Journal of Ambient Intelligence and Humanized Computing (2023) 14:16011–16024
https://doi.org/10.1007/s12652-022-04380-0

ORIGINAL RESEARCH

A deep feature-level fusion model for masked face identity recommendation system

Tipajin Thaipisutikul[1] · Phonarnun Tatiyamaneekul[1] · Chih-Yang Lin[2] · Suppawong Tuarob[1]

Received: 1 August 2021 / Accepted: 30 July 2022 / Published online: 19 September 2022

Abstract

The widespread occurrences of airborne outbreaks (e.g., COVID-19) and pollution (e.g., PM2.5) have urged people in the affected regions to protect themselves by wearing face masks. In certain areas, wearing masks amidst such health-endangering times is even enforced by law. While most people wear masks to guard themselves against airborne substances, some exploit such excuses and use face masks to conceal their identity for criminal purposes such as shoplifting, robbery, drug transport, and assault. While automatic face recognition models have been proposed, most of these models aim to identify clear, unobstructed faces for authentication purposes and cannot effectively handle cases where masks cover most facial areas. To mitigate such a problem, this paper proposes a deep-learning-based feature-fusion framework, *FIREC*, that combines additional demographic-estimated features such as age, gender, and race into the underlying facial representation to compensate for the information lost due to mask obstruction. Given an image of a masked face, our system recommends a ranked list of potential identities of the person behind the mask. Empirical results show that the best configuration of our proposed framework can recognize bare faces and masked faces with the accuracy of 99.34% and 97.65% in terms of Hit@10, respectively. The proposed framework could greatly benefit high-recall facial identity recognition applications such as identifying potential suspects from CCTV or passers-by's cameras, especially during crisis times when people commonly cover their faces with protective masks.

Keywords COVID-19 · AI-driven technology · Deep learning · Face identity recommendation system · Masked face recognition

1 Introduction

Given the current situation of the COVID-19 pandemic worldwide, it is critical to facilitate contactless operations in all business units, particularly in contact-inexorable places such as airports, retail shops, working offices, hospitals, and department stores. Among multiple biometrics used for person recognition, face recognition systems have been lauded as a reliable and non-contact method of validating a person's identity since it provides solid discriminative features for recognition under unconstrained environments. As a result, the face recognition system has become an essential tool for enhancing the mechanical abilities of security and surveillance systems and thousands of other applications in our daily lives. In combating outbreaks such as the currently widespread COVID-19 infection, people have been required by law in many countries to wear face masks in public places to avoid the spread of transmissible diseases in certain areas. Such mask-wearing practices have become the *New Normal* that brings a huge challenge for researchers, especially in computer vision, since the performance and trust of contactless identity verification through traditional face recognition can be significantly hindered as a majority of facial area is covered by a mask. The lack of accurate identity recognition for mask-wearing people has imposed enormous challenges

✉ Suppawong Tuarob
suppawong.tua@mahidol.edu

Tipajin Thaipisutikul
tipajin.tha@mahidol.edu

Phonarnun Tatiyamaneekul
phonarnun.tai@student.mahidol.edu

Chih-Yang Lin
andrewlin@saturn.yzu.edu.tw

[1] Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom, Thailand

[2] Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan

Fig. 1 Example scenario of a security concern that behooves automatic recognition of masked faces

for authorities when monitoring a huge population for suspicious behaviors (e.g., shoplifting, robbery, and assaults). As shown in Fig. 1, such urgent security needs call for an automated masked face identity recommendation system capable of narrowing down possible identities of a mask-wearing person of interest.

However, masked face identity recommendation is problematic for various reasons, primarily from the following perspectives. Firstly, masks obscure many important features deemed crucial for identity recognition from faces, such as mouths and noses, which could degrade the performance of existing face recognition systems. Secondly, using conventional face representations with local descriptors to encode the masked face features such as local phase quantization (LPQ) (Ojansivu and Heikkilä 2008; Ahonen et al. 2008), local binary patterns (LBP) (Ahonen et al. 2006), binarized statistical image features (BSIF) (Chu et al. 2013), and dual-cross patterns (DCP) (Ding et al. 2016) is too shallow to discriminate the complex nonlinear facial appearance variations. Thirdly, although recent works have shifted towards utilizing deep-learning approaches such as convolutional neural networks (CNNs) (Taigman et al. 2014; Chen et al. 2018) to automatically extract important features that are robust to the nonlinear appearance variation of face images, the additional features, especially the person's demographic information included in the modern facial feature extraction tools are not widely researched in existing literature, limiting the expressiveness of the model performance. Lastly, most existing face identity recognition systems are designed as black-box models with limited explainability. Thus, face identity recommendation systems with an interpretable explanation of demographic information such as gender, age, and race could prove crucial to law-enforcement officers when narrowing down plausible suspects.

Motivated by the above challenges, in this study, we propose a Deep Feature-Level Fusion Model For Masked **F**ace **I**dentity **Rec**ommendation System (FIREC) that can be utilized by the authorities to facilitate mitigation, prevention, evaluation, and action planning of security protocols amidst the COVID-19 and future similar situations (e.g., new airborne pandemic or pollution) where mask-wearing is commonly enforced. The model architecture of *FIREC* is elaborately designed with four primary modules to optimize the performance of the masked face identity recommendation task. First, data management is included to augment the full frontal face images to masked face images. Second, the feature extraction module is designed to efficiently extract not only the face representation but also the complementary semantic features such as age, race, and gender. Third, a feature-level fusion module is integrated to fuse all features into the richer feature representation. Lastly, the recommendation module will return the top-N ranked list of person identities with corresponding features. Since *FIREC* incorporates the additional demographic features to compensate for the concealed features (i.e., mouth, nose) hidden by face masks, extensive experiments on one public dataset and our two custom datasets demonstrate the superiority of our proposed model over other baselines in all metrics. To the best of our knowledge, this is the first published approach in the masked face identity recommendation domain that provides quantitative results and the qualitative explanation based on personal demographic information to the end-user while maintaining high accuracy of roughly 95% in recognition rate.

We summarize the contributions of our study as follows:

- We investigate the effect of wearing face masks on the performance of the current face recognition system.
- We propose an optimal solution to improve the existing studies by incorporating estimated demographic information for masked face identity recommendations into the proposed model.
- We introduce additional semantic features such as gender, race, and age into a proposed feature-level fusion approach to enhance the performance of the mask face recommendation. These additional features are estimated from a pre-trained DeepFace model.
- We conduct extensive experiments on three databases: bare face, masked face, and mixed face datasets to evaluate the proposed method. A comparative analysis with the baseline methods has been made to validate the effec-

tiveness of the proposed model. The code is available in the GitHub repository.[1]

The remainder of this manuscript is organized as follows: Sect. 2 reviews the related works. The methodology and the proposed model architecture are presented in Sect. 3. Sections 4 and 5 show the experimental settings and results of our study. Section 6 concludes the paper.

2 Background and related work

The COVID-19 pandemic has resulted in a variety of real-world challenges that have compelled the attention of the scientific and research communities. Exemplifications of direct research challenges include investigations of trends and analysis of pertinent information (Gupta and Katarya 2021a, b; Katarya et al. 2021), as well as automated ways to identify COVID-19 patients (Kedia and Katarya 2021; Gupta et al. 2021; Dantas et al. 2021; Kusakunniran et al. 2021). However, the indirect problems that have arisen as a result of government initiatives in the fight against the epidemic have received little attention to date. This study focuses on the issues that have developed as a result of the face-mask policies that have been widely implemented throughout the world, which require individuals to wear surgical masks at all times while outside. Criminals may be encouraged to commit crimes as a result of such enforcement since they will have more opportunities to conceal their identities while conducting illicit activities. In order to solve this challenge, it is necessary to be able to identify individuals based on their facial images even while masks are worn. As a result, we suggest that, in addition to the usual face recognition task, we extract additional information from the face-masked photos of individuals and use them to predict their identities. The fact that these applications, particularly in the law enforcement domain, require high-recall results leads us to frame this problem as a recommendation task (Katarya and Saini 2022; Gupta and Katarya 2021c, d; Katarya and Arora 2020; Katarya et al. 2013), where potential identities are ranked and returned, rather than returning only one candidate as followed by the traditional face recognition protocol.

Literature on automatic face recognition from facial images is vast (Naveen and Sivakumar 2021; Li et al. 2020; Ahmed et al. 2020; VenkateswarLal et al. 2019). Therefore, we focused on the recent studies directly relevant to our research problem. The summary of existing studies related to our work is shown in Table 1.

