Multimodal Biometrics Recognition Using a Deep Convolutional Neural Network with Transfer Learning in Surveillance Videos

Hsu Mon Lei Aung 1, Charnchai Pluempitiwiriyawej 2,*, Kazuhiko Hamamoto 3 and Somkiat Wangsiripitak 4

1 Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand; hsumonlaiaung2012@gmail.com
2 Multimedia Data Analytics and Processing Research Unit, Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
3 School of Information and Telecommunication Engineering, Tokai University, Tokyo 108-8619, Japan; hama@keyaki.cc.u-tokai.ac.jp
4 Faculty of Information Technology, King Mongkut’s Institute of Technology Ladkrabang, Bangkok 10520, Thailand; somkiat@it.kmitl.ac.th

* Correspondence: charnchai.p@chula.ac.th

Abstract: Biometric recognition is a critical task in security control systems. Although the face has long been widely accepted as a practical biometric for human recognition, it can be easily stolen and imitated. Moreover, in video surveillance, it is a challenge to obtain reliable facial information from an image taken at a long distance with a low-resolution camera. Gait, on the other hand, has been recently used for human recognition because gait is not easy to replicate, and reliable information can be obtained from a low-resolution camera at a long distance. However, the gait biometric alone still has constraints due to its intrinsic factors. In this paper, we propose a multimodal biometrics system by combining information from both the face and gait. Our proposed system uses a deep convolutional neural network with transfer learning. Our proposed network model learns discriminative spatiotemporal features from gait and facial features from face images. The two extracted features are fused into a common feature space at the feature level. This study conducted experiments on the publicly available CASIA-B gait and Extended Yale-B databases and a dataset of walking videos of 25 users. The proposed model achieves a 97.3 percent classification accuracy with an F1 score of 0.97 and an equal error rate (EER) of 0.004.

Keywords: multimodal biometrics; human recognition; deep CNNs; transfer learning

1. Introduction

The identification of people is a critical task in biometric authentication, security control, and video surveillance systems. Although the face, iris, and fingerprints each have been used as a primary physical biometric to recognize people [1–4], it is a challenge to use each of them reliably in a typical surveillance system in an unconstrained environment. For example, a face biometric taken at a considerable distance from the camera is ineffective due to the low resolution of the face images. Gait traits, the unique walking style of an individual which cannot be faked easily, are the only biometric that is not sensitive to distance and the quality of capturing devices. However, it has some limitations due to clothing, carrying bags, or environmental circumstances [5]. Due to various encountered difficulties in obtaining feature patterns, the recognition performance in a unimodal biometric is decreased.

A surveillance system with multimodal biometrics supports more reliable information extraction than unimodal systems [6–8] for recognition accuracy improvement. Different fusion-level methods are employed for the information fusion of various modalities. Since the feature set contains richer information about the input biometrics pattern than the score
level [9], integrating biometrics information at the early-stage fusion may improve the recognition performance.

Machine learning (ML) methods have been widely used for biometric feature extraction from raw data and also as classifiers for recognition, but they have some limitations on feature discrimination and selection in various domains. Deep learning (DL), a new subcategory of machine learning, has been developed using artificial neural networks (ANN) with multiple hidden layers for extracting low-level features to abstract-level features. DL techniques include parallel and distributed data computing, adaptive feature learning, reliable fault tolerance, and hardy robustness characteristics [10]. Deep convolutional neural networks (deep CNN) have recently been used in biometric recognition systems [11]. The extensive training time, the massive amount of data and expensive and powerful GPUs for processing requirements are significant problems with deep CNN models. However, transfer learning (TL) could solve these problems by reusing the trained model for new tasks. It is also a machine learning method that can take learned features on one task and leverage them on a new task. TL is performed well for a task with a small amount of data for training a whole model from scratch. It also spectacularly reduces the cost and the training time, which results in an improved performance on related problems.

As a result of significant performance of deep learning methods in various recognition tasks, this paper proposed a transfer learning-based deep convolutional neural network (deep CNN) model to identify a person using two biometric traits, namely face and gait, with a small amount of training data. Our proposed network comprises eight convolutional layers, five pooling layers, and two fully connected (FC) layers. The network input is LR face or gait energy (GE) images of size $224 \times 224 \times 1$ pixels. The LR face images are captured directly from the video for recognition to avoid the LR face image creation and HR face image requirement. GE images are obtained from the silhouette images of a surveillance video sequence. The first convolutional layer extracts features from the input image by convolving the input to $3 \times 3$ filters. Next, a max subsampling layer is followed to reduce the feature map dimensionality and retain the most critical information. The extracted features of the previous layers are connected as every neuron in the last layer to every neuron on the next layer by the FC layer. The last FC layer is an output layer with the neurons associated with the total number of classes present in the datasets.

The proposed network is more straightforward in the architecture with more convolutional layers with fewer FC layers and fewer neurons in FC layers with 0.3 million trainable parameters. The deep CNNs with billions of learnable parameters require a large training dataset with many training images per class. Thus, the proposed network is preferable for the training and testing of small datasets. The proposed model is trained and tested individually on the popular CASIA-B gait dataset and Yale-B face dataset containing approximately eighty images per person. As for the recognition accuracies, over 97.1 percent in gait recognition and 95.8 percent in face recognition and short training time-consuming, the trained proposed model is used as a pre-trained model to transfer learned features knowledge to our target task. The model performed as a feature extractor by freezing the model layers to preserve existing learning. The classification model with an FC layer is developed to learn additional information and classification for unimodal biometric recognition. The extracted gait features and face features are combined using the feature-level concatenation method. The integrated features are used as input to our classification model for multimodal recognition.

