Multi-modal Authentication Model for Occluded Faces in a Challenging Environment

Dahye Jeong, Eunbeen Choi, Hyeongjin Ahn, Ester Martinez-Martin, Eunil Park, and Angel P. del Pobil

***Abstract*—Authentication systems are crucial in the digital era, providing reliable protection of personal information. Most authentication systems rely on a single modality, such as the face, fingerprints, or password sensors. In the case of an authentication system based on a single modality, there is a problem in that the performance of the authentication is degraded when the information of the corresponding modality is covered. Especially, face identification does not work well due to the mask in a COVID-19 situation. In this paper, we focus on the multi-modality approach to improve the performance of occluded face identification. Multi-modal authentication systems are crucial in building a robust authentication system because they can compensate for the lack of modality in the uni-modal authentication system. In this light, we propose DemoID, a multi-modal authentication system based on face and voice for human identification in a challenging environment. Moreover, we build a demographic module to efficiently handle the demographic information of individual faces. The experimental results showed an accuracy of 99% when using all modalities and an overall improvement of 5.41%–10.77% relative to uni-modal face models. Furthermore, our model demonstrated the highest performance compared to existing multi-modal models and also showed promising results on the real-world dataset constructed for this study.**

***Index Terms*—Human authentication, user identification, multi-modalities, face, voice, demographic information.**

I. Introduction

With the increase in computer applications in every sector of society, the need for a reliable authentication system has become increasingly important. PC frameworks, workstations, PDAs, ATMs, mobile terminal-based payments, and even buildings require appropriate authentication systems to allow access to only qualified users [1]. Without appropriate and robust authentication systems, these frameworks are vulnerable to the deceits of an attacker [1]. For instance, credit card extortion can cost a business large sums of money every year without a powerful client authentication system [1]. Authentication enables organizations to maintain the security of their networks by permitting only authenticated users or processes to gain access to their protected resources, ranging from computer systems, networks, databases, and websites to other network-based applications or services. For such reasons, several systems with robust authentication methods have been proposed to maintain secure management of information [2].

However, with the outbreak of COVID-19, an infectious disease caused by severe acute respiratory syndrome (SARS-CoV-2) [3], authentication systems have encountered great difficulty in maintaining reliable performance due to the occluded information [4]. In fact, authentication applications relying on face recognition have nearly failed since mask mandates have effaced the prime element of authentication. Because such systems are trained to pay attention to important facial features, such as the eyes, nose, lips, and face edges, the system cannot maintain acceptable performance when these features are not visible due to masks. Moreover, authentication systems, which require physical contacts, such as fingerprints or password sensors, are no longer considered a good form of authentication since they can facilitate the spread of the disease. Generally, this COVID-19-specific situation has severely impacted current security systems and entailed measures to work toward a system that avoids unnecessary contact and does not assume a fully revealed face.

To efficiently handle partial occlusion or uncertainties in the COVID-19 situation, several artificial intelligence approaches have been developed to predict and analyze pandemic challenges. For example, [5] attempted to estimate the structure and scale of the uncertainties in the pandemic casualties in Turkey by constructing three different identification approaches. The results confirmed the validity of the developed models, which analyze unknown uncertainties and predict future COVID-19 casualties. Prior studies developed an artificial intelligence-based long-term policy-making algorithm to minimize COVID-19 losses [6]. Based on the proposed algorithm of [7], multifaceted interventions of non-pharmacological policy, including lockdowns, curfews, or schools' closures, are produced to be applied to the ever-changing pandemic with consideration of unknown uncertainties.

Despite these efforts, there were still difficulties in the COVID-19 situation, and further improvement was needed, especially in the authentication system. Because of the above

Manuscript received 13 March 2023; revised 4 February 2024; accepted 21 March 2024. This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea, funded by the Ministry of Education under Grant NRF-2022R1A6A3A13063417, in part by the Government of the Republic of Korea (MSIT), and in part by the National Research Foundation of Korea under Grant NRF-2023K2A9A1A01098773. *(Corresponding authors: Eunil Park; Angel P. del Pobil.)*

Dahye Jeong, Hyeongjin Ahn, and Eunil Park are with the Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul 03063, South Korea (e-mail: eunilpark@skku.edu).

Eunbeen Choi and Angel P. del Pobil are with the Department of Computer Science and Engineering, Jaume I University, 12006 Castellon, Spain (e-mail: pobil@uji.es).

Ester Martinez-Martin is with the Department of Computer Science and Artificial Intelligence, University of Alicante, 03690 Alicante, Spain.

Recommended for acceptance by G. Huang.

Digital Object Identifier 10.1109/TETCI.2024.3390058

difficulties, several authentication methods have been developed, such as physiological Doppler rader [8], acoustical palmprint [9], radio-frequency identification [10] to identify a user in a COVID-19 environment. Among various authentication methods, we focus on the multi-modal learning approach to efficiently identify the occluded faces. Multi-modal authentication enhances the security of user authentication by combining two or more identifiers, such as fingerprint, face, and iris [11]. Since multi-modal authentication uses multi-modality fusion, it can compensate for the lack of modality in a challenging environment (e.g., an occluded face by a mask and an unclear voice due to noise).

Therefore, we propose a multi-modal authentication system based on face and voice to preserve the individual's facial features and improve the performance of occluded face identification. We consider demographic information (e.g., gender, age, and race) to extract the individual's facial characteristics and combine it with face and voice. Our proposed model can identify users despite occluded faces and can be used in a contactless service for infection prevention. The proposed model code and dataset are publicly available.[1] Our research question and main contributions are presented as follows:

- *Research Question:* Will multi-modal authentication models using face, voice, and demographic factors perform better than uni-modal identification models in a challenging environment?

We propose DemoID, a multi-modal authentication model using facial, vocal, and demographic features for the occluded faces in a challenging environment. To efficiently handle the demographic information (i.e., age, gender, and race), we suggest a demographic module designed to compute the co-attention between demographic details and facial features. The experimental results show that the proposed model surpasses current human identification methods, particularly in recognizing occluded faces with remarkable accuracy. Moreover, the model demonstrates robust performance when evaluated on a real-world dataset.

II. Related Work

A. Multi-Modal Authentication

User authentication has been widely used to protect information technology systems against unauthorized user activities [12]. Authentication is "*the process of identifying someone or something to provide access control for systems through a matching process of users' data with the data stored in an authorized database*" [13].

Among various current authentication and authorization techniques, biometric authentication methods have attracted significant. Biometric authentication is the method of using anatomical or physical characteristics (e.g., face, iris, palm print, hand geometry, hand vein, fingerprints, finger vein, ear shape, tooth shape, and electrocardiogram) and behavioral characteristics (e.g., voice, gait, signature, and keystroke dynamics) to identify users [13]. Such technology is considered to achieve a higher level of security than other traditional authentication systems because biometrics cannot be copied, shared, lost, forgotten, manipulated, or forged [14]. The applications of biometric authentication include various fields, such as criminal investigation [15], logical and physical access control [16], surveillance [17], healthcare [18], and Internet-of-Things applications [19].

Most recent biometric authentication techniques take advantage of a uni-modal biometric authentication scheme to execute their authenticating process. However, uni-modal biometric authentication suffers from various challenges in terms of noise artifacts, data insufficiency, and data security [13]. Additionally, uni-modal authentication systems tend to have poor security reliability when a modality corresponding to the system is defective. As a countermeasure, several researchers have proposed multi-modal biometric authentication methods [20], [21], [22]. For example, Zhang et al. [23] proposed a fused model of face and voice information to provide reliable and convenient biometric authentication in a smartphone environment. Moreover, Joseph et al. [24] employed a multi-modal authentication system that integrates fingerprint, iris, and palm print biometric features. They converted the fused features into the hash of strings and numbers using an MD-5 hashing algorithm.