2.1 Face image representation

There are two main categories for face representations: local descriptor-based representations and deep learning-based representations (Beveridge et al. 2015). The former category can be further divided into two groups: the handcrafted and the learning-based descriptors. Ahonen et al. (2006) encoded the relationship of neighboring pixels by employing the texture descriptor for face representation. The Dual-Cross Patterns (DCP) descriptor was proposed by Ding et al. (2016) to capture the high-order features of facial components. Later, other handcrafted local descriptors such as Local Phase Quantization (LPQ) (Ojansivu and Heikkilä 2008) and Binarised Statistical Image Features (BSIF) (Chu et al. 2013) were proposed. In contrast with the handcrafted descriptors, the learning-based descriptors utilize the encoding pattern by machine learning techniques. An example of deep learning-based representation methods includes the DeepFace model proposed by Taigman et al. (2014) for face recognition which can achieve high recognition accuracy on par with that of humans. Chen et al. (2018) employed CNN to boost the discriminative power in extracting face features and combined two losses of identification and verification for more efficient training. Dosaj et al. (2018) proposed a face recognition system with steerable pyramid transform (SPT) and local directional pattern (LDP) for e-health's secured login in the cloud domain.

2.2 Face recognition based on multiple features

Utilizing only a single face representation from the image input could restrict the expressiveness of model capability in the face recognition problem. Therefore, recent research has included multiple additional features to provide better discriminative power to the machine learning models. Ding et al. (2016) offered to include three holistic-level features and six component-level features extracted from face representation to enhance the face recognition performance. In their proposed model, Chen et al. (2018) concatenated 25 in-depth features with Principle Component Analysis (PCA). Sarangi et al. (2022) proposed a face recognition based on the multimodal biometric approach using ear and face descriptors as the main features. Besides features from the facial area, multiple features have also been used in various applications in multimedia and computer vision, e.g., visual tracking (Ma and Xiang 2015), image classification (Xu et al. 2014; Qi et al. 2009; Enkhbat et al. 2020; Chen et al. 2021), and customer behavior analysis (Yolcu et al. 2019).

Our proposed model *FIREC* is different from the existing methods in many aspects. First, we propose a solution to improve the current face recognition to perform masked face identity recommendations. Second, we extract both face

[1] https://github.com/Zenonist/FIREC.

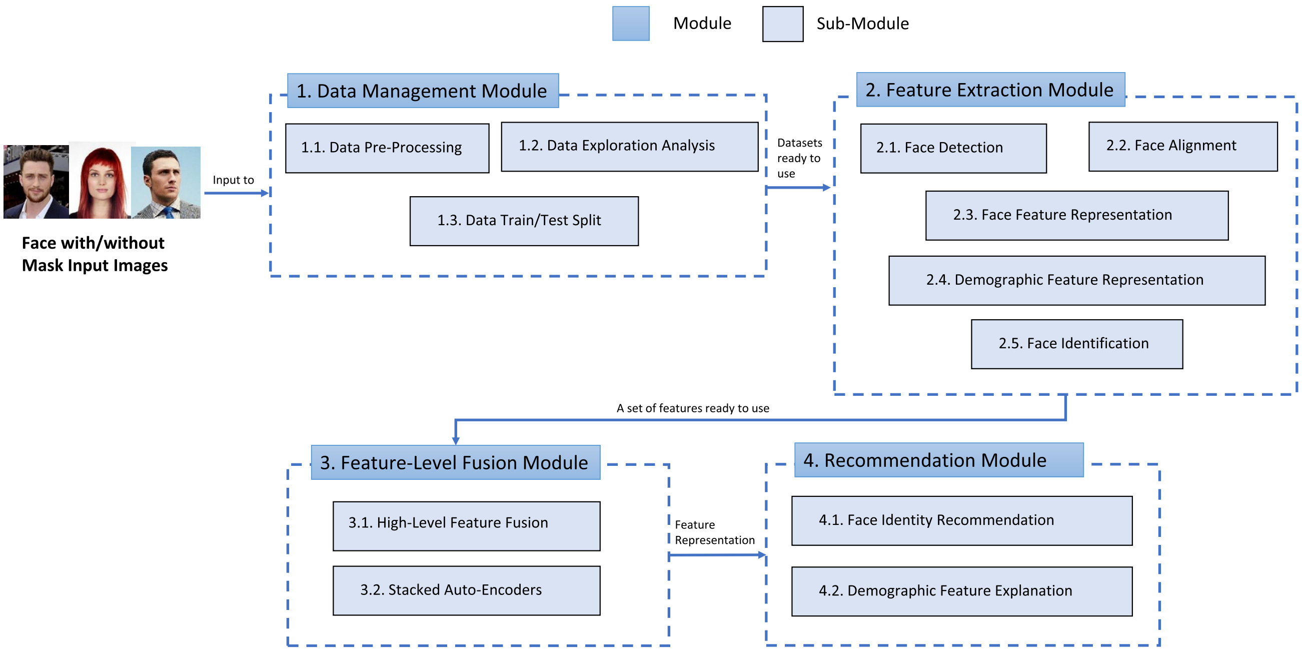
Module
Sub-Module
Face with/without Mask Input Images
Input to
1. Data Management Module
1.1. Data Pre-Processing
1.2. Data Exploration Analysis
1.3. Data Train/Test Split
Datasets ready to use
2. Feature Extraction Module
2.1. Face Detection
2.2. Face Alignment
2.3. Face Feature Representation
2.4. Demographic Feature Representation
2.5. Face Identification
A set of features ready to use
3. Feature-Level Fusion Module
3.1. High-Level Feature Fusion
3.2. Stacked Auto-Encoders
Feature Representation
4. Recommendation Module
4.1. Face Identity Recommendation
4.2. Demographic Feature Explanation

Fig. 2 Illustration of the proposed *FIREC* model pipeline consisting of four core modules: (1) data management module; (2) feature extraction module; (3) feature-level fusion module; and (4) recommendation module

Table 1 The summary of comparative analysis of existing studies

Paper	Techniques used	Advantages	Disadvantages	Dataset
Guo (2017)	LBP and KNN	Robust	Illumination conditions	LFW
Bonnen et al. (2013)	MRF, MLBP, and cosine similarity	Robust	Landmark extraction	AR (scream)
Karaaba et al. (2015)	HOG and MMD	Aligning difficulties	Low recognition accuracy	FERET
Arigbabu et al. (2015)	PHOG and SVM	Head pose variation	Complexity	LFW
Zhu et al. (2007)	PCA-FCF and correlation filter	Occlusion-insensitive	Linear method	CMU-PIE
Simonyan et al. (2013)	Fisher + SIFT and Mahalanobis matrix	Robust	Single feature type	LFW
Jose et al. (2012)	SPCA-KNN	Expression variation	Processing time	ESSEX
Sun et al. (2018)	CNN-LSTM	Automatically learn feature representations	Processing time	OPPORTUNITY

and other semantic features and propose the feature-level fusion module to fuse these features into a compact representation. Last but not least, we not only can deliver the quantitative recommendation list of prospective identities to a given masked face but also can provide an additional explanation on why our framework recommends these face identities as our final results.

3 Methodology

This section initially presents an overview of the pipeline used in our proposed framework: A Deep Feature-Level Fusion Model for Masked Face Identity Recommendation System (FIREC), as shown in Fig. 2. The *FIREC* model comprises four core modules: (1) **Data Management Module** to preprocess and augment the original images into the proper format for further processing. The input of this module is the set of facial images from publicly accessible online datasets. We perform the image augmentation procedure to mask the original frontal faces. As a result, we obtain three datasets with identity labels, including bare face, masked face, and mixed face datasets. Then, we perform a train/test split for each dataset to prepare the samples for training and testing the proposed model. (2) **Feature Extraction Module** to extract the essential features from the given input face images. There are five sub-modules to achieve this task: face detection, face alignment, face feature representation, demographic feature representation, and face identification. (3) **Feature-Level Fusion Module** fuses multiple features