Our proposed method is a combination of a face and gait recognition system using a deep transfer learning-based CNN approach. The proposed model performance is evaluated on a publicly available human walking video dataset.

2. Related Work

While unimodal systems have proven reliable in many studies, they have limitations such as noisy data sensing, the absence of distinguishable data representation, and non-universal properties of the modalities. As a result of single modality-based recognition prob-
lems, multimodal human recognition methods that combine multiple modalities of data have been developed. Several studies have proposed multimodal biometrics recognition systems with various combinations of different biometrics. This section discusses studies that applied different machine learning approaches in multimodal biometric systems.

According to the aforementioned machine learning approaches in biometric systems, the modalities require specialized feature extraction and selection methods for various types of biometrics. In deep learning-based biometric systems, the network automatically extracts the features from the input modalities. These deep learning abilities have solved the limitations of other machine learning algorithms. Therefore, deep learning-based biometric recognition methods have been developed recently.

Kim et al. [12] proposed a deep CNN of finger-vein and finger shape modalities. The near-infrared camera sensor captured the finger images. The matching distance scores of features were fused by weighted sum, product, and perceptron methods. Also, in [13], face, fingerprint, and iris biometrics recognition systems using deep learning template matching techniques were designed. Contourlet transform and local derivative ternary methods were utilized for feature extraction. The weighted rank-level algorithm fused the extracted features. In [14], the extracted elements of multi-stream CNNs with face, iris, and fingerprint modalities were fused by multi-level feature abstraction for identification. Al-Waisy et al. [15] proposed a deep belief network (DBN)-based facial feature extraction and recognition process and a CNN-based left and right iris feature extraction and classification process. The resulting scores are combined by multimodal biometrics identification for score-level and rank-level fusion methods. Ding et al. [16] proposed a CNN-based deep learning framework for multimodal face information-based recognition. A set of CNNs was implemented for extracting features of multiple face modalities. A stacked autoencoder was used for feature-level fusion.

However, deep CNN models achieved low performance when insufficient training data problems occurred. Transfer learning (TL) is often used to handle inadequate biometric data issues by transferring pre-trained knowledge from a source to related target domains. In [17], combined iris and periocular region modalities-based deep transfer learning methods were proposed for recognition. A VGG model was used for feature extraction, and the feature selection applied the binary particle swarm algorithm. The fusion of two modalities used matching score-level and feature-level fusion methods. Lumini et al. [18] presented various deep learning models with transfer learning for an automated plankton recognition system. Tao et al. [19] proposed a bidirectional feature extraction and transfer learning-based deep CNN for finger vein recognition. The original finger-vein image features and the rotated image features were fused by feature-level concatenation. Therar et al. [20] used a CNN with a transfer learning approach for multimodal right iris and left iris biometrics recognition. The extraction and classification of iris features were performed using the proposed deep learning and multi-class support vector machine (SVM) algorithm. In Ref. [21], finger vein and finger knuckle print features were extracted using three CNNs, alexnet, VGG16, and resnet50, interacting with transfer learning. The proposed fusion approaches combined these features for classification. Zhu et al. [22] developed a CNN-based deep transfer learning for the human identification framework. The framework learned and transferred optimal motion dynamics representation knowledge from the source domain tasks to the target domain.

Based on the automatic discriminative feature extraction ability of deep CNN methods, we propose a deep transfer learning-based CNN model to perform multimodal recognition with a small amount of data in this study. This research developed a pre-trained model as feature extractors and a classification model for face, gait, and multimodal recognition using the early-stage fusion method.
3. Materials and Methods

In this section, the general architecture of the proposed multimodal biometric recognition system is explained in Section 3.1. The detailed descriptions of the proposed approach follow in the subsequent Sections 3.2–3.4.

3.1. Proposed System Overall Procedure

A typical multimodal recognition framework mainly comprises biometric data acquisition and pre-processing, feature extraction, feature fusion, and classification components. The overall proposed multimodal system is shown in Figure 1. In the first step, a sequence of walking people frames is acquired from a surveillance video, and then human silhouette images are extracted using background subtraction. A person is walking in various directions in many video sequences. The adequately aligned single sequence of the silhouette of gait is averaged into a GE image. At the same time, the proposed system detects the face region in a video frame and extracts it as an LR face input image. Next, face and gait features are acquired by the proposed deep CNN model and concatenated these features directly at the feature level fusion. The unified feature vector is passed through the classification stage for multimodal recognition. The step-by-step flowchart of the proposed system is shown in Figure 2.