Since multi-modal biometric authentication combines different biometrics to overcome the problems of uni-modal authentication systems, they have been expected to be more reliable and convenient than conventional uni-modal biometrics [25], [26], [27]. However, previous multi-modal authentication studies have not taken into account the user's demographic information, which could capture facial characteristics. Demographic information extracted from modality can be vital information to identify the user if there is a lack of modality or if features of modalities are similar. Moreover, Demographic information can be hidden features in the occluded face. Therefore, we extract demographic information from the modality and use it as an authentication key.

B. Face Recognition

Face recognition is a well-known biometric technique for identity authentication, which has been widely applied in many areas, including military, finance, and daily life [28]. For such reasons, face recognition has been a trending research topic in computer vision. Traditionally, face recognition utilizes margin-based metrics to increase intra-class compactness and train models with a huge amount of data to improve their performance [29]. In 2012, AlexNet reshaped the research landscape of face recognition after winning the ImageNet competition using a deep learning technique [30]. Deep convolutional neural networks automatically learn the lower-level features from the initial layers, and the higher-level abstract features from the deeper layers, significantly improving the level of performance. Inspired by this work, researchers shifted their focus toward deep learning-based approaches. In 2014, DeepFace [31] attained SOTA accuracy on the famous labeled faces in the wild benchmark [32], approaching human performance on the unconstrained condition for the first time (DeepFace: 97.35% vs. Human: 97.53%), by training a nine-layer model on four million facial images. In only three

[1] https://github.com/jeongdahye3427/multi-modal-identification

years, the accuracy of face recognition techniques dramatically increased to above 99.80% using deep learning-based methods, surpassing humans in several scenarios [31], [33], [34].

C. Masked Face Recognition

Several models present outstanding performance in face recognition. However, a large proportion of the models perform well using unoccluded face data. The maximum amount of occlusions of the face due to sunglasses, mustache, bangs, or hats under which these models could perform well was only a small part of the face [35], [36]. Since COVID-19 masks occlude around 70% of the face area [37] (e.g., mouth, chin, and nose), specific studies and methods aimed at masked face recognition have arisen since the outbreak of COVID-19.

One of the first necessities for masked face recognition under COVID-19 is the creation of international datasets of real or simulated face masks. Because of the lack of training and testing datasets with masked face images, such masked datasets are a prerequisite before any modeling. Therefore, Wang et al. [38] proposed three types of masked face datasets: masked face detection dataset, real-world masked face recognition dataset, and synthetic masked face recognition dataset. Moreover, they claimed to have enhanced recognition accuracy from 50% to 95%. Similarly, Anwar et al. [39] presented an open-source tool, MaskTheFace, to mask faces effectively and create a large dataset of masked faces that they used to train a facial recognition system with target accuracy for masked faces. They reported an increase of about 38% in the true positive rate for the Facenet system [39].

Other studies have focused on enhancing the recognition performance of masked faces. For example, Mundial et al. [40] trained the support vector machine (SVM) classifier using the feature vector embeddings provided by the FaceNet model on a collected small database, reporting 99% accuracy. Li et al. [41] introduced a de-occlusion distillation framework, where appearance information is recovered using a generative inpainting network, after which rich structural knowledge is transferred from a pre-trained general recognizer in a teacher-student model. Vu et al. [42] proposed a method for extracting deep features based on ArcFace [43] from the detected and normalized face images. They combined them with local binary pattern features extracted from the eyes and eyebrows. Other works focused on benchmarking the performance of masked-face-recognition methods [37], [44], [45], [46]. However, some quite competitive algorithms with full faces still fail to authenticate between 10% and 40% of masked images [37]. Additionally, some studies do not provide particular design details of the tested algorithms or database or a clear evaluation protocol and its procedures [37], [46], [47]. Generally, masked facial recognition models do not yield a sufficient level of performance to be used alone to authenticate securely and robustly. Moreover, the use of a multi-modal approach to compensate for the low performance of masked facial recognition models has yet to be considered. Because of these points, this paper proposes a multi-modal authentication model trained on basic facial and masked facial datasets to boost the recognition performance of masked faces. Since the proposed model uses multi-modalities (e.g., face, voice, and demographic information), it can also compensate for the low performance due to the mask.

D. Voice Recognition

Voice biometric recognition leverages unique features of the human voice to identify and authenticate a user [48]. The construction of the articulatory apparatus and its use that generate and modulate the voice of a talker, including the lungs, vocal cords, and articulators, are uniquely configured for each individual [48], making voice a powerful authentication method [49]. Accordingly, several voice recognition studies have been conducted using different methods. There are two main stages in voice recognition: feature extraction and matching. Extracting the speech signal is the main task of speaker recognition systems. Two popular sets of features often employed in the analysis of the speech signal are the Mel frequency cepstral coefficients (MFCCs) and linear prediction cepstral coefficients. Moreover, voice recognition techniques can be classified into four categories:

- Systems with vector quantification [50], [51], [52];
- Systems with Gaussian mixture models [53], [54], [55];
- Systems with factor analysis [56];
- Recent systems with deep neural networks [57], [58].

Recent studies have focused on strengthening the voice-based authentication system against replay and impersonation attacks. Shang et al. [59] examined the vibrations of the vocal cords to prevent voice spoofing, requiring users to press the phone against their throat. Chen et al. [60] employed the magnetic field information emitted from loudspeakers to defend against replay attacks. They also require users to move the phone in a specific trajectory during authentication. Other studies employed information, such as pop noise, lip movement, articulatory gestures, and vocal system, as biometrics to create systems robust against potential replay and impersonation attacks [61], [62], [63], [64]. Although there are several studies on decreasing voice recognition performance concerning noise, this paper attempts to approach it using a multi-modal authentication method. Most previous studies rely on uni-modal authentication, and prior multi-modal models did not consider users' demographic information. Thus, this paper aims to combine the user's demographic information with the biometric modality (face and voice) to model it to have good performance even with defects of a specific modality.

III. Dataset

A. Face Dataset

We used the VGGFace2 dataset [65] for face recognition. The VGGFace2 dataset consists of about 3.31 million high-diversity images (i.e., pose and age), divided into 9,131 classes. Each class represents the identity of a different person, with an average of about 361 photos per identity [65]. The images of the training set had an average resolution of 137×180 pixels, with less than 1% at a resolution below 32 pixels. From this dataset, we randomly sampled ten classes (4,464 images) and divided each class into

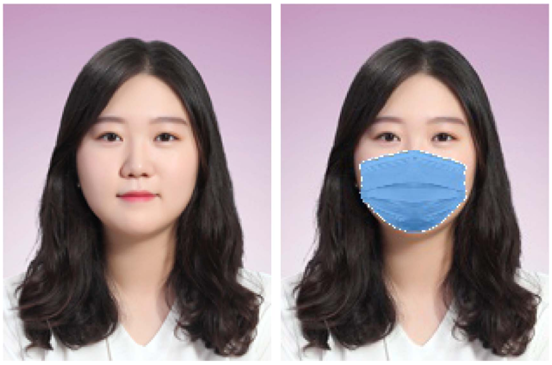

Fig. 1. Examples of the face and masked face images (surgical mask type): We used five mask types (surgical, cloth, empty, KN95, and N95).