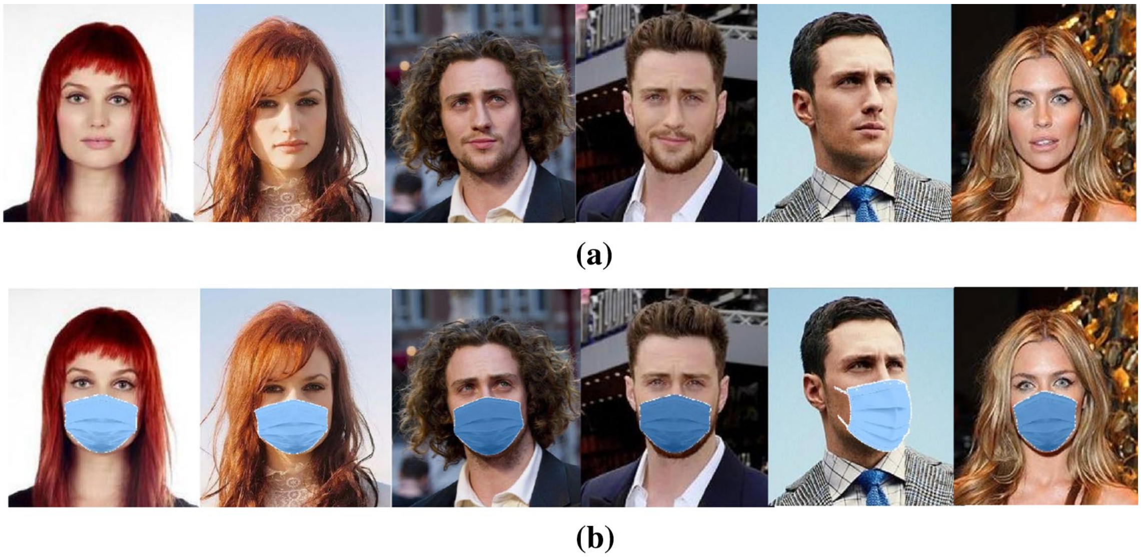

Fig. 3 Examples of input and output images of the Data Pre-processing sub-module. **a** Represents the original input images, denoted as $D_{bareface}$. **b** Represents the output images after performing the face-mask augmentation, denoted as $D_{maskedface}$

learned from the previous module into a higher-order feature representation. Then, the stacked auto-encoders process the fused feature to map the features into the correct person identity. (4) **Recommendation Module** to return the top-N identity list with demographic feature explanation based on the given input images, which could help the end-user to make the final decision. In contrast with traditional face recognition systems targeting authentication-related applications where the system returns only one predicted identity to a given face image, we frame our problem into a face identity recommendation problem, where the system returns a list of identities based on the likelihood of matching. Such a recommendation protocol would be more beneficial in the situation that focuses on high recall, such as narrowing down the suspects of a shoplifter, so law-enforcement officers could efficiently use their judgment to further pinpoint and investigate each potential suspect. We elaborate on the details of each module below.

3.1 Data management module

This section explains how we generate the masked face images from the original full-face images, our dataset characteristics, and split the dataset for training and testing to validate our proposed models.

3.1.1 Data-preprocessing sub-module

In this sub-module, we perform the face-mask augmentation on the original full frontal face images ($D_{bareface}$) to obtain the masked face images ($D_{maskedface}$). We utilize MaskTheFace (Anwar and Raychowdhury 2020), an open-source Python script to generate a mask on a target image. Example images before and after applying the MaskTheFace algorithm are shown in Fig. 3 using sample pictures from the public VGGFace[2] dataset. In this study, we create another dataset containing both bare and masked faces for further experiments and name it as $D_{mixedface}$.

3.1.2 Data exploration analysis sub-module

In this sub-module, we visualize the semantic information and summarize the insights from the overall features used in our experiments. These features are extracted from the feature extraction module. As shown in Fig. 4, we observe that the number of male images is higher than female samples. Also, most of our samples are in the age range from 20 to 40. The white race contributes about 62% out of all samples while Indian, Asian, Black, Middle Eastern, and Latino Hispanic are distributed around 6–10%. From these observations, we believe that our datasets contain various demographic features that can be used to investigate the impact of fusing these features to enhance face identification performance.

3.1.3 Data train/test split sub-module

After obtaining three datasets: $D_{bareface}$, $D_{maskedface}$, and $D_{mixedface}$ from the data-preprocessing sub-module, we further divide these datasets into 80% training and 20% testing

[2] http://www.robots.ox.ac.uk/~vgg/data/vgg_face/.

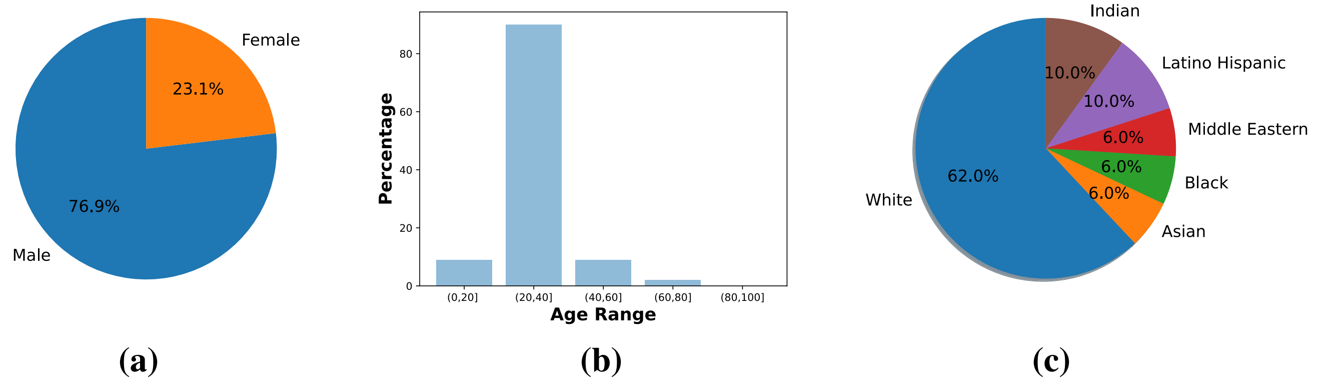

Fig. 4 Visualization of overall demographic features estimated by the DeepFace model used in our experiments. **a** Represents the gender distribution, **b** represents the age distribution and **c** represents race distribution from all samples

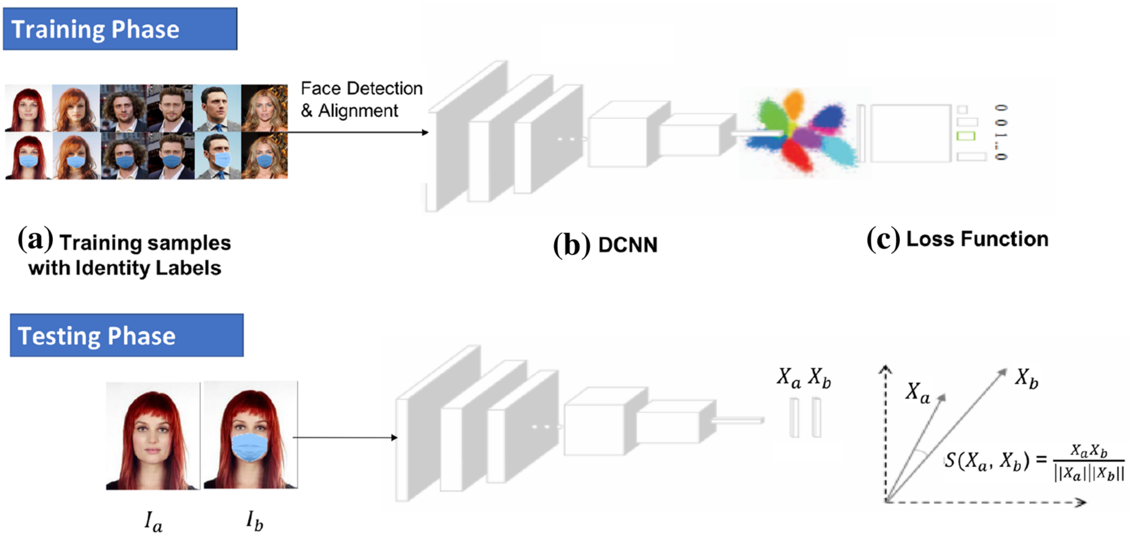

Fig. 5 Feature extraction pipeline based on DeepFace pre-trained model architecture. In the training phase, the approach executes a variant of face preprocessing (detect and align) and trains a deep convolutional neural network (DCNN) to categorize a pool of subjects. **a** Shows the input images with labels. **b** Shows DCNN network of DeepFace to map image intensity into a feature. **c** Shows the softmax loss function to discriminate between subjects. In the testing phase, the network is used as a feature extractor to obtain matched descriptors to perform masked face recognition

datasets. The output of this sub-module is therefore $D^{train}_{bareface}$, $D^{train}_{maskedface}$, $D^{train}_{mixedface}$, $D^{test}_{bareface}$, and $D^{test}_{maskedface}$.