![Figure 1. General structure of the proposed multimodal biometrics system.](image)

3.2. Detection and Extraction of LR Face and Gait Energy Image as Pre-Processing

First, the pre-processing step involves the face detection of a person for acquiring an LR face image from the input frames. A cutting-edge deep learning-based facial detector, the retinaface [23], is used to detect the face. It is a robust single-stage pixel-wise facial localization method that samples face locations and scales on feature pyramids. The two trained models, mobilenet 0.25 with a 1.7 M model size and Resnet 50 with a 30 M size, are used as the pre-trained models for face detection. They predict the face score, face region by bounding box, and five facial landmarks (right eye, left eye, nose tip, right mouth, and left mouth) on the face. The face detection accuracy is approximately 87 percent on the input videos. Based on the face detection result, the system obtains the LR face region images from the frames.

The proposed system uses a GE averaging action to represent a human walking gesture sequence in a single image for gait feature extraction and recognition. The GE image can retain the original gait sequence data and has less sensitivity to a silhouette image noise. The first step in the GE image extraction is the foreground moving object detection. The simplified self-organized background subtraction method (simplified SOBS) [24] is adopted for sensing foreground objects from the background model. The median filter with consecutive frames is used to construct an initial background model. The Euclidean distance [25] calculates the minimum distance between the input pixel and the current background model by the image’s HSV hexagonal colour space \((h, s, v)\) to find the best match, as shown in Equation (1). The required distance value must not be larger than the...
thresholds set automatically using Otsu’s method. The founded best match is defined as a background pixel, and other pixels are defined as the foreground object component.

\[
d(b, I(x,y)) = \sqrt{(v bs_b \cos(h_b) - (v_is_i \cos(h_i)))^2 + (v bs_b \sin(h_b) - (v_is_i \sin(h_i)))^2 + (v_b - v_i)^2},
\]

where \(b\) is the intensity of the background pixel, and \(I(x,y)\) is the intensity of the non-background pixel at location \((x,y)\).

After getting the human walking binary silhouette image sequences, the proposed system forms horizontally centre-aligned and size-normalized silhouette images. The grey level of the GE image is an efficient spatio-temporal gait representation approach to human walking properties in a complete gait cycle for individual recognition by gait. Each silhouette image of a walking person describes the space-normalized energy image. GE is the average cycle of the silhouette images into one embodiment as the time-normalized accumulated energy image. A human gait cycle is composed of two phases: the stance and the swing. The step starts with the heel strike of one foot and continues until the
preparation of the same foot’s heel strike for the next step. This is defined as a complete gait cycle. A GE image is defined from gait silhouette image sequences as follows:

\[ G(x, y) = \frac{1}{N} \sum_{t=1}^{N} B_t(x, y) \]  

(2)

where \( N \) is the number of frames in a silhouette gait sequence of a cycle, \( t \) represents the frame number in the image sequence, and \( x \) and \( y \) are values in the 2D image coordinate. The proposed system considered an \( N \) value range from 20 to 25 frames per cycle. \( B_t(x, y) \) is the gait silhouette image frame \( t \) in the sequence. The resulting LR face region image and GE image are shown in Figure 3.

![Figure 3. Example of gait energy image and low-resolution face image extraction.](image)

3.3. Proposed Deep Convolutional Neural Network Architecture

The study used LR face and GE images instead of HR face and raw gait sequence as the input for the deep convolutional neural network. The detailed architecture of our proposed CNN model is shown in Table 1. The model consists of nine learned layers and uses a fixed input image size of 224 × 224 pixels per channel. The first layer is the convolutional layer, the main building block of the network, with 32 filters. The input volume is convolved with these 3 × 3 filters to produce an output feature map. The size of the feature maps for each layer is calculated as follows:

\[ \text{FeatureMap}_{\text{out}} = \frac{W - F + 2P}{S} + 1 \]  

(3)

where \( W \) is the input volume size, \( F \) is the filter size; \( P \) is the padding borders, and \( S \) is the stride size. The first layer is followed by a dropout with 0.5 rates and a rectified linear unit (relu) transfer function as an activation to prevent slower learning speed and overfitting in the learning process. Trainable parameters of each layer are the number of well learnable parameters of the layer affected by the backpropagation process. The formula in Equation (4) is used to calculate the learnable parameters in each layer. These parameters situate only in the convolutional and fully connected layers of the network.
Table 1. Description of our proposed base network model.

| Layer Number | Type of Layer | Number of Filters | Size of Feature Map | Trainable Parameters |
|--------------|---------------|-------------------|--------------------|---------------------|
| 0            | Image Input   | -                 | $224 \times 224 \times 1$ |                     |
| 1            | Conv2d-1      | 32                | $224 \times 224 \times 32$ | 320                 |
| 2            | dropout       | $p = 0.5$         |                     |                     |
| 3            | MaxPool2d_1   | 1                 | $112 \times 112 \times 32$ | 9248                |
| 4            | Conv2d-2      | 32                | $110 \times 110 \times 32$ |                     |
| 5            | MaxPool2d_2   | 1                 | $55 \times 55 \times 32$ |                     |
| 6            | Conv2d-3      | 64                | $53 \times 53 \times 64$ | 18,496              |
| 7            | Conv2d-4      | 64                | $51 \times 51 \times 64$ | 36,928              |
| 8            | MaxPool2d-3   | 1                 | $25 \times 25 \times 64$ |                     |
| 9            | Conv2d-5      | 128               | $25 \times 25 \times 128$ | 73,856              |
| 10           | Conv2d-6      | 128               | $23 \times 23 \times 128$ | 147,584             |
| 11           | MaxPool2d-4   | 1                 | $11 \times 11 \times 128$ |                     |
| 12           | Conv2d-7      | 256               | $9 \times 9 \times 256$ | 295,168             |
| 13           | Conv2d-8      | 256               | $7 \times 7 \times 256$ | 590,080             |
| 14           | MaxPool2d-5   | 1                 | $3 \times 3 \times 256$ |                     |
| 15           | Flatten       | -                 | $2304 \times 1$      |                     |
| 16           | Fc1           | -                 | $1024 \times 1$      | 2,360,320           |
| 17           | dropout       | $p = 0.5$         |                     |                     |
| 18           | Output Layer  | -                 | #Classes            |                     |

Every pooling layer of the proposed model performs the maximum subsampling operation to reduce the feature map dimensionality and retain the most critical information.