8:1:1 for training (2,978 images), validating (742 images), and testing (744 images). We chose a limited number of classes to synchronize face and voice data to construct a unique virtual identity for each individual.

We employed the MaskTheFace model to generate a dataset of faces with masks. MaskTheFace is a publicly available package that converts face images into masked-face images[2]. The model uses the dlib-based face landmark detector to identify face tilt and six key features of the face, which are necessary for applying the mask. Based on the face tilt, we applied ten different types of mask templates on the face. After the masking process, we conducted preprocessing to input the images into the model. First, we removed the images that failed to mask because the bounding box of the face was not recognized because of the small size of the face. Second, we set the size of all the images as 152×152×3 since the input size of the face module needs to be unified. Finally, we reduced the number of mask types from ten to five and removed infrequently used mask types, such as gas masks, yielding a total of 15,996 training, 4,007 validation, and 4,026 testing images, with 24,029 augmented images. Fig. 1 shows an example of the face and mask images.

B. Voice Dataset

For voice recognition, we utilized the LibriSpeech corpus dataset[3]. LibriSpeech is a corpus of approximately 1000 hours of 16 kHz English speech derived from audiobooks. We used a subset of the train-clean-100.tar.gz [6.3G]. This subset consists of speaking data for an average of 23.72 minutes, and we randomly sampled 251 identities to match the face data. Moreover, we split them into record files with an average duration of 14 seconds. Then, we extracted five types of vocal features: MFCCs, Chromagram, Mel-scaled spectrogram, spectral contrast, and tonal centroid features (tonnetz), from each sound file and saved them into a NumPy array. Finally, We concatenated the extracted features and split the array into sets as follows: 70% (train; 9,188), 20% (validation; 2,625), and 10% (test; 1,312). All sets were scaled to a value between 0 and 1 according to a standard normal distribution.

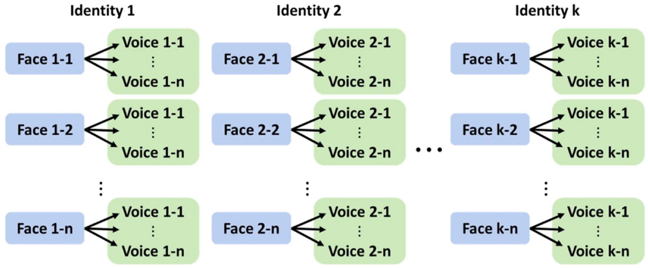

Fig. 2. Our generating virtual identities method utilizing separate face and voice datasets; k denotes an index of the specific identity.

C. Combination Dataset

Since we utilized face images and voice files from separate datasets, we aligned the face and voice data to create a combination dataset. In other words, we generated a virtual identity with one unique face and one unique voice. We carefully selected ten classes of facial and vocal data to ensure that the gender and age groups were similar, aiming to create virtual identities that resemble the same person as closely as possible. In the mapping procedure, the first class of the face (F_1) was matched with the first class of the voice (V_1), and the second class of the face (F_2) was matched with the second class of the voice (V_2). More specifically, the first image (F_{1-1}) of F_1 was combined with n pairs of all voice files $V_{1-1\ldots 1-n}$ of V_1. In addition, the second image F_{1-2} of F_1 was combined with n pairs of all voice files $V_{1-1\ldots 1-n}$ of V_1. In this way, we created ten virtual identities with unique face and voice data. Consequently, we obtained 229,557 training cases and 7,256 test cases for face data by applying this combination method. Furthermore, 1,377,342 training and 43,536 testing cases were created for the masked data (see Fig. 2).

D. Multi-Modal Dataset

We additionally employed a multi-modal dataset named Lip Reading Sentence 3 (LRS3) [66], which is a large-scale English sentence-level audio-visual dataset. The LRS3 dataset, originating from TED talks, is a compilation of videos paired with textual transcripts. To process the videos, we segmented them into individual frames to isolate images of the speaker's face. Our final dataset comprised 10,282 training cases from 3,995 speakers for the training set and 1,201 utterances from 409 speakers for the testing set. Furthermore, 61,692 training and 7,206 testing cases were created for the masked data.

E. Demographic Information

We used demographic information to complement the face modality. We extracted demographic information on the age, gender, and race of the face images using *lightface*[4] package [67]. *Lightface* is a face recognition and facial attribute analysis framework that extracts age as a continuous variable, gender as a binary class, and race as Asian, Black, Latino Hispanic, Middle Eastern, and White. After extracting demographic

[2] https://github.com/aqeelanwar/MaskTheFace
[3] http://www.openslr.org/12
[4] https://github.com/serengil/deepface

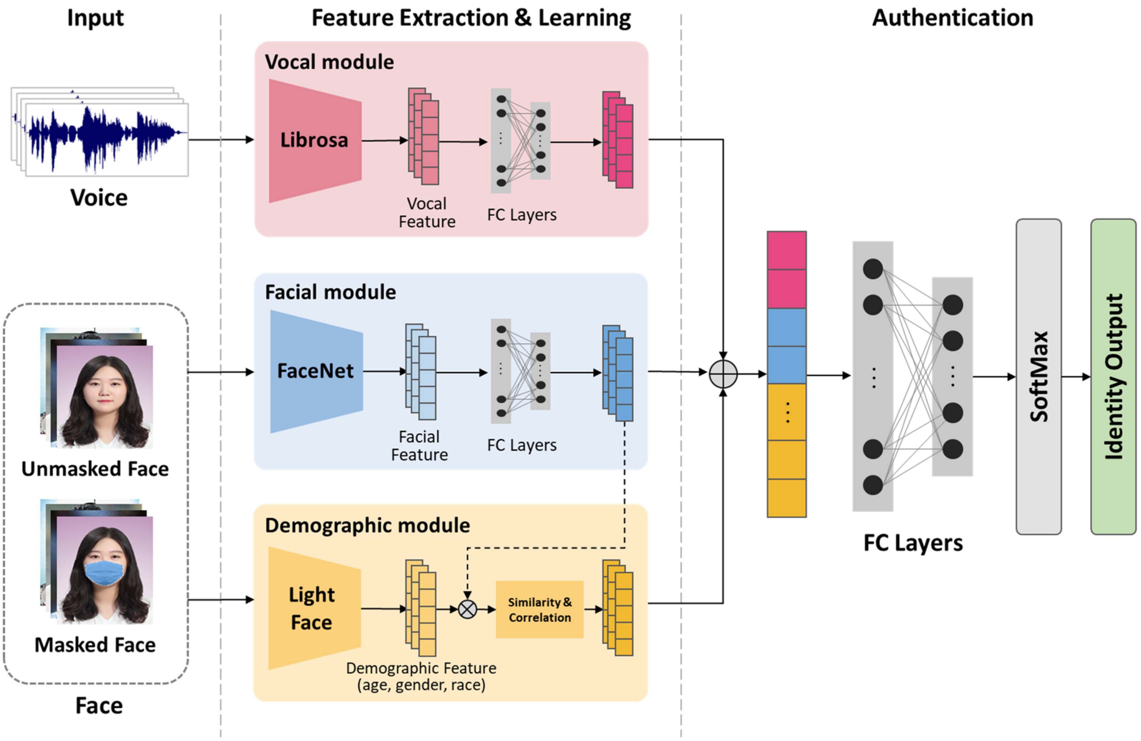

Fig. 3. Architecture of our proposed multi-modal authentication model, named *DemoID*, incorporates three distinct modules: a facial module, a vocal module, and a demographic module. The facial module analyzes features from both masked and unmasked facial images, the vocal module processes voice data for unique vocal characteristics, and the demographic module assesses demographic attributes like age, gender, and race. The combined output from these modules is then synthesized to produce a precise ID prediction.