3.2 Feature extraction module

The backbone of the face feature extraction module is developed based on the state-of-the-art deep learning suite, LightFace, a hybrid deep face recognition framework (Serengil and Ozpinar 2020). To compute facial representation, we deploy DeepFace (Taigman et al. 2014), a method based on deep convolutional neural network (DCNN) that was developed by Facebook AI Research with improved accuracy and speed. DeepFace is selected as our transfer pre-trained learning model since it provides rich information about the demographic features of each identity, which have not been utilized in the previous studies. Note that the demographic attributes for each picture are predicted by the pre-trained DeepFace model rather than from actual demographic information since the proposed framework is intended to be generalized even in the absence of such personal information. This motivated us to investigate the impact of a deep feature-level fusion approach on face identity performance. We re-train DeepFace with three different datasets ($D_{bareface}$, $D_{maskedface}$, and $D_{mixedface}$) to obtain the optimal model weights. As a result, we could obtain our three custom feature extraction models denoted as $M_{bareface}$, $M_{maskedface}$, and $M_{mixedface}$. As shown in Fig. 5, a standard pipeline proceeds

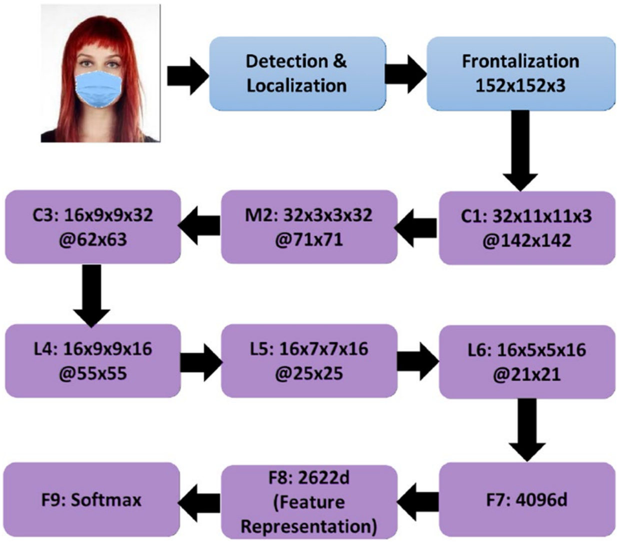

Fig. 6 The DCNN model architecture based on DeepFace used in this study

in learning face representations, starting from defining a DCNN architecture and a loss function for classification. Then, the DCNN is trained on a closed pool of subjects and used as a descriptor extractor on unseen faces. A typical testing pipeline boils down to exploiting the activations in the layer prior to the classification layer as a descriptor to encode the input. These descriptors are then compared and ranked using cosine similarity with the input test sample. The top-ranked identities are then returned as a recommendation. Further detail about the DCNN architecture used in this study is explained below.

The detail of Fig. 5b is elaborated in Fig. 6. As shown Fig. 6, we re-train the DCNN network with our datasets to classify the face images. Firstly, a 3-channels RGB face image of size `152x152` pixels is fed to a convolutional layer (C1) with 32 filters of size `11x11x3`. We symbolize this as `32x11x11x3@152x152`. Afterward, the output of C1 is given to a max-pooling layer (M2), which pools the max value over `3x3` spatial neighborhoods of stride 2 for each channel. Then, the output of C2 is given to C3, which has 16 filters of size `9x9x16`. This is followed by the subsequent layers (L4, L5, and L6), which execute similar operations but are different in the size of input pixels and filters used. The purpose of these six layers is to extract low-level features, such as simple edges and texture. Finally, the top three layers (F7, F8, and F9) are fully connected layers to connect all the inputs. These layers could capture correlations among features from different parts of the face images. The last fully-connected layer (F9) is a SoftMax layer that produces a distribution over K identity labels, where K is the number of identities. Given an input image I_t, the probability assigned to the k-th class denoted as o_k is the output of the SoftMax function. In other words, the detail of the loss function shown in Fig. 5c is shown as follows:

$$p_k = \frac{exp(o_k)}{\sum_h exp(o_h)}. \tag{1}$$

In the testing phase, the last layer (F9) is removed, and the output of the fully-connected layer (F8) in the network will be used as our raw face representation feature vector throughout this paper. Given two images I_a and I_b as inputs, the well-trained model now could compare the face representation X_a and X_b in cosine similarity space, computed as:

$$S(X_a, X_b) = \frac{X_a \cdot X_b}{\|X_a\| \cdot \|X_b\|}. \tag{2}$$

We calculate the cosine similarity scores of the input image against all candidates in our corpus and then rank the candidate identities based on the cosine similarity scores. A candidate image with a higher cosine similarity score is likely to belong to the same identity as the test image. The first top-N identities are then returned as a recommendation list. Besides the face feature representation, LightFace also predicts demographic features such as age, race, and gender from a given face image. The age feature is represented with the floating scalar value, while race and gender are represented as probability mixtures of [Indian, Asian, black, Middle Eastern, Latino Hispanic] and [male, female], respectively. In summary, the output of this module is a well-trained feature extraction models: $M_{bareface}$, $M_{maskedface}$, and $M_{mixedface}$.

3.3 Feature-level fusion module

We symbolize the features extracted by the feature extraction module as f_{face}, f_{age}, f_{gender}, and f_{race} for face embedding, age, gender, and race, respectively. These features represent semantic information for face identity recommendations. We conduct feature-level fusion to obtain a distinct signature for each face image by concatenating all the features as:

$$\hat{f} = [f_{face}:f_{age}:f_{gender}:f_{race}] \tag{3}$$

f is a high dimensional vector, which is impractical for real-world face recognition applications. Therefore, we further propose to lessen the dimension of f in non-linear feature transformations approach by stacked auto-encoders (SA). In this study, we employ a four-layer SA. The numbers of neurons of the four auto-encoders are 2048, 1024, 256, and 128, respectively. The last encoder's output is used as the compact signature of the face image before further processing in the Recommendation module. The structure of the designed SA is illustrated in Fig. 7.

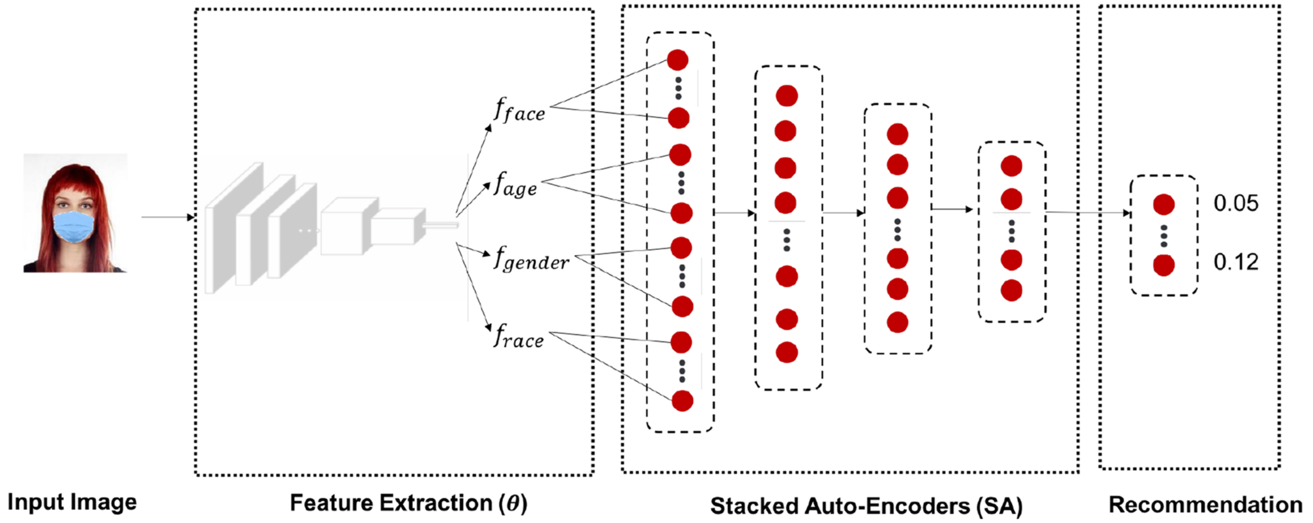

Fig. 7 Model architecture of *FIREC* consisting of feature extraction using DeepFace as a backbone network and stacked auto-encoders (SA) to fuse multiple feature-level into higher-level feature representation followed by Recommendation process

3.4 Recommendation module

In this last module, the final output is a dense layer whose size equals the total number of face identities (k) followed by a SoftMax function which outputs the probability of each identity, given the face and demographic feature interactions as:

$$P(o_k \in I_k | f_{face}; f_{age}; f_{gender}; f_{race}) \tag{4}$$

where o_k is the output probability of being identity k and I_k is a ground truth.