Trainable Parameters = $((\text{filter\_size} \times \text{Depth}) + 1) \times \text{number\_of\_filters}$  \hspace{1cm} (4)

The output from the previous convolutional layers is flattened before connecting to the last FC layers. These layers contain the weight and bias together with the neurons for connecting these neurons between two different layers.

The flattened vector is then passed through an FC layer and a 0.5-dropout layer that randomly drops out 50 percent of the nodes from the network to prevent the cause of overfitting issues. The last layer of the proposed model is also an FC layer. The output layer of the proposed model represents the output classification result of the classes of the related datasets.

The proposed network was trained and tested on the publicly available CASIA-B gait dataset [26] for gait recognition and the Extended Yale-B [27] for face recognition. The datasets include 110 GE images of each person for 124 people in the gait dataset and 60 side-face pictures for 38 people individually in the face dataset. We split these datasets into a 70/20/10 ratio for train/validation/test, respectively. The proposed model [28] had over 97 percent gait recognition accuracy and over 95 percent face recognition on the test data. The details of comparison of the proposed model and state-of-the-art (SOTA) models and recognition accuracies on the Yale-B dataset are shown in Section 4.

3.4. The Proposed CNN Model-Based Feature Extraction, A Classification Model Architecture and Multimodal Feature Fusion

The proposed model has performed recognition on reasonably large datasets. We also consider using it for other tasks with a small amount of data. Deep learning models require a large amount of data to get better performance than other techniques, need expensive GPUs, and take a long time to train, which increases the computational cost. The transfer learning (TL) technique can overcome these limitations of deep learning methods by reusing a trained model on a specific task as part of the training process for a different task.

In this step, the proposed deep CNN model was performed as a feature extractor by transferring knowledge from the base model to the classification model, as shown in Figure 4. There are two independent feature extractors—one is for gait, and the other
is for the face—that do not share their weight values. The proposed method freezes the pre-trained layers of a base model to preserve existing learning generic features in feature extraction. Then, the learned features are used as input for a new, smaller classification model. The model was comprised of two FC layers to learn additional information on a new dataset and perform a classification process for unimodal recognition.

The proposed system fuses the extracted features from the two feature extractors for multimodal biometric recognition and conducts a feature-level fusion method on all possible combinations of face features and gait features to generate the maximum number of concatenated feature vectors. The integrated feature vector is passed through the classifier layer for the multimodal recognition process.

4. Experimental Environment and Results

This study was intended to design an effective multimodal biometric recognition system in surveillance environments where the object is far from the cameras. We assessed the proposed multimodal biometric recognition system using feature-level fusion features of two modalities and benchmarked it with the performance of unimodal recognition systems: LR face modality only and gait modality only.

4.1. Dataset

The proposed system evaluated the walking category of the recognition of human actions video dataset [29]. The walking video dataset is composed of 25 subjects in four different scenarios, as shown in Figure 5, such as (d1) outdoors, (d2) with scale variation, (d3) outdoors with different clothes, and (d4) indoors. There are 100 video sequences of 25 people in total.

The study also used the LR face region images, as shown in Figure 6, corresponding to 25 subjects with 45 images per person. The human silhouette images are extracted by detecting foreground objects from 100 video sequences corresponding to 25 subjects. The proposed method extracted four sequences of silhouette images for each person in a
scenario. The range of 20–25 silhouette frames is used for an average GE image, as shown in Figure 7 and explained in Section 3.2 for each person.

Figure 6. Result samples of LR face region images for one person from video sequences.

Figure 7. Result samples of GE images for one person from video sequences.

4.2. Comparison of Architectures and Parameters of CNN Models

The architecture and essential parameters of the proposed network model and other state-of-the-art (SOTA) models are compared in Table 2. The proposed model has nine depth layers of 33.73 MB in size, and the input image is 224 pixels in height and width with a channel. The proposed model has lower learned layers with the lower trainable parameters as a small network. The small network can avoid the big network problem that causes slowness to lead to overfitting and even bigger problems. Many depth layers increased the weights that led to the model complexity. A small number of the training sets with the vast network caused overfitting and reduced accuracy.

Table 2. The depth, trainable parameters, and size of pre-trained models.

| Model            | Depth Layers | Input Image Size (h, w, c) | Size (MB) |
|------------------|--------------|---------------------------|-----------|
| AlexNet [30]     | 8            | $227 \times 227 \times 3$ | 219       |
| DenseNet121 [31] | 121          | $224 \times 224 \times 3$ | 26.92     |
| InceptionV3 [32] | 48           | $299 \times 299 \times 3$ | 96.58     |
| VGG16 [33]       | 16           | $224 \times 224 \times 3$ | 513.73    |
| proposed network | 9            | $224 \times 224 \times 1$ | 33.73     |

4.3. Comparison of Recognition Accuracy

First, the proposed network model evaluation was performed for unimodality face and gait recognition. More than two classes in the machine learning prediction are multi-class classification tasks.

