern Recogn Lett 91:17–26. https://doi.org/10.1016/j.patrec.2017.04.002
3. BBC News. In: BBC. http://www.bbc.com/news
4. Bharati MH, Liu JJ, MacGregor JF (2004) Image texture analysis: methods and comparisons. Chemom Intell Lab Syst 72:57–71. https://doi.org/10.1016/j.chemolab.2004.02.005
5. Cao Z, Yin Q, Tang X, Sun J (2010) Face recognition with learning-based descriptor. In: Int Conf Comput Vis Pattern Recognit (CVPR). IEEE, San Francisco, CA, USA, p 2707–2714
6. Cao Y, Steffey S, Jianbiao H, Xiao D, Tao C, Chen P, Müller H (2015) Medical image retrieval: a multimodal approach. Cancer Informat 13:125–136. https://doi.org/10.4137/CIN.S14053
7. Castrillón-Santana M, Lorenzo-Navarro J, Ramón-Balmaseda E (2016) On using periocular biometric for gender classification in the wild. Pattern Recogn Lett 82:181–189. https://doi.org/10.1016/j.patrec.2015.09.014
8. Chen Y, Yang J, Wang C, Liu N (2016) Multimodal biometrics recognition based on local fusion visual features and variational Bayesian extreme learning machine. Expert Syst Appl 64:93–103. https://doi.org/10.1016/j.eswa.2016.07.009
9. Dalal N, Triggs W (2005) Histograms of oriented gradients for human detection. In: Int Conf Comput Vis Pattern Recognit (CVPR). IEEE, San Diego, CA, USA, p 886–893
10. Delac K, Grgic M, Kos T (2006) Sub-image homomorphic filtering technique for improving facial identification under difficult illumination conditions. In: Int Conf Syst, Signals Image Process. Budapest, Hungary, p 95–98
11. Devasena CL, Revathí R, Hemalatha M (2011) Video surveillance systems - a survey. Int J Comput Sci 8: 635–642
12. Elhamifar E, Vidal R (2011) Robust classification using structured sparse representation. In: Int Conf Comput Vis Pattern Recognit (CVPR). IEEE, Colorado Springs, CO, USA, p 1873–1879
13. Fan CN, Zhang FY (2011) Homomorphic filtering based illumination normalization method for face recognition. Pattern Recogn Lett 32:1468–1479. https://doi.org/10.1016/j.patrec.2011.03.023
14. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: Int Conf Comput Vis Pattern Recognit (CVPR). IEEE, Las Vegas, Nevada, USA, p 1933–1941
15. Goswami G, Mittal P, Majumdar A, Vatsa M, Singh R (2016) Group sparse representation based classification for multi-feature multimodal biometrics. Inf Fusion 32:3–12. https://doi.org/10.1016/j.inffus.2015.06.007
16. Goswami G, Singh R, Vatsa M, Majumdar A (2017) Kernel group sparse representation based classifier for multimodal biometrics. In: Int Joint Conf Neural Networks. IEEE, Anchorage, AK, USA, p 2894–2901
17. Hayat M, Bennamoun M, An S (2015) Deep reconstruction models for image set classification. IEEE Trans Pattern Anal Mach Intell 37:713–727. https://doi.org/10.1109/TPAMI.2014.2353635
18. Hayat M, Khan SH, Bennamoun M (2017) Empowering simple binary classifiers for image set based face recognition. Int J Comput Vis 123:479–498. https://doi.org/10.1007/s11263-017-1000-3
19. He Q, Zhang C, Liu DC (2015) Nonlinear image enhancement by self-adaptive sigmoid function. Int J Signal Process Image Process Pattern Recognit 8:319–328. https://doi.org/10.14257/ijsip.2015.8.11.29
20. Hu G, Yang Y, Yi D, Kittler J, Christmas W, Li SZ, Hospedales T (2015) When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition. In: Int Conf Comput Vis Workshop (ICCVW). IEEE, Santiago, Chile, p 142–150
21. Internet Movie Database. In: IMDB. http://www.imdb.com
22. Jagadiswary D, Saraswady D (2016) Biometric authentication using fused multimodal biometric. Procedia Comput Sci 85:109–116. https://doi.org/10.1016/j.procs.2016.05.187
23. Jain AK, Nandakumar K, Ross A (2016) 50 years of biometric research: accomplishments, challenges, and opportunities. Pattern Recogn Lett 79:80–105. https://doi.org/10.1016/j.patrec.2015.12.013
24. Kafai M, An L, Bhanu B (2014) Reference face graph for face recognition. IEEE Trans Inf Forensics Secur 9:2132–2143. https://doi.org/10.1109/TIFS.2014.2359548
25. Kahou SE, Bouthillier X, Lamblin P, Al E (2016) EmoNets: multimodal deep learning approaches for emotion recognition in video. J Multimodal User Interfaces 10:99–111. https://doi.org/10.1007/s12193-015-0195-2
26. Karpathy A, Joulin A, Fei-Fei L (2014) Deep fragment embeddings for bidirectional image sentence mapping. In: Int Conf Neural Info Process Syst. ACM, Montreal, Canada, p 1889–1897