TABLE I
SUMMARY OF EXPERIMENTAL DATASETS

Dataset	Data types	Train	Validation	Test
Combination	Face	2,978	742	744
	Masked Face	15,996	4,007	4,026
	Voice	9,188	2,625	1,312
	Face-Voice		229,557	7,256
	Masked Face-Voice		1,377,342	43,536
LRS3	Face-Voice		10,282	1,201
	Masked Face-Voice		61,692	7,206

information from face images, we compared the difference between the information extracted from the masked and unmasked face images. In the case of age, there was an average difference of 5.406 years between the face and masked face images. To address this issue, we categorized the age intervals: 10–19 were categorized as 15; 20–29 were categorized as 25; 30–39 were categorized as 35, and so on.

IV. OUR PROPOSED METHOD

In this section, we present the details of our proposed method and its modules.

A. DemoID: Multi-Modal Authentication Model

We introduce *DemoID*, a multi-modal authentication system that leverages facial, vocal, and demographic information to accurately identify individuals, even when faces are partially obscured. The system is structured into three dedicated modules: facial, vocal, and demographic. Each module is specialized in processing its respective modality and is designed to work in harmony with the others. Specifically, the demographic module is engineered to determine the interplay between demographic characteristics and facial features. During the identification process, *DemoID* processes each modality through its corresponding module. The extracted features–face f_i, voice v_i, and demographic d_i–are then aggregated into a unified feature vector $o_i = [v_i, f_i, d_i]$, where o_i belongs to the dimensional space $\mathbb{R}^d$ and $i = [1, 2, \ldots, M]$, where *M* means the number of the dataset in Table I. Subsequently, a fully connected layer, followed by a K-way softmax function, computes the probability distribution over potential class labels. The identity predicted by *DemoID* corresponds to the class with the highest probability, indicating the individual whose facial, vocal, and demographic characteristics most closely align with the input data. Fig. 3 shows a schematic representation of the *DemoID* architecture.

B. Vocal Module

In the vocal module, vocal feature extraction is performed using *Librosa* Python package, as outlined in Section III-B. From each audio sample, we extract five distinct types of vocal features, resulting in feature vectors represented as 193-dimensional NumPy arrays. For feature learning, we employ a deep neural network architecture, designed with multiple fully connected layers to process these features. Initially, the feature vectors are input into a dense layer of 193 units, ensuring a direct mapping from the raw input features. This is immediately followed by a dropout layer set at 0.1, introduced to mitigate overfitting by randomly deactivating a fraction of the neurons during training. Subsequently, the networks' flow advances through two additional dense layers, each consisting of 128 units. These layers are interspersed with dropout layers, with dropout rates increasing progressively to 0.25 and then to 0.5, further enhancing the networks' generalization capabilities.

C. Facial Module

To select the most effective model for facial feature extraction, we conducted a comparative analysis of two models (i.e., *DeepFace* [31] and *FaceNet* [68]). *DeepFace* and *FaceNet* are the most widely used models in the field of face recognition.

DeepFace is examined as follows [31]. We employed an image (Size: 152, 152, 3) as the input to the first convolutional layer (C1) with 32 filters of size 11×11×3. The resulting 32 feature maps are then fed to a max-pooling layer (M2) with a pooling size of 3 × 3 and a stride of 2 per channel, followed by another convolutional layer (C3) with 16 filters of size 9 × 9 × 16. These three layers extract low-level features, such as simple edges and textures. The subsequent layers are local-connected convolutional layers (L4, L5, and L6), which learn different sets of filters in every location of the feature map. This is based on the assumption that the areas between the eyes and eyebrows exhibit very different appearances and have much higher discrimination than those between the nose and mouth. The network is followed by two fully connected layers (F7, F8), with one drop-out layer between.

FaceNet [68] is designed to process facial images with an input size of 160×160×3. The core of this model is a deep Convolutional Neural Network (CNN), which undergoes a series of convolutions and pooling operations to extract high-level features from the input image. Following the feature extraction, the model applies L_2 normalization to the output of the deep CNN. This normalization step is crucial as it maps the facial images onto a compact Euclidean space, where distances directly correspond to a measure of facial similarity. The deep CNN is trained using Stochastic Gradient Descent (SGD) with standard back-propagation and AdaGrad. And, the triplet loss of *FaceNet* encourages the model to ensure that an anchor image of a person's face is closer to all other positive images of the same person than any negative images of different persons in the learned feature space.

After evaluating the performance of both models, we selected the one that yielded the best results for facial feature extraction, FaceNet. Subsequently, we employed feature learning to further refine the models' understanding of the extracted features. Similar to the vocal module, this feature learning process utilizes a deep neural network architecture based on fully connected layers.

D. Demographic Module

To assign weights according to the significance of the demographic information, we calculated the weighted sum from the demographic information and extracted image features, which were output through the final layer of the facial module. Fig. 4 shows the details of the demographic module. The specific procedures are summarized as follows. We encoded the categorical information (gender and race) into integer values using the LabelEncoder function of scikit-learn library and concatenated all the demographic information (age a_i, gender g_i, and race r_i). Demographic information w_i is indicated as $w_i = [a_i, g_i, r_i]$, with $w_i \in \mathbb{R}, i = 1, 2, \ldots, M$, where M represents the number of datasets in Table I. We then input into a dense layer with an embedding size of 128 and tanh activation. To measure the similarity between the face and demographic features, we calculated the correlation vector using an element-wise multiplication as follows:

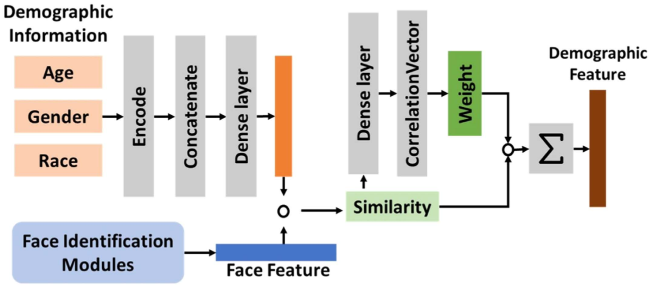

Fig. 4. Details of the suggested demographic module.

$$s_i = tanh(f \odot w_i), \tag{1}$$

where $\odot$ means element-wise multiplication, and s_i represents the correlation vector between the face feature and *i*-th demographic information. Based on the correlation vector s_i, we calculated the weights of each demographic feature as follows:

$$a^s = softmax(W_s^T s + b_s), \tag{2}$$

where $W_s \in \mathbb{R}^d$, $b_s \in \mathbb{R}^K$ is a vector containing the weights of demographic information. Finally, we calculated the weighted sum of each demographic feature by multiplying the similarity and weights through element-wise multiplication as follows:

$$d = \sum_{i=1}^{K} a_i^s s_i. \tag{3}$$

V. Experiments and Results

A. Implementation Details

For details of the facial module, the input image size is 160×160×3. We utilized two 2D convolutional layers and three locally connected 2D layers, all incorporating a *ReLU* activation function. The first Conv2D layer has 32 filters and 11 kernel sizes. The first Conv2D layer comprises 32 filters with a kernel size of 11, while the second Conv2D layer consists of 16 filters and a kernel size of 9. Then, all locally connected 2D layers feature 16 filters with kernel sizes of 9, 7, and 5 sequentially. For the vocal module, the input size is 1×193. The three fully connected layers in the vocal module have embedding sizes of 193, 128, and 128, with corresponding dropout rates of 0.1, 0.25, and 0.5. For the demographic module, the input size is 1×3, encompassing encoded values for age, gender, and race. During