Many evaluation metrics occurred on the confusion matrix that made the records of the number of counts of actual and the predicted classification. According to the derived face and gait image results from video sequences, the dataset used in this study was a balanced dataset type. The performance analysis of classification on multi-class balanced datasets used recognition accuracy as the most relevant metric because class distribution did not need to be considered [34]. The accuracy returned an overall measurement of the correct prediction of individual class among the total number of classes. The total number of the actual positive value of the whole system is obtained through the following equation:

$$TTP_{\text{all}} = \sum_{j=1}^{n} x_{jj}$$  \hspace{1cm} (5)
where \( n \) is the number of classes in a system and \( j \) is the row or column number of a matrix [35]. The overall accuracy of the model is computed using the following equation:

\[
\text{Accuracy} = \frac{\text{TTP}_{\text{all}}}{\text{Total of test samples}}
\]  

(6)

4.3.1. Unimodal Recognition

For the performance measurement of the proposed model and SOTA models of the unimodal recognition system, the Yale-B face is used to train the proposed and Alexnet model from scratch and train the popular SOTA plain ImageNet-train network models as the first experiment.

In Table 3, we compared the popular SOTA network models with our proposed model for face recognition on the Yale-B dataset. It contained an average of 60 images per person of 38 people.

**Table 3.** Comparison of the proposed model and SOTA models.

| Models         | Total Parameters | Trainable Parameters | Non-Trainable Parameters | Accuracy  | Training Time (Seconds) |
|----------------|------------------|----------------------|--------------------------|-----------|-------------------------|
| Alexnet        | 59,345,446       | 59,342,694           | 2752                     | 73.6 percent | 44                      |
| DenseNet121    | 19,892,582       | 12,855,078           | 7,037,504                | 95.6 percent | 718                     |
| InceptionV3    | 55,367,238       | 33,564,454           | 21,802,784               | 95.7 percent | 1513                    |
| VGG16          | 21,147,238       | 6,432,550            | 14,714,688               | 83.7 percent | 101                     |
| proposed network | 3,570,950       | 3,570,950            | 0                        | 95.8 percent | 14                      |

Inference time in DL algorithms refers to the time taken by DL to make predictions on new unseen data. The number of layers in the network, the number of neurons in each layer, and network model complexity can impact this time. Commonly, inference time rises with network size and complexity. Moreover, the more trainable parameters, the more extensive the network size, which then consumes a large number of computing resources.

The proposed network has the smallest number of trainable parameters and number of layers. It achieves higher accuracy with lower training time than other SOTA methods, leading to a decreased inference time for the prediction process. Our experiments are conducted on a GPU Tesla V100-SXM2-16GB Google Colab Cloud Platform with 50 GB RAM.

In the second experiment, we compared the accuracy of unimodal recognition with popular multi-class classification algorithms on the extracted face features from our proposed model-based feature extractor. The multi-class classification categorized the test data into multiple class labels present in the trained data as a model prediction.

Binary classification algorithms such as logistic regression and SVM (support vector machine) with one-vs-rest (OvR) and one-vs-one (OvO) strategies were also used as multi-class classifiers. The recognition accuracy comparison results are shown in Tables 4 and 5. Table 4 shows that the proposed model achieved higher accuracy with 89 percent in the OvR logistic regression classifier for gait recognition. In Table 5, the face recognition accuracies of different classifiers are described. The logistic regression and SVM with OvO strategy gained 86 percent more accuracy than other classifiers for face recognition.
Table 4. Gait recognition accuracy comparison on different classifiers.

| Classification Algorithms | Accuracy (Percent) |
|---------------------------|--------------------|
| K-nearest neighbors       | 64                 |
| Decision tree             | 42                 |
| Naive Bayes               | 59                 |
| Random forest             | 73                 |
| Gradient boosting         | 60                 |
| Logistic regression (OvR) | 89                 |
| Logistic regression (OvO) | 88                 |
| SVM (OvR)                 | 84                 |
| SVM (OvO)                 | 84                 |

Table 5. Face recognition accuracy comparison on different classifiers.

| Classification Algorithms | Accuracy (Percent) |
|---------------------------|--------------------|
| K-nearest neighbors       | 74                 |
| Decision tree             | 49                 |
| Naive Bayes               | 73                 |
| Random forest             | 73                 |
| Gradient boosting         | 69                 |
| Logistic regression (OvR) | 80                 |
| Logistic regression (OvO) | 86                 |
| SVM (OvR)                 | 85                 |
| SVM (OvO)                 | 86                 |

4.3.2. Multimodal Recognition

Multimodal biometric recognition was performed as a third experiment. The fused features were classified using multi-class classification algorithms, and the recognition accuracies are compared in Table 6. The highest accuracy was 97 percent on the OvR logistic regression classifier.