27. Kasar MM, Bhattacharyya D, Kim T-H (2016) Face recognition using neural network: a review. Int J Secur Appl 10:81–100. https://doi.org/10.14257/ijsia.2016.10.3.08
28. Kumar N, Berg AC, Belhumeur PN, Nayar SK (2009) Attribute and simile classifiers for face verification. In: Int Conf Comput Vis (ICCV). IEEE, Kyoto, Japan, p 365–372
29. Levi G, Hassner T (2015) Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In: Int Conf Multimodal Interaction. ACM, Seattle, Washington, USA, p 503–510
30. Li H, Lin Z, Shen X, Brandt J, Hua G (2015) A convolutional neural network approach for face detection. In: Int Conf Comput Vis Pattern Recognit (CVPR). IEEE, Boston, MA, USA, p 5325–5334
31. Liu Y, Guo Y, Georgiou T, Lew MS (2018) Fusion that matters: convolutional fusion networks for visual recognition. Multimed Tools Appl. https://doi.org/10.1007/s11042-018-5691-4
32. Lumini A, Nanni L (2017) Overview of the combination of biometric matchers. Inf Fusion 33:71–85. https://doi.org/10.1016/j.inffus.2016.05.003
33. Martinez A, Benavente R (1998) The AR face database, Barcelona
34. Min R, Kose N, Dugelay J-L (2014) KinectFaceDB: a Kinect face database for face recognition. IEEE Trans Syst Man, Cybern Syst 44:1534–1548. https://doi.org/10.1109/TSMC.2014.2331215
35. Mokhayeri F, Granger E, Bilodeau G (2015) Synthetic face generation under various operational conditions in video surveillance. In: Int Conf Image Process (ICIP). IEEE, Quebec City, QC, Canada, p 4052–4056
36. Naver News. In: Naver. http://news.naver.com/
37. Ng HW, Winkler S (2014) A data-driven approach to cleaning large face datasets. In: Int Conf Image Process (ICIP). IEEE, p 343–347
38. Nigam I, Vatsa M, Singh R (2015) Ocular biometrics: a survey of modalities and fusion approaches. Inf Fusion 26:1–35. https://doi.org/10.1016/j.inffus.2015.03.005
39. Ojala T, Pietikäinen M, Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24:971–987. https://doi.org/10.1109/TPAMI.2002.1017623
40. Padole CN, Proenca H (2012) Periocular recognition: Analysis of performance degradation factors. In: IAPR Int Conf Biometrics (ICB). IEEE, New Delhi, India, p 439–445
41. Park U, Jillela RR, Ross A, Jain AK (2009) Periocular biometrics in the visible spectrum: A feasibility study. In: Int Conf Biometrics: Theory, Appl, Syst (BTAS). IEEE, p 1–6
42. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: British Machine Vision Conf. p 1–12
43. Pietikäinen M, Hadid A, Zhao G, Ahonen T (2011) Local binary patterns for still images. In: Computer vision using local binary patterns. Springer, Berlin, pp 1689–1699
44. Poria S, Cambria E, Gelbukh A (2015) Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proc Conf Empirical Methods in Natural Language Process. Lisbon, Portugal, p 2539–2544
45. Raghavendra R, Busch C (2016) Learning deeply coupled autoencoders for smartphone based robust periocular verification. In: Int Conf Image Process (ICIP). IEEE, Phoenix, Arizona, USA, p 325–329
46. Ramachandram D, Taylor GW (2017) Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process Mag 34:96–108
47. Ranjan R, Sankaranarayanan S, Castillo CD, Chellappa R (2017) An all-in-one convolutional neural network for face analysis. In: Int Conf Automatic Face and Gesture Recognit. IEEE, Washington, DC, USA, p 17–24
48. Ross A, Jain AK (2004) Multimodal biometrics: an overview. In: European Signal Process Conf. Vienna, Austria, p 1221–1224
49. Schroff F, Kalenichenko D, Philbin J (2015) FaceNet: A unified embedding for face recognition and clustering. In: Int Conf Comput Vis Pattern Recognit (CVPR). IEEE, Boston, MA, USA, p 815–823
50. Shahamat H, Pouyan A (2014) Face recognition under large illumination variations using homomorphic filtering in spatial domain. J Vis Commun Image Represent 25:970–977
51. Shekhar S, Patel VM, Nasrabadi NM, Chellappa R (2014) Joint sparse representation for robust multimodal biometrics recognition. IEEE Trans Pattern Anal Mach Intell 36:113–126. https://doi.org/10.1109/TPAMI.2013.109
52. Simonyan K, Zisserman A (2014) Two-Stream convolutional networks for action recognition in videos. arXiv Prepr. 568–576
53. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv Prepr. 1–14
54. Srivastava N, Salakhutdinov R (2012) Learning representations for multimodal data with deep belief nets. In: Int Conf Mach Learning Workshop. Edinburgh, Scotland, UK
55. Štruc V, Pavešić N (2010) The complete Gabor-fisher classifier for robust face recognition. EURASIP J Adv Signal Process 2010:1–26. https://doi.org/10.1155/2010/847680