4 Experimental setting

This section first introduces the research questions we aim to answer. Then, we provide more details on the datasets used for evaluating the recommendation performance of our proposed *FIREC* model. Furthermore, we summarize all the baselines and evaluation metrics we use for the empirical analysis. Lastly, the implementation details are given.

4.1 Research questions

RQ1. How does wearing a mask on a face affect face recognition's performance? **RQ2.** Does the proposed model *FIREC* outperform other baselines? How can the proposed feature-level fusion module enhance the face identity recommendation's performance? **RQ3.** Can our proposed model *FIREC* provide an intuitive explanation for decision-makers in pinpointing the actual identity from the list of recommended identities?

4.1.1 Datasets

We use VGGFace[3] dataset published by the Oxford group (Chang et al. 2020). 50 person identities are randomly sampled from this dataset for evaluating our *FIREC* model. On average, each identity has around 166 distinct images, ranging from 80–200 images per identity. After performing the data-preprocessing, we obtain three datasets: Bare Face only dataset ($D_{bareface}$), Masked Face only dataset ($D_{maskedface}$), and mixed face dataset ($D_{mixedface}$). We split the dataset into 80% and 20% portions for training and testing, respectively. The summary of datasets used in our experiments is provided in Table 2.

4.2 Baseline

Several baselines on face identity recommendation are used for comparison as follow.

- $M_{bareface}$: our feature extraction model trained with $D_{bareface}$

Table 2 Statistics of the datasets used in our experiments

Dataset	Total	Train	Test
Bare face ($D_{bareface}$)	10,419	8296	2123
Masked face ($D_{maskedface}$)	10,262	8175	2087
Mixed face ($D_{mixedface}$)	20,681	16,471	4210

[3] http://www.robots.ox.ac.uk/~vgg/data/vgg_face.

- $M_{maskedface}$: our feature extraction model trained with $D_{maskedface}$
- $M_{mixedface}$: our feature extraction model trained with $D_{mixedface}$.

Our proposed *FIREC* model:

- $M^{age}_{mixedface}$: our feature extraction model trained with $D_{mixedface}$ and fused with age feature
- $M^{gender}_{mixedface}$: our feature extraction model trained with $D_{mixedface}$ and fused with gender feature
- $M^{race}_{mixedface}$: our feature extraction model trained with $D_{mixedface}$ and fused with race feature
- $M^{(age+gender)}_{mixedface}$: our feature extraction model trained with $D_{mixedface}$ and fused with age and gender features
- $M^{(age+race)}_{mixedface}$: our feature extraction model trained with $D_{mixedface}$ and fused with age and race features
- $M^{(gender+race)}_{mixedface}$: our feature extraction model trained with $D_{mixedface}$ and fused with gender and race features
- $M^{all}_{mixedface}$: our feature extraction model trained with $D_{mixedface}$ and fused with age, gender, and race features.

4.3 Evaluation metrics

To evaluate the performance of each method for the Masked Face Identity Recommendation System, we adopt four evaluation metrics and one statistical test named Wilcoxon signed-rank test, commonly used to evaluate recommendation systems as follows.

Hit Rate (H@N) quantifies the coverage of the recommendation in the recommended results. For each instance in the testing set, H@N is 1 if the ground truth identity appears in the set of top-N recommended identities and 0 otherwise. The overall H@N is computed as the average value of all testing instances. Note that H@1 is equivalent to the accuracy measure used for validating traditional face recognition models where only the top result is taken into account (Kar et al 2020; Taigman et al. 2014).

Precision (P@N) measures the ability of a model to predict the correct identity in the top-N recommended results.

$$P@N = \frac{1}{|N|}\sum_{i\in N}\frac{|Recom_{set_i} \in Test_{set_i}|}{|Recom_{set_i}|} \tag{5}$$

where $Test_{set_i}$ denotes the set in the testing samples, and $Recom_{set_i}$ denotes the set of recommended identities.

Mean Reciprocal Rank (MRR) evaluates the ranking capability of the recommendation. MRR is the average of multiplicative inverse of ranks of the first correctly predicted identity, defined as:

$$MRR = \frac{1}{|N|}\sum_{i\in N}\frac{1}{rank_i} \tag{6}$$

where $rank_i$ refers to the rank position of the first correctly predicted identity in the recommendation list.

Mean Rank (MR) is the arithmetic average of the ranks of the correctly predicted identities in the lists, calculated as:

$$MR = \frac{1}{|N|}\sum_{i\in N} rank_i \tag{7}$$

where $rank_i$ refers to the rank position of the first correctly predicted identity in the recommendation list.

Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used in this study to compare two recommendation ranked lists to determine whether they are statistically significantly different. The confidence level (alpha) is 0.05, below which we can reject the null hypothesis H_0, meaning that our proposed model is substantially different from the baseline.

In this paper, we choose $N = 1, \ldots, 10$ to illustrate the model's performance at different numbers of top results (*N*). The selection of N's values also corresponds to the real-world applications where a law-enforcement officer is presented with roughly ten plausible short-lists of suspects to investigate.

4.4 Implementation

For the model architecture of the feature extraction module, we follow the experiments and hyper-parameters settings as proposed in the original papers. All experiments were conducted on a Linux machine with an RTX 2080 Ti GPU and 64 GB of RAM. Our program is implemented with both Keras[4] and Tensorflow.[5] We use grid-search on the validation set to fine-tune our hyper-parameters of the models. We use Early Stopping with Adam optimization to obtain the best model for each dataset. Finally, we use 100, 100, and 100 epochs for $D_{bareface}$, $D_{maskedface}$, and $D_{mixedface}$ datasets, respectively.

5 Experimental results

In this section, experiment results are reported and discussed to answer the aforementioned research questions.

[4] https://keras.io/.

[5] https://www.tensorflow.org/.

5.1 Performance comparison (RQ1)

Table 3 presents the performance comparison among our proposed three models on two testing datasets. We have included some state-of-the-art methods for comparison on both datasets.

We make the following observations. First, it is noticeable that the traditional face recognition system trained with bare face images is not robust enough to yield acceptable accuracy on the masked face testing dataset. The sharp decline of H@1 performance from 95.52 to 55.53% (41.87% declination) indicates that the traditional face recognition trained only with bare face images is no longer efficient to deploy in the collaborative environment.

Second, when the model is trained with only masked face images, the H@1 performance on the testing masked-face dataset shows a significant improvement from 55.53 to 85.67% (54.28% improvement). However, the performance on the bare face testing dataset drops from 95.52 to 72.54% (24.06% declination). This indicates the limitation of this model on the generalization ability.

Third, we can observe the comparable H@1 performance to the best case when training the model with the mixed face images. Thus, we employ this model as our base model for the next section to investigate the impact of feature-level fusion on the performance gain. Therefore, we can conclude that wearing masks affects face recognition performance, and it is apparent on all investigated datasets. Also, the most optimal and generalized model from this experiment is the model trained with mixed-face images since it can handle both bare and masked faces with acceptable accuracy.

Finally, on the bare-face dataset, our proposed technique, which was trained on both the bare-face and mixed-face datasets, outperforms the existing state-of-the-art methods. Concerning the masked-face dataset, our proposed models trained on the masked-face and mixed-face datasets show a significant improvement of approximately 10% over the existing models trained on the masked-face dataset.

Table 3 H@1 (%) performance comparison of three models on two testing datasets

Model	$D^{test}_{bareface}$	$D^{test}_{maskedface}$
$M_{bareface}$	95.52	55.53
$M_{maskedface}$	72.54	85.67
$M_{mixedface}$	94.49	85.34
Vera-Rodriguez et al. (2019)	94.00	–
Vera-Rodriguez et al. (2019) with gender info.	94.40	–
Aswal et al. (2020)	–	75.70

5.2 Impact of the feature-level fusion module (RQ2)

As shown in Fig. 8, a significant improvement in terms of H@N is observed when we fuse additional semantic features into the learning model. Specifically, the best configuration of our proposed model yields 99.34% and 97.65% in terms of Hit@10. Overall, this indicates that our model can capture the correct identities within the top ten ranked results roughly 97–99 out of 100 samples (both bare and masked faces). Such a high-recall performance is vital to security applications.