TABLE II
ACCURACY WITH DIFFERENT INPUT DATA FOR *DEEPFACE* AND *FACENET*; THE RESULTS OF THE FACE AND MASKED FACE TEST DATA IN TABLE I

Modality	DeepFace (%)	FaceNet (%)
Face	68.28	85.69
Face, Demography	**73.25**	**90.01**
Masked Face	40.29	64.83
Masked Face, Demography	46.90	68.38
Face, Masked Face	64.43	80.85
Face, Masked Face, Demography	70.26	88.23

TABLE III
COMPARISON WITH EXISTING FACE MODELS ON TEST DATA OF FACE AND MASKED FACE

Model	Face (%)	Face & Masked Face (%)
CNN	56.78	56.28
PCA + SVM	59.17	55.14
DeepFace	68.28	64.43
FaceNet	85.69	80.85
FAN	84.53	82.38
DeepFace + Ours	73.25	70.26
FaceNet + Ours	**90.01**	**88.23**

feature concatenation, the demographic feature undergoes resizing to a 16-dimensional embedding size through a dense layer employing a *tanh* activation function. Lastly, we employed fully connected layers with an embedding size of 4096 and utilized *ReLU* activation. A softmax layer was employed with the number of identities as its output. The ultimate output of *DemoID* represents the identity of a specific person corresponding to the input face and voice. The identity is encoded as a one-hot list, indicating the unique number assigned to the person. In the training procedure, optimal hyper-parameters were chosen through grid search, a method that systematically evaluates model performance across a predefined range of hyper-parameter values to determine the best combination. All hyperparameters were fixed based on those yielding the best results. Specifically, a batch size of 64 and a momentum of 0.9 were selected. The learning rate was computed by multiplying 0.01 and dividing the batch size by the embedding sizes. This dynamic computation ensures an adaptive learning rate that scales with the chosen batch size, enhancing the optimization of the training process. The training parameters were optimized using the SGD optimizer and categorical cross-entropy.

B. Face Identification

To evaluate the efficacy of the facial module, we utilized *DeepFace* and *FaceNet* as feature extractors in separate configurations on the combination dataset. We also assess the contribution and effectiveness of the demographic module. Our evaluation encompassed various modality conditions to analyze the performance of each facial module setup:

1) Training and testing with face images;
2) Training and testing with face images and demographic information;
3) Training with face images and testing on masked images;
4) Training on face images with demographic information and testing on masked images;
5) Training and testing on face and masked images;
6) Training and testing on face and masked images with demographic information.

As demonstrated in Table II, we attained test set accuracies of 68.28% and 85.69% for the *DeepFace*-based model and *FaceNet*-based model, respectively. These results were achieved by training and evaluating both models using face images. Second, we achieved 73.25% and 90.01% test set accuracies for the *DeepFace* and *FaceNet* models, respectively, when integrated with our demographic module. These results were achieved by training and testing the models with face images and demographic information. Third, we achieved 40.29% and 64.83% accuracies on the test set by training the models with face images and testing them with masked face images, respectively. Fourth, we achieved 46.90% and 68.38% accuracies on the test set by training the models on face images with demographic information, respectively. Furthermore, we achieved 64.43% and 80.85% accuracies on the test set by training and testing on the face and masked face images, respectively. Finally, we achieved 70.26% and 88.23% accuracies on the test set by training and testing the models on the face and masked face images with demographic information. We obtained 3.55%–7.38% improvement on the face identification model using the demographic module. By retraining the masked images, we identified the masked images with an increase in performance of 16.02%–24.14% using the proposed model. These results show that demographic information can be one of the key components of human authentication. Moreover, the results indicate that using demographic information in the identification models can compensate for the lack of facial features.

Furthermore, we compared the performance of our model in the face identification task with the following models.

- **CNN** [69]: We followed the CNN for face recognition. Each image is transferred to 2D convolutional and max pooling layers of 152, 152, and 3 sizes. We employed categorical cross-entropy for the loss function.
- **PCA with SVM** [70]: We employed the principle component analysis and SVM for the face recognition problem.
- **DeepFace** [31]: *DeepFace* consists of convolutional layers, a max pooling layer, and locally connected convolutional layers. In this model, we used images of 152, 152, and 3 sizes.
- **FaceNet** [68]: *FaceNet* is one of the state-of-the-art face recognition methods with deep CNN. In this model, we used images of 160, 160, and 3 sizes.
- **FAN** [71]: *Face attention network* is an effective face detector for occluded faces based on the anchor-level attention algorithm.

Although the face recognition model using MTCNN and FaceNet [72] is the popular model, it cannot be employed as a baseline model because MTCNN does not detect face regions well in masked faces. Table III summarizes the accuracy of each model with face and masked face data. *FaceNet* based our model outperformed other models with 90.01% and 88.23% for the face

TABLE IV
RESULTS OF *DEMOID* WITH DIFFERENT INPUT MODALITIES

Modality	DemoID (%)
Face, Voice	95.45
Face, Voice, Demography	96.24
Face, Masked Face, Voice	94.86
Face, Masked Face, Voice, Demography	**99.00**

and (face+masked face) data, respectively, when integrated with our demographic module. The results show that the employed demographic module can consider the correlation between facial features and demographic information; thus, it is suitable for processing demographic information.

C. Multi-Modal Identification

To demonstrate the performance according to each modality and test the effectiveness of multi-modality, we conducted four experiments under the following conditions:

1) Face + Voice
2) Face + Voice + Demography
3) Face + Masked Face + Voice
4) Face + Masked Face + Voice + Demography

Table IV presents the results of our proposed model, *DemoID*. Our proposed *DemoID* achieved the greatest accuracy under the following condition: face+masked face+voice+demography (99.00%). This result shows that this modality combination can improve the performance of human authentication models. We also confirm that additional usage of demographic modality makes up for the performance degradation by wearing the mask (face+voice:95.45%, face+voice+demography: 96.24%, face+masked face+voice: 94.86%, face+masked face+voice+demography: 99.00%). Compared to other unimodal face identification models, our proposed model achieved an overall improvement of 5.41%–6.23% and 6.63%–10.77% for the face and masked face data, respectively. It indicates that multiple modalities can be useful for building a robust human authentication system.

In Table II, the addition of masked face data resulted in decreased accuracy, while in Table IV, the inclusion of such data led to increased accuracy, presenting seemingly contradictory results. One reason for these outcomes may lie in the diversity of the data and the complexity of the models. When masked face data is included, the model is faced with learning from a broader array of variations, which can lead to reduced performance in simpler models. However, when employing more complex models or integrating additional modalities such as voice data, the model can utilize this extra information to make more accurate predictions.

We compared the performance of *DemoID* with existing models in the multi-modal identification task. We selected the following multi-modal models using both face and voice modalities. Moreover, to ascertain the robustness and reliability of our model, we expanded our evaluation by incorporating the LRS3 dataset for multi-modal identification.