56. Tan X, Triggs B (2010) Recognition under difficult lighting conditions. IEEE Trans Image Process 19: 1635–1650. https://doi.org/10.1109/TIP.2010.2042645
57. Tensorflow Library. In: TensorFlow. https://www.tensorflow.org/
58. Tiong LCO, Kim ST, Ro YM (2017) Multimodal face biometrics by using convolutional neural networks. J Korea Multimed Soc 20:170–178. https://doi.org/10.9717/kmms.2017.20.2.170
59. Wolf L, Hassner T, Maoz I (2011) Face recognition in unconstrained videos with matched background similarity. In: Int Conf Comput Vis Pattern Recognit (CVPR). IEEE, Colorado Springs, CO, USA, p 529–534
60. Woodard DL, Pundlik SJ, Lyle JR, Miller PE (2010) Periocular region appearance cues for biometric identification. In: Int Conf Comput Vis Pattern Recognit Workshop (CVPRW). IEEE, San Francisco, CA, USA, p 162–169
61. Wu X, He R, Sun Z, Tan T (2018) A light CNN for deep face representation with noisy labels. IEEE Trans Inf Forensics Secur 13:2884–2896. https://doi.org/10.1109/TIFS.2018.2833032
62. Xu Y, Lu Y (2015) Adaptive weighted fusion: a novel fusion approach for image classification. Neurocomputing 168:566–574. https://doi.org/10.1016/j.neucom.2015.05.070
63. Xu Y, Li Z, Pan JS, Yang JY (2013) Face recognition based on fusion of multi-resolution Gabor features. Neural Comput Appl 23:1251–1256. https://doi.org/10.1007/s00521-012-1066-3
64. Yang M, Zhang D, Feng X (2011) Fisher discrimination dictionary learning for sparse representation. In: Int Conf Comput Vis (ICCV). IEEE, Barcelona, Spain, p 543–550

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Leslie Ching Ow Tiong is currently pursuing his Ph.D. degree in KAIST. He received his BSc (Hons) degree in Computer Science from Sunway University in 2012, and MSc degree in Computer Science (by research) from Lancaster University in 2014. His research interests include biometrics, biometrics template protection, image processing, and deep learning.

Seong Tae Kim received the B.S. degree from Korea University, Seoul, South Korea, in 2012 and the M.S. degree from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2014. He is currently working towards the Ph.D. degree in the Image and Video Systems Laboratory at KAIST. In 2015, he was a visiting researcher at the University of Toronto in Toronto, Ontario, Canada. His research interests include computer vision, pattern recognition, biometrics, and medical image analysis. Mr. Kim received the Best 10% Paper Award of IEEE ICIP in 2015, the Honorable Mentioned Poster Award of SPIE Medical Imaging in 2015, the Samsung HumanTech Paper Award in 2017, and the Robert F. Wagner All Conference Best Student Paper Final Lists Award of SPIE Medical Imaging in 2018. He served as a session chair in IEEE International Conference on Image Processing 2015.

Yong Man Ro received the B.S. degree from Yonsei University, Seoul, Korea, and the M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. In 1987, he was a Visiting Researcher with Columbia University, New York, USA, and from 1992 to 1995, he was a Visiting Researcher with the University of California, Irvine, CA, USA, and KAIST. He was a Research Fellow with the University of California, Berkeley, CA, USA, and a Visiting Professor with the University of Toronto, Toronto, ON, Canada, in 1996 and 2007, respectively. He is currently a Professor in the School of Electrical Engineering, KAIST, where he is directing the Image and Video Systems Laboratory. His current research interests include image/video processing, 3-D video processing, image/video understanding, and visual recognition. Dr. Ro received the Young Investigator Finalist Award of ISMRM in 1992 and the Year's Scientist Award, Korea, in 2003. He was an Associate Editor in IEEE SIGNAL PROCESSING LETTERS. He has organized and served on technical program committees in many international conferences, including as the Program Chair of IWDW 2004. He co-organized special sessions on digital photo album technology in AIR 2005, social media in DSP 2009, and human 3-D perception and 3-D video assessments in DSP2011. He is currently an Associate Editor in Transactions on Data Hiding and Multimedia Security (Springer-Verlag).

Terms and Conditions