Figures 9 and 10 show the performance of P@1, P@5, and P@10 on both testing datasets. Overall, we observe an upward trend when integrating more features into the proposed model. Interestingly, we notice that using all features does not make a substantial improvement as compared to using two features such as age, race, age, gender, and race, gender. However, with these observations, we still can conclude that a feature-level fusion module can enhance the face identity recommendation performance.

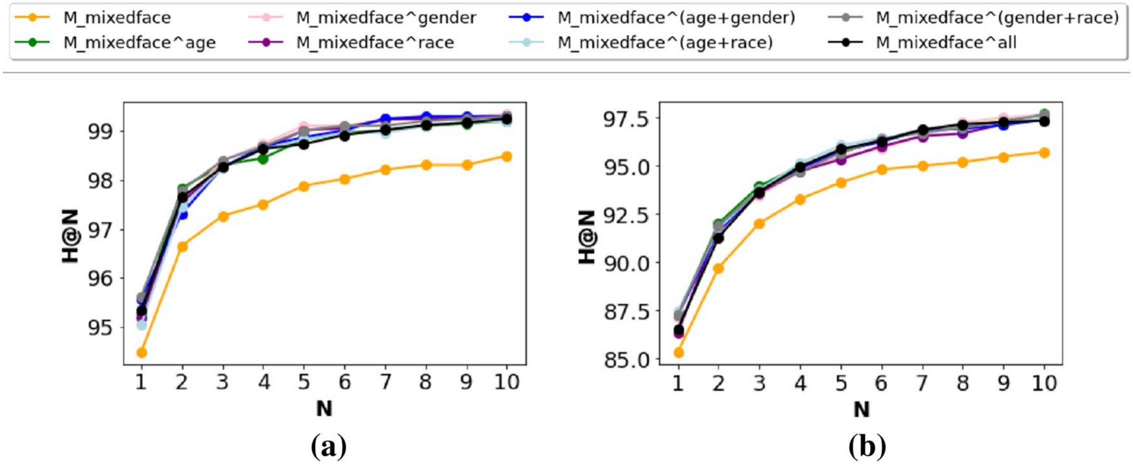

Fig. 8 Model performance comparison in terms of H@N on **a** bare face dataset and **b** masked face dataset

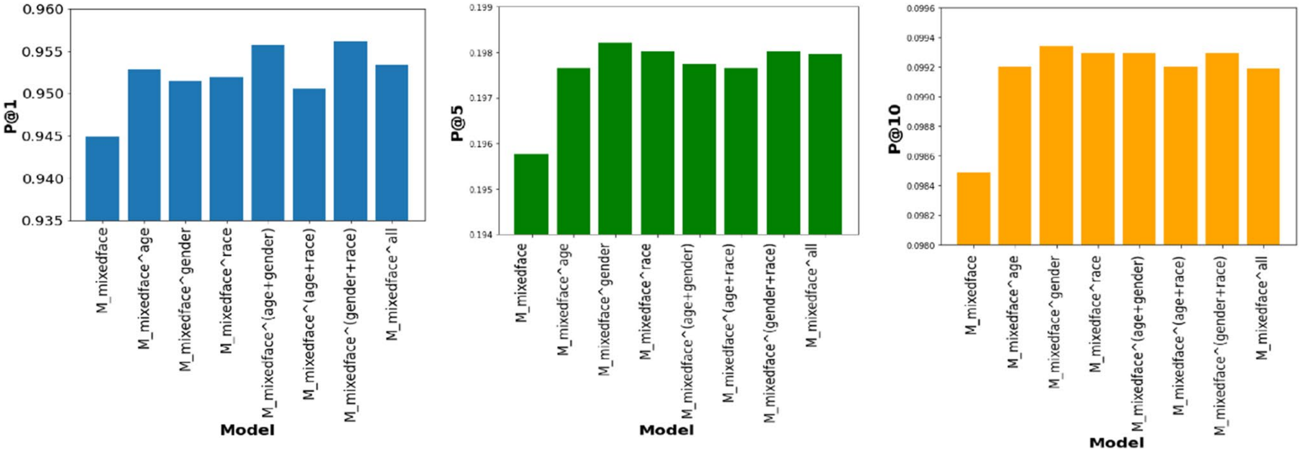

Fig. 9 The comparison of P@1, P@5, and P@10 of our proposed model variants on a bare face dataset

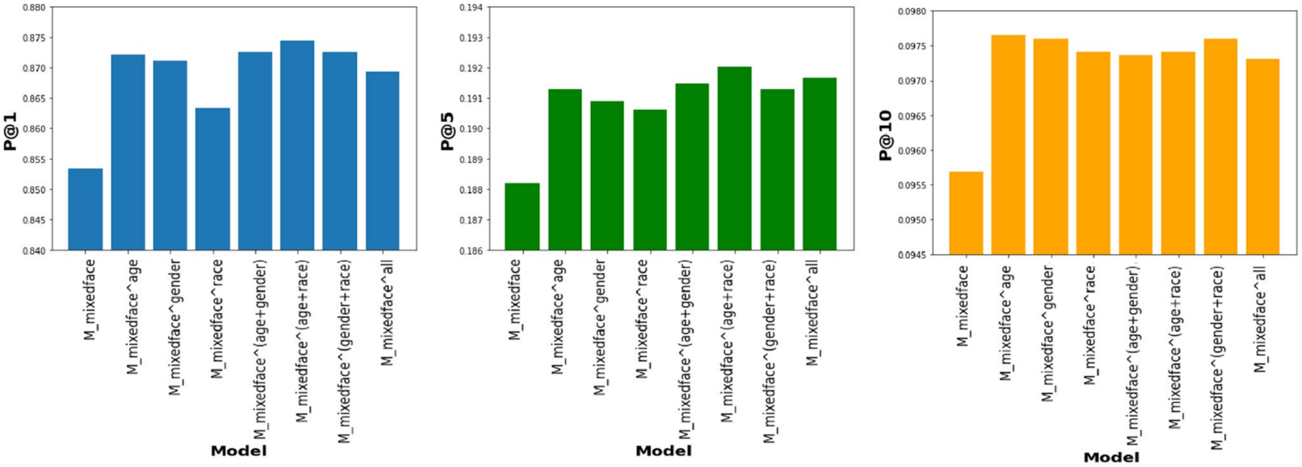

Fig. 10 The comparison of P@1, P@5, and P@10 of our proposed model variants on a masked face dataset

Fig. 11 The MRR performance comparison of our proposed model variants

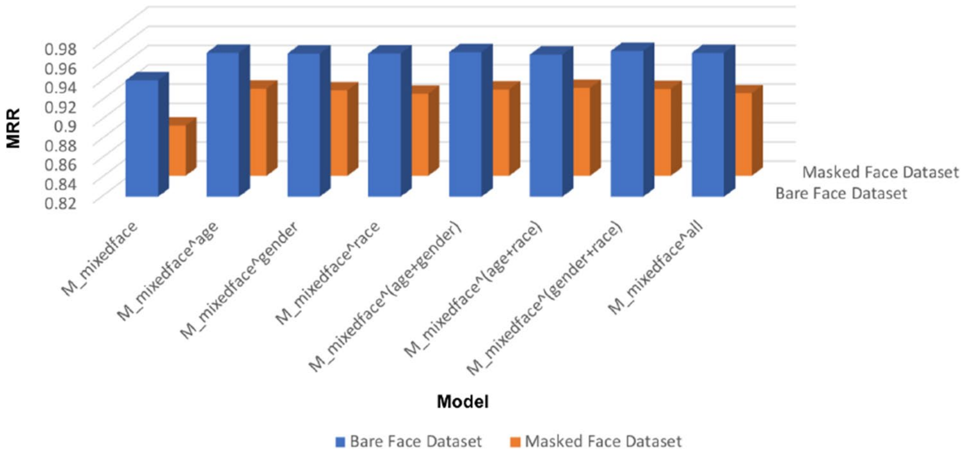

Figure 11 shows the mean reciprocal rank performance to evaluate the capability of our proposed model in ranking the correct prediction in the top position of the recommendation list. The observation is also aligned with the previous metrics where our *FIREC* model could obtain a higher MRR value than the base model without any feature fusion on