Table 6. Multimodal recognition accuracy comparison on different classifiers.

| Classification Algorithms | Accuracy (Percent) |
|---------------------------|--------------------|
| K-nearest neighbors       | 65                 |
| Decision tree             | 51                 |
| Naive Bayes               | 79                 |
| Random forest             | 88                 |
| Gradient boosting         | 75                 |
| Logistic regression (OvR) | 97                 |
| Logistic regression (OvO) | 96                 |
| SVM (OvR)                 | 88                 |
| SVM (OvO)                 | 94                 |

Logistic regression and SVM classifier-based models obtained higher accuracy in both unimodal and multimodal recognition. After these two classification algorithms, random forest accurately classified the gait and face traits with 73 percent and 88 percent accuracy in multimodal biometric recognition. The decision tree classifier provided a less accurate result for all our recognition systems. Although the classification using naive Bayes and gradient boosting methods worked better on the face than gait classification, K-nearest neighbors-based gait recognition was higher than face recognition. Multimodal recognition using various classifiers except K-nearest was more effective than the unimodal system.

As seen in Table 7, the proposed multimodal recognition method was compared with other existing approaches. Our multimodal system showed comparable accuracy as opposed to other multimodal systems. Although Derbel’s method used the frontal view face and gait silhouette images extracted from near the camera, the proposed multimodal
system applied combined side-face LR and GE images from a long-distance camera with various clothes and environment situations. Moreover, Derbel’s unimodal face trait was more effective than the gait trait, and our proposed unimodal was equally effective on both biometrics. Therefore, our proposed model can be more reliable in a typical surveillance unconstrained environment at a considerable distance from the camera.

Table 7. Recognition accuracy comparison of various multimodal methods.

| Papers             | Modalities      | Method                             | Recognition Accuracy (Percent) |
|--------------------|-----------------|------------------------------------|-------------------------------|
| Zhou et al. [36]   | Face (side) + gait | MDA, PCA                           | 87                            |
| Hossain et al. [37]| Face (side) + gait | PCA, LDA, holistic or hierarchical fusion | 75                            |
| Zhang et al. [38]  | Face + gait     | LBP, CCA                           | 90                            |
| Xing et al. [39]   | Face + gait     | Projection of heterogeneous features, KNN | 94                            |
| Derbel et al. [40] | Face + gait     | Merge descriptors, PCA, NN         | 97.4                          |
| Our proposed method | Face + gait     | Deep CNN with transfer learning    | 97.3                          |

4.4. Comparison of F1-Score

The F1-score is the weighted average of precision and recall used in all classification algorithms. An F1-score value becomes one that is considered perfect. F1-score is a better evaluation metric if there are imbalanced class distribution in the datasets. The result of the F1-score for gait, face and multimodal recognition are described in Table 8.

Table 8. F1-score comparison on different classifiers of unimodal and multimodal recognition.

| Classification Algorithms | Gait F1-Score | Face F1-Score | Multimodal F1-Score |
|--------------------------|--------------|--------------|---------------------|
| K-nearest neighbours     | 0.61         | 0.72         | 0.63                |
| Decision tree            | 0.41         | 0.50         | 0.49                |
| Naïve Bayes              | 0.58         | 0.72         | 0.79                |
| Random forest            | 0.71         | 0.72         | 0.87                |
| Gradient boosting        | 0.59         | 0.69         | 0.72                |
| Logistic regression (OvR)| 0.85         | 0.79         | 0.97                |
| Logistic regression (OvO)| 0.87         | 0.85         | 0.96                |
| SVM (OvR)                | 0.84         | 0.84         | 0.87                |
| SVM (OvO)                | 0.80         | 0.86         | 0.94                |

The resulting score of logistic regression with OvR provided the highest F1-score value for multimodal biometrics. The score of the first two classifiers in the face modality had a more significant result than the other two modalities. However, different classifiers obtained a higher score in multimodal than unimodal biometrics.

4.5. Result in Terms of Error Rate

The equal error rate (EER) is one of the evaluation metrics for measuring and comparing the biometric systems’ performance. It can be obtained by calculating the error value of the false acceptance rate (FAR) and false rejection rate (FRR) equal. The FAR, a synonym of
false-positive rate (FPR), and FRR, also known as false-negative rate (FNR), are calculated by the following equations:

\[
\text{FAR} = \frac{FP}{(FP + TN)}
\]

\[
\text{FRR} = \frac{FN}{(TP + FN)}
\]

where \(FP\) is the number of false positives, \(TN\) is the number of true negatives, \(TP\) is the number of true positives, and \(FN\) is the number of false negatives. EER is the inverse of accuracy, which means the lower the EER value, the better the accuracy of the biometric system. Table 9 describes the EER values of the proposed biometrics system. The proposed multimodal system’s EER was smaller than the unimodal biometrics system.

Table 9. Equal error rates for the proposed biometric recognitions method.

| Biometrics | EER  |
|------------|------|
| Gait       | 0.032|
| Face       | 0.055|
| Multimodal | 0.004|

In this paper, we proposed a deep CNN model with small trainable parameters for biometric recognition. The model trained with a small amount of data provided high recognition accuracy on two biometrics. Then, the proposed deep CNN model was performed as a feature extractor by extracting features from LR face images and GE images acquired from low-quality surveillance videos as inputs to the classification model for biometrics recognition on the small amount of data. Our proposed system uses a simplified SOBS method based on moving object detection and silhouette image extraction. Then, GE images are subtracted from the horizontally centre-aligned and size-normalized silhouette images. It overcomes the problems with the misalignment of inaccurate human body region segmentation, which exists in other GE-based methods for gait recognition.