TABLE V
COMPARISON WITH EXISTING MULTI-MODAL MODELS ON TWO DATASETS

Model	Combination (%)	LRS3 (%)
CTC/Attention	84.12	80.55
AVSR + AVSE	90.41	91.63
MMST	94.88	93.73
DemoID (Ours)	**99.00**	**96.48**

TABLE VI
RESULTS OF DEMOID FOR REAL-WORLD DATASET

Modality	*DemoID* (%)
Face, Masked Face, Voice	90.60
Voice, Demography	97.69
Face, Masked Face, Demography	98.09
Face, Masked Face, Voice, Demography	**98.35**

- **CTC/Attention** [73]: A pioneering hybrid CTC/attention architecture, offering significant improvements in word error rate, especially under noisy conditions
- **AVSR+AVSE** [74]: An advanced audio-visual speech recognition (AVSR) system integrating an AVSR conformer-based model with hybrid decoding and an audio-visual speech enhancement (AVSE) U-net model with recurrent neural network (RNN) attention
- **MMST** [75]: Multi-modal sparse transformer network designed to enhance audio-visual speech recognition by utilizing a sparse self-attention mechanism

As shown in Table V, our method outperformed other identification models. Since the combination dataset was artificially constructed with ten virtual identities created by us, it generally presented better performance than the LRS3 dataset, which contains a larger number of class identities. Moreover, our model demonstrated superior multi-modal identification performance on both datasets, achieving the highest accuracies of 99.00% and 96.48%, respectively. These results substantiate the effectiveness of our proposed model's face-voice fusion approach and the utilization of demographic information in enhancing human identification. Lastly, the training and testing time of our model is about 7 hours 16 minutes and 1 minute 47 seconds, respectively.

D. Additional Study

1) Results of DemoID *on Real-World Dataset:* We conducted additional experiments with real-world datasets. We recruited 15 participants, collected their faces and voices, and extracted their demographic information. The face dataset consists of about 100 images for each participant, including face and masked face images. In our voice dataset, we collected speech files organized by a 5-minute book reading session for each participant. We split the collected dataset into 8:2, trained and tested *DemoID* in four conditions: face+masked face+voice, face+masked face+demography, voice+demography, face+masked face+voice+demography.

Table VI shows the results of *DemoID* with the real-world dataset. First, we achieved 90.60% accuracy on the test set by training and testing *DemoID* with face and voice. Second,

TABLE VII
RESULTS OF 3-FOLD VALIDATION ON COMBINATION DATASET

Dataset	DemoID (%)
Virtual identity 1	**99.00**
Virtual identity 2	98.68
Virtual identity 3	98.88

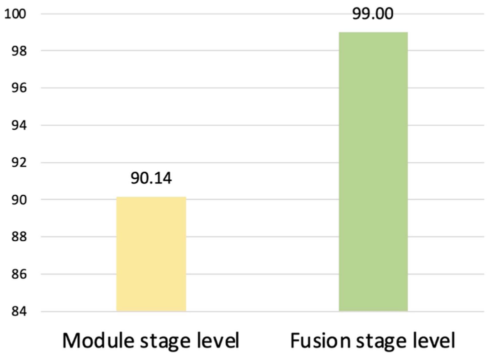

Fig. 5. Results of stage-level approaches.

we achieved 97.69% accuracy using voice and demographic information. Furthermore, we achieved 98.09% accuracy on the test set using face and demographic information. Finally, we achieved 98.35% accuracy for face, voice, and demographic information. These results demonstrate that *DemoID* is suitable for real-life applications. We also found that the proposed tree modalities can help improve the performance of user authentication.

2) Cross Validation on Combination Dataset: To enhance the robustness of our experimental findings and preclude the potential for coincidental outcomes in virtual identity experiments, we generated two additional datasets featuring virtual identities. Then, we subjected the *DemoID* model to a 3-fold validation process. Table VII shows the summary of the results with 3-fold validation approaches.

3) Stage Level Study: We compared the authentication result predicted at the fusion and module stage levels. The fusion stage level implies that the concatenated feature from each module is used to identify a person. In other words, the fusion stage level is the algorithm of the proposed model. Meanwhile, the module stage level is a voting method for someone predicted by each module. Fig. 5 summarizes the result of the stage-level study. The result shows that the fusion stage level performs better than the module stage level. We can confirm that a model fusion of modality is more optimal than the uni-modal method through stage-level study.

VI. DISCUSSION AND CONCLUSION

In this study, we have proposed a multi-modal authentication model for occluded faces in a challenging environment. The proposed model leverages facial, audio, and demographic information (i.e., age, gender, and race) for personal identification. We also have suggested a demographic module to efficiently utilize demographic information and consider the correlation between demographic information and facial features. The experimental results show that the face identification model with the demographic module performs better than the baseline models, indicating that demographic information is helpful for a human identification model. We also observe that the proposed multi-modal identification model outperforms the uni-modal face model in all modality combinations. This means that the multi-modal identification model can be more useful than uni-modal models in specific situations, such as the COVID-19 pandemic.

From an academic perspective, we contribute to the field of person identification, particularly for occluded faces and non-face-to-face human identification. Previous single-modality identification systems based on fingerprints or PINs require physical contact or depend on information that must be memorized, which can lead to usability and hygiene issues. Especially in challenging environments such as a pandemic, where physical contact between individuals is limited and wearing masks is mandatory, prior single-modality models are difficult to apply. Furthermore, in cases of low video quality or long distances, the extraction of facial features becomes challenging, potentially compromising the effectiveness of face recognition systems. Our proposed model can prevent performance degradation by using multi-modality in these challenging environments where specific modalities may be obscured or impaired. In addition, the experimental results of the proposed demographic network module have demonstrated the utility of demographic information in person identification, confirming the efficacy of the module. This finding indicates the groundwork for future expansions in utilizing demographic information across various domains. In addition, the proposed module can be utilized with other modalities (e.g., gait, voice).

From a practical perspective, our model can be further developed and employed in future authentication systems to help society cope with security vulnerabilities that can arise in challenging environments. Particularly, the experimental results on the wild dataset showed that the proposed model is immediately applicable to authentication systems in real-world settings, significantly enhancing their reliability and efficiency. Moreover, the multi-modality approach to person identification can be the solution to evolving security threats within the rapidly changing digital society.

Although this study holds notable implications and findings, our current work has several limitations. First, only a few datasets have been considered for the application of this study. Second, the proposed multi-modal authentication model may require longer computational time than other approaches. Finally, since this algorithm tested the closed set recognition method, it is difficult to apply the proposed model in real-time. Based on the findings and limitations of this study, we suggest future research to focus on expanding the dataset diversity to enhance the models' robustness across various scenarios and investing in computational optimizations to facilitate real-time processing and immediate applicability of the multi-modal authentication model in dynamic environments.

REFERENCES

[1] A. Kathed et al., “An enhanced 3-tier multimodal biometric authentication,” in *Proc. IEEE Int. Conf. Comput. Commun. Inform.*, 2019, pp. 1–6.

[2] R. Ryu, S. Yeom, S.-H. Kim, and D. Herbert, “Continuous multimodal biometric authentication schemes: A systematic review,” *IEEE Access*, vol. 9, pp. 34541–34557, 2021.

[3] C. I. Paules, H. D. Marston, and A. S. Fauci, “Coronavirus infections–more than just the common cold,” *JAMA*, vol. 323, no. 8, pp. 707–708, 2020.

[4] P. Kumari and K. Seeja, “A novel periocular biometrics solution for authentication duringCOVID-19,” *J. Ambient Intell. Humanized Comput.*, vol. 12, pp. 10321–10337, 2021.

[5] O. Tutsoy, K. Balikci, and N. F. Ozdil, “Unknown uncertainties in the COVID-19 pandemic: Multi-dimensional identification and mathematical modelling for the analysis and estimation of the casualties,” *Digit. Signal Process.*, vol. 114, 2021, Art. no. 103058.

