Multimedia Tools and Applications (2020) 79:11125–11141
https://doi.org/10.1007/s11042-019-7323-z

Effective multiple person recognition in random video sequences using a convolutional neural network

Niraimathi Puhalanthi[1] · **Daw-Tung Lin**[1]

Received: 17 June 2018 / Revised: 3 January 2019 / Accepted: 31 January 2019 /
Published online: 9 February 2019

Abstract

Effective and efficient face recognition through pervasive networks of surveillance cameras is one of the most challenging objectives of advanced computer vision. This study developed a real-time person recognition system (PRS) for the effective identification of multiple people in video sequences. We focused on identifying approximately 9000 celebrities by intelligent preprocessing, training, and deployment of a deep-learning convolutional neural network (CNN). The proposed PRS method comprises the following three major steps. In the first step, multiple faces present in a given frame as well as their associated landmarks are detected. This must be precise because the accuracy of this step dictates the accuracy of the complete PRS. In the second step, the extracted facial regions of interest are then aligned using affine warping, based on their respective identified landmark positions. The alignment process is meant to ensure correct identification of a person, because a wide range of faces entails intrinsic interclass similarities. Finally, in the third step, a VGG-19 CNN is trained to classify the aligned facial images for person recognition. In the training phase of the PRS, we utilized images from the CASIA WebFace database, which contains nearly 9000 classes, and aligned them using their respective facial landmarks. Subsequently, we used the aligned images to train a VGG-19 CNN classifier. For the purpose of validation, the trained classifier was tested with the standard Labelled Faces in the Wild (LFW) database by extracting the features for the LFW images using the trained VGG. Specifically, the VGG-extracted LFW features were used to train support vector machine classifiers, and the obtained resultant classification accuracy of approximately 96% was very close to the currently existing benchmark for the LFW database. During the testing phase, alternate frames of the input video were extracted and the identified faces (post-alignment) were used as inputs into the trained VGG to recognize the people in a given frame. When tested on random samples of video images, the proposed PRS offered robust recognition performance for most of the facial regions that had reasonable facial orientations and sizes. Furthermore, the average recognition time per person was approximately 370 milliseconds. The proposed deep learning-based PRS is the first of its kind to exhibit real-time performance for person recognition with significant accuracy, without involving any prior knowledge of the people involved in a video.

✉ Daw-Tung Lin
dalton@mail.ntpu.edu.tw

Extended author information available on the last page of the article.

Keywords Person recognition · Deep learning · Convolutional neural network · Landmark identification · Face alignment · VGG · CASIA WebFace database · LFW database

1 Introduction

Today's digital world is highly connected and a large proportion of it is under the watchful gaze of countless surveillance cameras. However, utilization of these video resources is comparatively limited because of the heavy complexity of analyzing their vast quantities of data; efficient real-time analysis would be particularly difficult. Person recognition and object positioning have been investigated over the past several years, and only recently have there been breakthroughs in person/object recognition over single and multiple cameras. This stands as a testimony to the cumbersome challenges involved in person/object recognition. Just some of the numerous applications of real-time person recognition are: 1) tracking lost people (especially children) in different regions, 2) locating criminals who are at large places, and 3) identifying and segregating a person in a crowd for security reasons. Notably, manual recognition of a person through a large network of cameras, over thousands of hours, is almost an impossible task, irrespective of the human resources involved. Hence, an automatic, efficient, and robust video-based person recognition system (PRS) is of paramount importance for the safety and security of people in the world's rapidly developing and busy societies.

Deep learning-based approaches have emerged as a major force and yielded excellent state-of-the-art results for many challenging computer vision problems and applications [22, 30, 33]. Because deep-learning architectures involve numerous parameters that must be updated during training, they entail high computational costs. However, simultaneous rapid advancements in the design of higher-end and computationally powerful GPUs have alleviated the computational-cost problem. Although deep-learning approaches are data-centric, their accuracy in many challenging image analysis tasks is high because of the availability of high-quality labelled databases contributed by various relevant research groups (e.g., ImageNet [31]). Moreover, the availability of powerful open tools should also be credited for the growing success of deep-learning applications. The ability of deep-learning networks to identify the required features by themselves (with appropriate design and training) is the key to solving complex problems that could not be solved by merely using handcrafted features.