Secondly, most previous studies used HR face images for LR face recognition by synthetically generating the corresponding LR image and mapping function. This study used the RetinaFace detection technique to directly detect the LR face region as an input image to the recognition process. It reduced the processing time and complexity. The proposed multimodal recognition system applied a feature-level fusion method to integrate learned features that fusion provided more discriminative information about input biometrics. The deep CNN-based feature extractor performed dimensionality reduction to overcome the curse of dimensionality and effectively transferred learned knowledge from one task to other related tasks. Therefore, the proposed system can provide better performance on the small amount of available data.

5. Conclusions

In this study, a multimodal biometrics recognition framework with a deep learning approach was proposed, along with the various classifiers, to achieve better person recognition accuracy. In the experiment, the extracted LR face and GE images from a publicly available videos database are utilized as input to the proposed CNN model trained on public datasets for learning and extracting practical features. Then, the extracted features were transferred to the classification model to classify a person. Seven kinds of classifiers were evaluated, namely, K-nearest neighbors, decision tree, naïve Bayes, random forest, gradient boosting, logistic regression and SVM, on three modalities. The experiment indicated that logistic regression was the most suitable classifier for human identification because it provided the average highest accuracy near 97 percent with multimodal biometrics, which was compatible in comparison to other multimodal recognition methods. Furthermore, the experiment was conducted on a public dataset to compare the recognition performance and the architecture and parameters of the proposed model and other SOTA, namely VGG
Net-16, Inception_v3, Densenet, and Alexnet. The proposed network has the smallest number of trainable parameters and achieves higher accuracy with lower training time than other SOTA. Moreover, in the EER metric-based comparison, the proposed method obtained an ER of 0.055 for the face modality, and the gait modality has an EER of 0.032 and 0.004 in a multimodal system.

**Author Contributions:** H.M.L.A., C.P. and K.H. designed and developed the study; H.M.L.A. performed the experiments, data analysis, evaluation, and manuscript writing; C.P., S.W. and K.H. coordinated the whole efforts as the supervisors of this research. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Acknowledgments:** The authors sincerely appreciate the ASEAN University Network (AUN/SEED-Net) for their financial support in contributing to this research.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Maltoni, D.; Maio, D.; Jain, A.K.; Prabhakar, S. *Handbook of Fingerprint Recognition*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.

2. Buddharpawar, A.S.; Subbaraman, S. Iris recognition based on pca for person identification. *Int. J. Comput. Appl.* 2015, 975, 8887.

3. Rosdi, B.A.; Shing, C.W.; Suandi, S.A. Finger vein recognition using local line binary pattern. *Sensors* 2011, 11, 1357. [CrossRef] [PubMed]

4. Joshi, A.; Gangwar, A.K.; Saquib, Z. Person recognition based on fusion of iris and periocular biometrics. In Proceedings of the 2012 12th International Conference on Hybrid Intelligent Systems (HIS), Pune, India, 4–7 December 2012; pp. 57–62.

5. Han, J.; Bhanu, B. Individual recognition using gait energy image. *IEEE Trans. Pattern Anal. Mach. Intell.* 2005, 28, 316–322. [CrossRef] [PubMed]

6. Kim, Y.G.; Shin, K.Y.; Lee, E.C.; Park, K.R. Multimodal biometric system based on the recognition of face and both irises. *Int. J. Adv. Robot. Syst.* 2012, 9, 65. [CrossRef]

7. Gomez-Barrero, M.; Galbally, J.; Fierrez, J. Efficient software attack to multimodal biometric systems and its application to face and iris fusion. *Pattern Recognit. Lett.* 2014, 36, 243–253. [CrossRef]

8. Liu, H.F.; Isa, D. Feature selection for support vector machine-based face-iris multimodal biometric system. *Expert Syst. Appl.* 2011, 38, 1105–1111. [CrossRef]

9. Bhanu, B.; Han, J. *Human Recognition at a Distance in Video*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2010.

10. Wang, P.; Fan, E.; Wang, P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognit. Lett.* 2021, 141, 61–67. [CrossRef]

11. Boucherit, I.; Zmirli, M.O.; Hentabli, H.; Rosdi, B.A. Finger vein identification using deeply-fused Convolutional Neural Network. *J. King Saud Univ. Comput. Inf. Sci.* 2020, 34, 346–656. [CrossRef]

12. Kim, W.; Song, J.M.; Park, K.R. Multimodal biometric recognition based on convolutional neural network using the fusion of finger-vein and finger shape using near-infrared (NIR) camera sensor. *Sensors* 2018, 18, 2296. [CrossRef]

13. Gunasekaran, K.; Raja, J.; Pitchai, R. Deep multimodal biometric recognition using contourlet derivative weighted rank fusion with human face, fingerprint and iris images. *Autom. Časopis Za Autom. Mjer. Elektron. Računarstvo I Komun.* 2019, 60, 253–265. [CrossRef]

14. Soleymani, S.; Dabouei, A.; Kazemi, H.; Dawson, J.; Nasрабadi, N.M. Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 3469–3476.