[6] O. Tutsoy, “Pharmacological, non-pharmacological policies and mutation: An artificial intelligence based multi-dimensional policy making algorithm for controlling the casualties of the pandemic diseases,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9477–9488, Dec. 2022.

[7] O. Tutsoy, “COVID-19 epidemic and opening of the schools: Artificial intelligence-based long-term adaptive policy making to control the pandemic diseases,” *IEEE Access*, vol. 9, pp. 68461–68471, 2021.

[8] S. M. M. Islam, O. Borić-Lubecke, Y. Zheng, and V. M. Lubecke, “Radar-based non-contact continuous identity authentication,” *Remote Sens.*, vol. 12, no. 14, 2020, Art. no. 2279.

[9] L. Wang, W. Chen, N. Jing, Z. Chang, B. Li, and W. Liu, “AcoPalm: Acoustical palmprint-based noncontact identity authentication,” *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 9122–9131, Dec. 2022.

[10] M. L. Mohd Tajuddin, K.-Y. Chan, W.-L. Pang, and A. S.-T. Ng, “RFID based identity authentication system in COVID-19 era,” in *Proc. Int. Conf. Comput. Sci. Technol.*, 2022, pp. 383–396.

[11] M. B. Stegmann, “Convenience security: The maths of multi-modal authentication,” 2020. [Online]. Available: https://www.fingerprints.com/2020/06/11/convenience-security-the-maths-of-multi-modal-authentication/

[12] A. Acar, H. Aksu, A. S. Uluagac, and K. Akkaya, “WACA: Wearable-assisted continuous authentication,” in *Proc. IEEE Secur. Privacy Workshops*, 2018, pp. 264–269.

[13] B. A. El-Rahiem, F. E. A. El-Samie, and M. Amin, “Multimodal biometric authentication based on deep fusion of electrocardiogram (ECG) and finger vein,” *Multimedia Syst.*, vol. 24, pp. 1–13, 2021.

[14] S. S. Sengar, U. Hariharan, and K. Rajkumar, “Multimodal biometric authentication system using deep learning method,” in *Proc. Int. Conf. Emerg. Smart Comput. Inform.,* 2020, pp. 309–312.

[15] K. Su et al., “Human identification using finger vein and ECG signals,” *Neurocomputing*, vol. 332, pp. 111–118, 2019.

[16] A. Herbadji et al., “Combining multiple biometric traits using asymmetric aggregation operators for improved person recognition,” *Symmetry*, vol. 12, no. 3, 2020, Art. no. 444.

[17] S. Fadl, A. Megahed, Q. Han, and L. Qiong, “Frame duplication and shuffling forgery detection technique in surveillance videos based on temporal average and gray level co-occurrence matrix,” *Multimedia Tools Appl.*, vol. 79, pp. 17619–17643, 2020.

[18] A. A. Abd El-Latif, M. S. Hossain, and N. Wang, “Score level multibiometrics fusion approach for healthcare,” *Cluster Comput.*, vol. 22, no. 1, pp. 2425–2436, 2019.

[19] X. Wang, S. Garg, H. Lin, M. J. Piran, J. Hu, and M. S. Hossain, “Enabling secure authentication in industrial IoT with transfer learning empowered blockchain,” *IEEE Trans. Ind. Inform.*,, vol. 17, no. 11, pp. 7725–7733, Nov. 2021.

[20] S. Soviany and M. Jurian, “Multimodal biometric securing methods for informatic systems,” in *Proc. IEEE 34th Int. Spring Seminar Electron. Technol.*, 2011, pp. 447–450.

[21] H. Aronowitz et al., “Multi-modal biometrics for mobile authentication,” in *Proc. IEEE Int. Joint Conf. Biometrics*, 2014, pp. 1–8.

[22] M. Hammad and K. Wang, “Parallel score fusion of ECG and fingerprint for human authentication based on convolution neural network,” *Comput. Secur.*, vol. 81, pp. 107–122, 2019.

[23] X. Zhang, D. Cheng, P. Jia, Y. Dai, and X. Xu, “An efficient android-based multimodal biometric authentication system with face and voice,” *IEEE Access*, vol. 8, pp. 102757–102772, 2020.

[24] T. Joseph, S. Kalaiselvan, S. Aswathy, R. Radhakrishnan, and A. Shamna, “A multimodal biometric authentication scheme based on feature fusion for improving security in cloud environment,” *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 6, pp. 6141–6149, 2021.

[25] D. Valdes-Ramirez et al., “A review of fingerprint feature representations and their applications for latent fingerprint identification: Trends and evaluation,” *IEEE Access*, vol. 7, pp. 48484–48499, 2019.

[26] D.-J. Kim, K.-W. Chung, and K.-S. Hong, “Person authentication using face, teeth and voice modalities for mobile device security,” *IEEE Trans. Consum. Electron.*, vol. 56, no. 4, pp. 2678–2685, Nov. 2010.

[27] S. Thavalengal, P. Bigioi, and P. Corcoran, “Iris authentication in handheld devices-considerations for constraint-free acquisition,” *IEEE Trans. Consum. Electron.*, vol. 61, no. 2, pp. 245–253, May 2015.

[28] M. Wang and W. Deng, “Deep face recognition: A survey,” *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[29] P. Dou, S. K. Shah, and I. A. Kakadiaris, “End-to-end 3D face reconstruction with deep neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5908–5917.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[31] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1701–1708.

[32] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database forstudying face recognition in unconstrained environments,” in *Proc. Workshop Faces ‘Real-Life’ Images Detection Alignment Recognit.*, 2008, pp. 1–15.

[33] C. Liu and H. Wechsler, “Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition,” *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[34] B. F. Klare et al., “Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1931–1939.

[35] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, “Occlusion robust face recognition based on mask learning with pairwise differential siamese network,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 773–782.

[36] D. Zeng, R. Veldhuis, and L. Spreeuwers, “A survey of face recognition techniques under occlusion,” *IET Biometrics*, vol. 10, pp. 581–606, 2021.

[37] M. L. Ngan et al., “Ongoing face recognition vendor test (FRVT) part 6B: Face recognition accuracy with face masks using post-COVID-19 algorithms,” 2020. [Online]. Available: https://doi.org/10.6028/NIST.IR.8331

[38] Z. Wang, B. Huang, G. Wang, P. Yi, and K. Jiang, “Masked face recognition dataset and application,” *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 5, no. 2, pp. 298–304, Apr. 2023.

[39] A. Anwar and A. Raychowdhury, “Masked face recognition for secure authentication,” 2020, *arXiv:2008.11104*.

[40] I. Q. Mundial, M. S. U. Hassan, M. I. Tiwana, W. S. Qureshi, and E. Alanazi, “Towards facial recognition problem in COVID-19 pandemic,” in *Proc. IEEE 4rd Int. Conf. Elect. Telecommun. Comput. Eng.*, 2020, pp. 210–214.

[41] C. Li, S. Ge, D. Zhang, and J. Li, “Look through masks: Towards masked face recognition with de-occlusion distillation,” in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3016–3024.

[42] H. N. Vu, M. H. Nguyen, and C. Pham, “Masked face recognition with convolutional neural networks and local binary patterns,” *Appl. Intell.*, vol. 52, pp. 5497–5512, 2021.

[43] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.

[44] N. Damer, J. H. Grebe, C. Chen, F. Boutros, F. Kirchbuchner, and A. Kuijper, “The effect of wearing a mask on face recognition performance: An exploratory study,” in *Proc. Int. Conf. Biometrics Special Int. Group*, 2020, pp. 1–6.