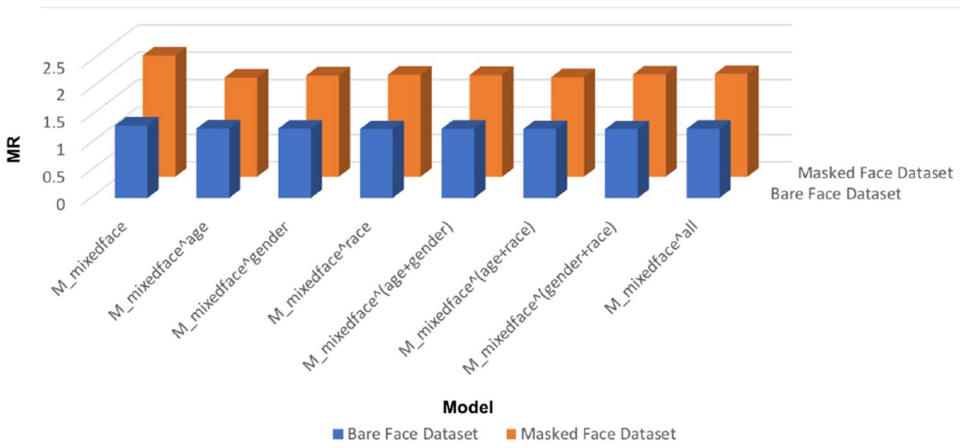

Fig. 12 The MR performance comparison of our proposed model variants

Table 4 Wilcoxon signed-rank test p-values against the base model ($M_{mixedface}$) on two testing datasets with $\alpha = 0.05$

Model	$D^{test}_{bareface}$	$D^{test}_{maskedface}$
$M^{age}_{mixedface}$	< 0.05	< 0.05
$M^{gender}_{mixedface}$	< 0.05	< 0.05
$M^{race}_{mixedface}$	< 0.05	< 0.05
$M^{(age+gender)}_{mixedface}$	< 0.05	< 0.05
$M^{(age+race)}_{mixedface}$	< 0.05	< 0.05
$M^{(gender+race)}_{mixedface}$	< 0.05	< 0.05
$M^{all}_{mixedface}$	< 0.05	< 0.05

both datasets. In addition, as shown in Fig. 12, our proposed *FIREC* model yields lower MRR, indicating that our *FIREC* model not only achieves high accuracy in terms of hit rate and precision but also ranks the correct prediction roughly within the top two results in the recommended list on average.

Besides, to confirm that our proposed *FIREC* model has a significant improvement over the base model that does not integrate a feature-level fusion module, we perform the Wilcoxon signed-rank test between each of our proposed feature-fusion variants against the base model (without feature fusion), as shown in Table 4. We can notice that on both datasets, all p-values are less than 0.05, indicating that the recommendations made by our proposed feature-fusion models are significantly different from those of the baseline (without future fusion). Hence, we can conclude in this section that our proposed *FIREC* model significantly outperforms other baselines in all metrics.

5.3 Case study (RQ3)

Since traditional face identity recognition systems lack the ability to explain justifications for the face identity recommendations, that could hinder the model's reliability and trustworthiness when used by relevant stakeholders. In this section, we show a case study using a test sample from a masked face dataset with $M^{all}_{mixedface}$ model for prediction. As shown in Fig. 13, we could provide an additional explanation along with the recommendation ranking list as follows. The possible top-5 identities of this person (input image) could be (1) *Davis* since the subject is male with 23 years old having white skin, (2) *Martin* since the subject is male with 23 years old having a middle eastern race, (3) *Alex* since the subject is male with 22 years old having white skin, (4) *Steve* since the subject is male with 23 years old having white skin, and (5) *Derick* since the subject is male with 27 years old having white skin. One of the limitations of this work is that photos of individuals in the database may appear to be younger than the real persons due to age differences. In the future, we could explore multiple deep learning techniques such as GAN (Generative Adversarial Nets) for change detection (Li et al. 2021). This kind of application can be used by government institutions, hospitals, airports, etc., to aid the authorities in their timely investigation of suspicious behaviors amidst crisis situations where wearing masks is a common or enforced practice.

6 Conclusions and future directions

In this study, we propose a Deep Feature-Level Fusion Model for Masked Face Identity Recommendation System (*FIREC*) with the design of multiple deep neural network layers to address the complicated masked face recognition problem. Particularly, we augment available full frontal face images with the data management module to obtain our custom masked and mixed face datasets. Then, we utilize the DeepFace network as a backbone to implement the feature extraction module to compute the face descriptor and additional semantic features. We further combine all features with the feature-level fusion module to progressively encode

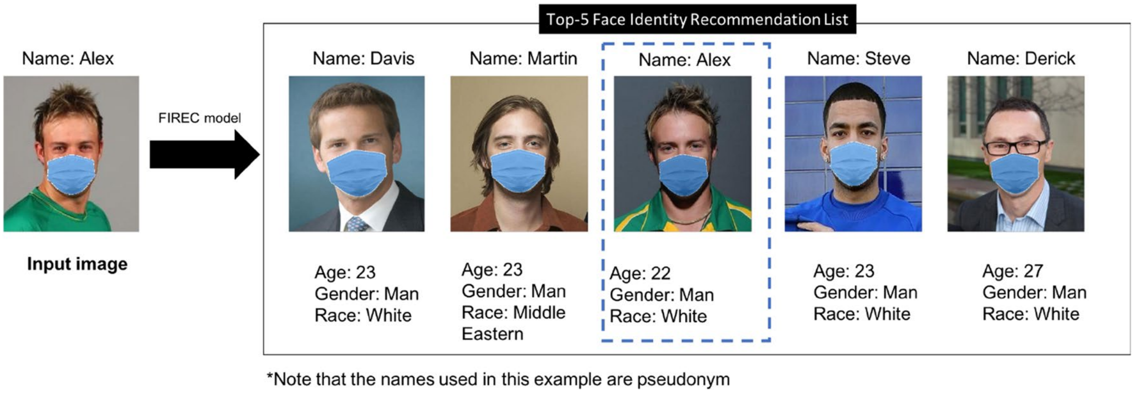

Fig. 13 The example of our proposed model in the real-life application in helping the surveillance worker to inspect the target person

the useful representation for the face identity recommendation task. Finally, the compact representation is fed to the recommendation module, and each identity's probability scores are given. As a result, the top-N highest scores are returned as a recommendation list. Through extensive experiments, we demonstrate that our proposed model outperforms all baselines in various metrics in terms of accuracy, ranking, and explainability. As future work, we aim to incorporate more biometric-indication features such as ear, eye, hair, and face shape into the model design to better handle non-frontal face images captured from different angles.

Acknowledgements This research project is supported by Mahidol University (Grant no. MU-MiniRC02/2564). Suppawong Tuarob is the corresponding author.

Code availability The source code is made available for research purposes at https://github.com/Zenonist/FIREC.

References

Ahmed T, Das P, Ali MF, Mahmud M-F (2020) A comparative study on convolutional neural network based face recognition. In: 2020 11th international conference on computing, communication and networking technologies (ICCCNT). IEEE, pp 1–5

Ahonen T, Hadid A, Pietikainen M (2006) Face description with local binary patterns: application to face recognition. IEEE Trans Pattern Anal Mach Intell 28(12):2037–2041. https://doi.org/10.1109/tpami.2006.244

Ahonen T, Rahtu E, Ojansivu V, Heikkila J (2008) Recognition of blurred faces using local phase quantization. In: 2008 19th International conference on pattern recognition. https://doi.org/10.1109/icpr.2008.4761847

Anwar A, Raychowdhury A (2020) Masked face recognition for secure authentication. arXiv preprint. arXiv:2008.11104

Arigbabu OA, Ahmad SMS, Adnan WAW, Mahmood S (2015) Soft biometrics: gender recognition from unconstrained face images using local feature descriptor. J Inf Commun Technol. https://doi.org/10.32890/jict2015.14.0.8159

Aswal V, Tupe O, Shaikh S, Charniya NN (2020) Single camera masked face identification. In: 2020 19th IEEE international conference on machine learning and applications (ICMLA). https://doi.org/10.1109/icmla51294.2020.00018

Beveridge JR, Zhang H, Draper BA, Flynn PJ, Feng Z, Huber P, Kittler J, Huang Z, Li S, Li Y et al (2015) Report on the FG 2015 video person recognition evaluation. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG). https://doi.org/10.1109/fg.2015.7163156