15. Al-Waisy, A.S.; Qahwaji, R.; Ipson, S.; Al-Fahdawi, S. A multimodal biometric system for personal identification based on deep learning approaches. In Proceedings of the 2017 Seventh International Conference on Emerging Security Technologies (EST), Canterbury, UK, 6–8 September 2017; pp. 163–168.

16. Ding, C.; Tao, D. Robust face recognition via multimodal deep face representation. *IEEE Trans. Multimed.* 2015, 17, 2049–2058. [CrossRef]

17. Silva, P.H.; Luz, E.; Zanolrensi, L.A.; Menotti, D.; Moreira, G. Multimodal feature level fusion based on particle swarm optimization with deep transfer learning. In Proceedings of the 2018 IEEE Congress on Evolutionary Computation (CEC), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8.

18. Lumini, A.; Nanni, L. Deep learning and transfer learning features for plankton classification. *Ecol. Inform.* 2019, 51, 33–43. [CrossRef]
19. Tao, Z.; Zhou, X.; Xu, Z.; Lin, S.; Hu, Y.; Wei, T. Finger-Vein Recognition Using Bidirectional Feature Extraction and Transfer Learning. *Math. Probl. Eng.* 2021, 2021, 6664809. [CrossRef]

20. Therar, H.M.; Mohammed, L.D.E.A.; Ali, A.J. Multibiometric System for Iris Recognition Based Convolutional Neural Network and Transfer Learning. In Proceedings of the IOP Conference Series: Materials Science and Engineering, The Fifth Scientific Conference for Engineering and Postgraduate Research (PEC 2020), Baghdad, Iraq, 21–22 December 2020; p. 012032.

21. Daas, S.; Yahi, A.; Bakir, T.; Sedhane, M.; Boughazi, M.; Bourennane, E.-B. Multimodal biometric recognition systems using deep learning based on the finger vein and finger knuckle print fusion. *IET Image Process.* 2020, 14, 3859–3868. [CrossRef]

22. Zhu, H.; Samtani, S.; Chen, H.; Nunamaker Jr, J.F. Human identification for activities of daily living: A deep transfer learning approach. *J. Manag. Inf. Syst.* 2020, 37, 457–483. [CrossRef]

23. Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; Zafeiriou, S. Retinaface: Single-stage dense face localisation in the wild. *arXiv* 2019, arXiv:1905.00641.

24. Maddalena, L.; Petrosino, A. A self-organizing approach to background subtraction for visual surveillance applications. *IEEE Trans. Image Process.* 2008, 17, 1168–1177. [CrossRef]

25. Selvarasu, N.; Nachiappan, A.; Nandhitha, N. Euclidean distance based color image segmentation of abnormality detection from pseudo color thermographs. *Int. J. Comput. Theory Eng.* 2010, 2, 514. [CrossRef]

26. Yu, S.; Tan, D.; Tan, T. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR’06), Hong Kong, China, 20–24 August 2006; pp. 441–444.

27. Georgiades, A.S.; Belhumeur, P.N.; Kriegman, D.J. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.* 2001, 23, 643–660. [CrossRef]

28. Aung, H.M.L.; Pluempitiviriyawej, C. Gait Biometric-based Human Recognition System Using Deep Convolutional Neural Network in Surveillance System. In Proceedings of the 2020 Asia Conference on Computers and Communications (ACCC), Singapore, 18–20 September 2020; pp. 47–51.

29. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 26 August 2004; pp. 32–36.

30. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 2012, 25, 1097–1105. [CrossRef]

31. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.

32. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.

33. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* 2014, arXiv:1409.1556.

34. Grandini, M.; Bagli, E.; Visani, G. Metrics for multi-class classification: An overview. *arXiv* 2008, arXiv:2008.05756.

35. Manliguez, C. Generalized Confusion Matrix for Multiple Classes. November 2016, pp. 2–4. Available online: https://www.researchgate.net/publication/310799885_Generalized_Confusion_Matrix_for_Multiple_Classes (accessed on 2 May 2022).

36. Zhou, X.; Bhanu, B. Feature fusion of face and gait for human recognition at a distance in video. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR’06), Hong Kong, China, 20–24 August 2006; pp. 529–532.

37. Hossain, E.; Chetty, G. Multimodal identity verification based on learning face and gait cues. In Proceedings of the International Conference on Neural Information Processing, Shanghai, China, 13–17 November 2011; pp. 1–8. Available online: https://researchprofiles.canberra.edu.au/en/publications/multimodal-identity-verification-based-on-learning-face-and-gait (accessed on 2 May 2022).

38. Zhang, D.; Wang, Y.; Zhang, Z.; Hu, M. Ethnicity classification based on fusion of face and gait. In Proceedings of the 2012 5th IAPR International Conference on Biometrics (ICB), New Delhi, India, 29 March–1 April 2012; pp. 384–389.

39. Xing, X.; Wang, K.; Lv, Z. Fusion of gait and facial features using coupled projections for people identification at a distance. *IEEE Signal Process. Lett.* 2015, 22, 2349–2353. [CrossRef]

40. Derbel, A.; Vivet, D.; Emile, B. Access control based on gait analysis and face recognition. *Electron. Lett.* 2015, 51, 751–752. [CrossRef]