[45] F. Boutros et al., “MFR 2021: Masked face recognition competition,” in *Proc. IEEE Int. Joint Conf. Biometrics*, 2021, pp. 1–10.

[46] J. Deng, J. Guo, X. An, Z. Zhu, and S. Zafeiriou, “Masked face recognition challenge: The insightface track report,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1437–1444.

[47] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 1695–1699.

[48] P. A. Thomas and K. P. Mathew, “A broad review on non-intrusive active user authentication in biometrics,” *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 1, pp. 339–360, 2023.

[49] Z. Meng, M. U. B. Altaf, and B.-H. F. Juang, “Active voice authentication,” *Digit. Signal Process.*, vol. 101, 2020, Art. no. 102672.

[50] J. Du, P. Liu, F. Soong, J.-L. Zhou, and R.-H. Wang, “Noisy speech recognition performance of discriminative HMMs,” in *Proc. 5th Int. Symp. Chin. Spoken Lang. Process.*, 2006, pp. 358–369.

[51] J. Ajmera, I. McCowan, and H. Bourlard, “Speech/music segmentation using entropy and dynamism features in a HMM classification framework,” *Speech Commun.*, vol. 40, no. 3, pp. 351–363, 2003.

[52] Z. Valsan et al., “Statistical and hybrid methods for speech recognition in Romanian,” *Int. J. Speech Technol.*, vol. 5, no. 3, pp. 259–268, 2002.

[53] S. Fine, J. Navratil, and R. A. Gopinath, “A hybrid GMM/SVM approach to speaker identification,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2001, pp. 417–420.

[54] D. Meuwly and A. Drygajlo, “Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM),” in *Proc. ODYSSEY Speaker Recognit. Workshop*, 2001, pp. 140–150.

[55] J. Yuan et al., “Speaker identification on the SCOTUS corpus,” *J. Acoustical Soc. Amer.*, vol. 123, no. 5, 2008, Art. no. 3878.

[56] A. Senior and I. Lopez-Moreno, “Improving DNN speaker independence with I-vector inputs,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 225–229.

[57] K. Aizat, O. Mohamed, M. Orken, A. Ainur, and B. Zhumazhanov, “Identification and authentication of user voice using DNN features and i-vector,” *Cogent Eng.*, vol. 7, 2020, Art. no. 1751557.

[58] F. Richardson, D. Reynolds, and N. Dehak, “A unified deep neural network for speaker and language recognition,” in *Proc. INTERSPEECH*, 2015, pp. 1146–1150.

[59] J. Shang, S. Chen, and J. Wu, “Defending against voice spoofing: A robust software-based liveness detection system,” in *Proc. IEEE 15th Int. Conf. Mobile Ad Hoc Sensor Syst.*, 2018, pp. 28–36.

[60] S. Chen et al., “You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones,” in *Proc. Int. Conf. Distrib. Comput. Syst.*, 2017, pp. 183–195.

[61] L. Lu et al., “LipPass: Lip reading-based user authentication on smartphones leveraging acoustic signals,” in *Proc. IEEE Conf. Comput. Commun.*, 2018, pp. 1466–1474.

[62] Q. Wang et al., “VoicePop: A pop noise based anti-spoofing system for voice authentication on smartphones,” in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 2062–2070.

[63] L. Zhang, S. Tan, and J. Yang, “Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 57–71.

[64] L. Zhang, S. Tan, J. Yang, and Y. Chen, “VoiceLive: A phoneme localization based liveness detection for voice authentication on smartphones,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 1080–1091.

[65] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” in *Proc. IEEE 13th Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 67–74.

[66] T. Afouras, J. S. Chung, and A. Zisserman, “LRS3-TED: A large-scale dataset for visual speech recognition,” 2018, *arXiv:1809.00496*.

[67] S. I. Serengil and A. Ozpinar, “Hyperextended lightface: A facial attribute analysis framework,” in *Proc. IEEE Int. Conf. Eng. Emerg. Technol.*, 2021, pp. 1–4.

[68] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[69] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, “Face recognition: A convolutional neural-network approach,” *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.

[70] M. O. Faruqe and M. A. M. Hasan, “Face recognition using PCA and SVM,” in *Proc. IEEE 3rd Int. Conf. Anti-Counterfeiting Secur. Identification Commun.*, 2009, pp. 97–101.

[71] X. Yin, Y. Tai, Y. Huang, and X. Liu, “FAN: Feature adaptation network for surveillance face recognition and normalization,” in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 301–319.

[72] R. Jin, H. Li, J. Pan, W. Ma, and J. Lin, “Face recognition based on MTCNN and FACENET,” 2021. [Online]. Available: https://url.kr/h5my64

[73] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, “Audio-visual speech recognition with a hybrid CTC/attention architecture,” in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 513–520.

[74] J.-W. Hwang, J. Park, R.-H. Park, and H.-M. Park, “Audio-visual speech recognition based on joint training with audio-visual speech enhancement for robust speech recognition,” *Appl. Acoust.*, vol. 211, 2023, Art. no. 109478.

[75] Q. Song, B. Sun, and S. Li, “Multimodal sparse transformer network for audio-visual speech recognition,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10028–10038, Dec. 2023.

Dahye Jeong received the B.Sc. degree from Kunsan National University, Gunsan-si, South Korea. She is currently working toward the Ph.D. degree with the Department of Applied Artificial Intelligence and Data Experience Lab, Sungkyunkwan University, Seoul, South Korea. Her research interests include machine learning in multimodal information and industrial data science.

Eunbeen Choi received the B.A. degree from Yonsei University, Seoul, South Korea. She is currently working toward the master’s degree with the Department of Interaction Science and Data Experience Lab, Sungkyunkwan University, Seoul. Her research interests include data psychology, user experience, and computational social science.

Hyeongjin Ahn received the M.S. degree from Sungshin Women’s University, Seoul, South Korea. She is currently working toward the Ph.D. degree with the Department of Applied Artificial Intelligence and Data Experience Lab, Sungkyunkwan University, Seoul. Her research interests include user experience, data analytics, and computational social science.

Ester Martinez-Martin received the Ph.D. degree from Jaume-I University, Castellon de la Plana, Spain. She is currently an Associate Professor with the University of Alicante, Alicante, Spain, and the Director Assistant with the University Institute for Computing Research. Her research focusses on lines related to the use of vision in robotic tasks, such as object detection and action recognition.

Eunil Park received the Ph.D. degree from the Korea Advanced Institute of Science Technology, Daejeon, South Korea. He is currently an Associate Professor with the Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, South Korea. He has been one of the interdisciplinary scientists in the research fields, and his research results have been published in numerous international social science journals, as well as in scientific journals. His research interests include data science, HCI, and user behavior. He was the inaugural recipient of the NRF-Elsevier Young Researcher Award in Korea (Interdisciplinary Studies).

Angel P. del Pobil received the Ph.D. degree from the University of Navarra, Pamplona, Spain. He is currently a Professor with Jaume I University (UJI), Castellon de la Plana, Spain, where he is the Founding Director of the UJI Robotic Intelligence Laboratory. He is co-Chair of the IEEE RAS Technical Committee on Performance Evaluation and Benchmarking of Robotic Systems, and a Member of the Governing Board of the Intelligent Autonomous Systems (IAS) Society and EURON (European Robotics Research Network of Excellence, 2001–2009). Prof. Pobil was co-organizer of some 50 workshops and tutorials at ICRA, IROS, RSS, ROMAN, IJCNN, and HRI. He has been the program or General Chair of international conferences, such as adaptive behavior (SAB 2014), artificial intelligence, and soft computing.