Bonnen K, Klare BF, Jain AK (2013) Component-based representation in automated face recognition. IEEE Trans Inf Forensics Secur 8(1):239–253. https://doi.org/10.1109/tifs.2012.2226580

Chang X, Wu J, Yang T, Feng G (2020) Deepfake face image detection based on improved VGG convolutional neural network. In: 2020 39th Chinese control conference (CCC). https://doi.org/10.23919/ccc50068.2020.9189596

Chen J-C, Ranjan R, Patel VM, Castillo CD, Chellappa R (2018) Unconstrained face identification and verification using deep convolutional features. In: Deep learning in biometrics, pp 33–64. https://doi.org/10.1201/b22524-2

Chen Y-N, Thaipisutikul T, Han C-C, Liu T-J, Fan K-C (2021) Feature line embedding based on support vector machine for hyperspectral image classification. Remote Sens 13(1):130. https://doi.org/10.3390/rs13010130

Chu W, Ying Z, Xia X (2013) Facial expression recognition based on binarized statistical image features. In: 2013 Ninth international conference on natural computation (ICNC). IEEE, pp 328–332

Dantas AJ, Jesus LD, Ramos ACB, Hokama P, Mora-Camino F, Katarya R, Verma OP, Grupta PK, Singh G, Ouahada K (2021) Using UAV, IoMT and AI for monitoring and supplying of COVID-19 patients. In: ITNG 2021 18th international conference on information technology-new generations. Springer, Berlin, pp 383–386

Ding C, Choi J, Tao D, Davis LS (2016) Multi-directional multi-level dual-cross patterns for robust face recognition. IEEE Trans Pattern Anal Mach Intell 38(3):518–531. https://doi.org/10.1109/tpami.2015.2462338

Dosaj A, Satapathy SC, Soundrapandiyan R, Kaur M, Hannoon N (2018) An efficient cloud based face recognition system for e-health secured login using steerable pyramid transform and local directional pattern. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-018-1115-6

Enkhbat A, Shih TK, Thaipisutikul T, Hakim NL, Aditya W (2020) Handkey: an efficient hand typing recognition using CNN for virtual keyboard. In: 2020—5th International conference on information technology (InCIT). https://doi.org/10.1109/incit50588.2020.9310783

Guo C (2017) Enhancing face identification using local binary patterns and k-nearest neighbors. J Imaging 3(3):37. https://doi.org/10.3390/jimaging3030037

Gupta A, Katarya R (2021a) A novel LDA-based framework to forecast COVID-19 trends. Available at SSRN 3833706

Gupta A, Katarya R (2021b) PAN-LDA: a latent Dirichlet allocation based novel feature extraction model for COVID-19 data using machine learning. Comput Biol Med 138:104920

Gupta G, Katarya R (2021c) EnPSO: an AutoML technique for generating ensemble recommender system. Arab J Sci Eng 46(9):8677–8695

Gupta G, Katarya R (2021d) A study of deep reinforcement learning based recommender systems. In: 2021 2nd international conference on secure cyber computing and communications (ICSCCC). IEEE, pp 218–220

Gupta A, Gupta S, Katarya R et al (2021) InstaCovNet-19: a deep learning classification model for the detection of COVID-19 patients using chest X-ray. Appl Soft Comput 99:106859

Jose JP, Poornima P, Kumar KM (2012) A novel method for color face recognition using KNN classifier. In: 2012 International conference on computing, communication and applications. IEEE, pp 1–3

Kar A, Pramanik S, Chakraborty A, Bhattacharjee D, Ho ESL, Shum HPH (2020) LMZMPM: local modified Zernike moment per-unit mass for robust human face recognition. IEEE Trans Inf Forensics Secur 16:495–509

Karaaba M, Surinta O, Schomaker L, Wiering MA (2015) Robust face recognition by computing distances from multiple histograms of oriented gradients. In: 2015 IEEE symposium series on computational intelligence. IEEE, pp 203–209

Katarya Rahul, Arora Yamini (2020) Capsmf: a novel product recommender system using deep learning based text analysis model. Multimedia Tools and Applications 79(47):35927–35948

Katarya R, Saini R (2022) Enhancing the wine tasting experience using greedy clustering wine recommender system. Multimedia Tools Appl 81(1):807–840

Katarya R, Verma OP, Jain I (2013) User behaviour analysis in context-aware recommender system using hybrid filtering approach. In: 2013 4th International conference on computer and communication technology (ICCCT). IEEE, pp 222–227

Katarya R, Gupta A, Sachdeva S, Dhamija T, Gupta S, Gupta A, Kedia P, Rai V et al (2021) A review of various mathematical and deep learning based forecasting methods for COVID-19 pandemic. In: 2021 7th International conference on advanced computing and communication systems (ICACCS), vol 1. IEEE, pp 874–878

Kedia P, Katarya R et al (2021) CoVNet-19: a deep learning model for the detection and analysis of COVID-19 patients. Appl Soft Comput 104:107184

Kusakunniran W, Karnjanapreechakorn S, Siriapisith T, Borwarnginn P, Sutassananon K, Tongdee T, Saiviroonporn P (2021) COVID-19 detection and heatmap generation in chest x-ray images. J Med Imaging 8(S1):014001

Li X-X, Hao P, He L, Feng Y (2020) Image gradient orientations embedded structural error coding for face recognition with occlusion. J Ambient Intell Humaniz Comput 11(6):2349–2367

Li X, Du Z, Huang Y, Tan Z (2021) A deep translation (GAN) based change detection network for optical and SAR remote sensing images. ISPRS J Photogramm Remote Sens 179:14–34

Ma Z, Xiang Z (2015) Robust visual tracking via binocular multi-task multi-view joint sparse representation. In: 2015 SAI intelligent systems conference (IntelliSys). IEEE, pp 714–722

Naveen P, Sivakumar P (2021) Adaptive morphological and bilateral filtering with ensemble convolutional neural network for pose-invariant face recognition. J Ambient Intell Humaniz Comput 12(11):10023–10033

Qi G-J, Hua X-S, Rui Y, Tang J, Zhang H-J (2009) Two-dimensional multilabel active learning with an efficient online adaptation model for image classification. IEEE Trans Pattern Anal Mach Intell 31(10):1880–1897. https://doi.org/10.1109/tpami.2008.218

Ojansivu V, Heikkilä J (2008) Blur insensitive texture classification using local phase quantization. In: International conference on image and signal processing. Springer, Berlin, pp 236–243

Sarangi PP, Nayak DR, Panda M, Majhi B (2022) A feature-level fusion based improved multimodal biometric recognition system using ear and profile face. J Ambient Intell Human Comput 13:1867–1898. https://doi.org/10.1007/s12652-021-02952-0

Serengil SI, Ozpinar A (2020) Lightface: a hybrid deep face recognition framework. In: 2020 Innovations in intelligent systems and applications conference (ASYU). IEEE, pp 23–27. https://doi.org/10.1109/ASYU50717.2020.9259802

Simonyan K, Parkhi O, Vedaldi A, Zisserman A (2013) Fisher vector faces in the wild. In: Proceedings of the British machine vision conference 2013,. https://doi.org/10.5244/c.27.8

Sun J, Fu Y, Li S, He J, Xu C, Tan L (2018) Sequential human activity recognition based on deep convolutional network and extreme learning machine using wearable sensors. J Sens 1–10:2018. https://doi.org/10.1155/2018/8580959

Taigman Y, Yang M, Ranzato MA, Wolf L (2014) Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1701–1708

VenkateswarLal P, Nitta GR, Prasad A (2019) Ensemble of texture and shape descriptors using support vector machine classification for face recognition. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-019-01192-7

Vera-Rodriguez R, Blazquez M, Morales A, Gonzalez-Sosa E, Neves JC, Proença H (2019) FaceGenderID: exploiting gender information in DCNNs face recognition systems. In: 2019 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 2254–2260. https://doi.org/10.1109/CVPRW.2019.00278

Xu C, Tao D, Xu C (2014) Large-margin multi-viewinformation bottleneck. IEEE Trans Pattern Anal Mach Intell 36(8):1559–1572. https://doi.org/10.1109/tpami.2013.2296528

Yolcu G, Oztel I, Kazan S, Oz C, Bunyak F (2019) Deep learning-based face analysis system for monitoring customer interest. J Ambient Intell Humaniz Comput 11(1):237–248. https://doi.org/10.1007/s12652-019-01310-5

Zhu X, Liao S, Lei Z, Liu R, Li SZ (2007) Feature correlation filter for face recognition. In: International conference on biometrics. Springer, Berlin, pp 77–86

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