In this study, we endeavored to develop a effective and efficient PRS. We began by identifying multiple faces in each frame, for which we employed the histogram of oriented gradients (HoG)-based approach. Subsequently, the landmark positions corresponding to each face were located. These landmark positions were then used to align the faces. We trained a VGG-19 deep-learning network [33] and used it for identifying the aligned faces periodically over successive frames. The overall procedures involved in the proposed framework are depicted in Fig. 1. A detailed description of the entire procedure is presented in the subsequent sections. As demonstrated later, such a hybrid technique was designed for jointly circumventing the aforementioned existing problems with currently available PRSs, and consequently yielded robust and real-time performance in person recognition.

The main contributions of this study can be summarized as follows:

– This study used a prime approach that incorporates a deep convolutional neural network (CNN) into the PRS framework for real-time recognition of nearly 9000 classes

(celebrities) from video images, without any prior knowledge of the scene or person involved.
– The proposed alignment-based training and testing significantly reduced false alarms caused by similarities between multiple classes.
– The proposed framework was verified as yielding accurate person recognition over multiple random celebrity videos, thereby demonstrating its reliability and applicability to practical person recognition.
– Finally, although we only focused on 9000 celebrity classes, we believe that our research will foster more relevant research for achieving recognition of a wide range of common citizens.

The remainder of the paper is organized as follows. Section 2 provides a brief literature review that discusses several intriguing aspects and challenges of PRS. The algorithms and protocols used for face detection and landmark identification are presented in Section 3, wherein the regions of interest (ROIs) are identified from a given video sequence. Preprocessing the ROI data is pivotal for the subsequent recognition process, and hence, Section 4 discusses the facial alignment procedure. Section 5 is dedicated to the design and training of a deep-learning network to be used for person recognition. Section 6 details the experimental analysis, results, and associated discussions. Finally, Section 7 presents the concluding remarks and potential future directions.

2 Related work

For several years, PRSs have been a pivotal area of research that is closely associated with studies on face recognition, because face recognition forms the core of a PRS. Face detection is the first step in face recognition, because the human face is an essential feature for identifying a person. Several approaches exist for detecting human faces, which mainly fall into two categories: 1) local feature-based methods, and 2) global methods. Local feature-based methods concentrate on detecting the critical features, such as the eyes, nose, mouth, and ears, which they then use to identify a face. By contrast, global methods involve the entire facial region, extracting the required features for detecting the face.

Viola et al. [39] proposed a famous face detection framework that achieves high detection rates. The key contributions of their method were the introduction of a new image representation method called the integral image and an AdaBoost learning algorithm used to select a very large set of potential features. Their method finally combined the classifiers in a cascade and discarded the background regions of the image to capture the face-like regions [39]. This method is a highly popular and efficient approach for face detection; however, it results in numerous false alarms. Keren et al. [20] proposed Antifaces, a multitemplate scheme for detecting arbitrary objects, including faces in images. Their method used a set of sequential classifiers to detect faces and reject nonfaces. Li et al. [24] proposed a support vector machine (SVM)-based multiview face detector in which they constructed a face pose estimator using support vector regression, and then trained separate face detectors for each face pose.

Now that we have discussed crucial studies on face detection, we proceed to discussing relevant studies in facial-feature extraction and recognition. Earlier face recognition studies have been based on handcrafted features and offered reasonable performance for recognizing several classes under a controlled environment; for instance, a local binary pattern (LBP) [1]. However, when the number of classes was increased and the imaging

environment was drastically varied, these methods suffered because of their intrinsic limitations. Subsequently, methods that used combinations of many handcrafted features were developed and their performances were much closer to the level of human accuracy [3, 5–7]. In particular, a binary face verification methodology was devised that used so-called reference sets, and showed that the extracted features were highly representative for identifying people [3]. A productive idea for deriving high dimensional features from a simple LBP was proposed in [7]. It has also been shown that such high dimensional feature representation resulted in enhanced recognition accuracy, even under challenging situations. Furthermore, Bayesian inference-based approaches such as [6] and [5] have exhibited strong recognition performance. Whereas a joint formulation procedure for modelling a face pair with appropriate priors was proposed in [6]. Cao et al. [5] proposed an enriched learning approach based on Bayesian classifiers and presented more interesting theoretical insights. All of these methods are considered later in this paper for performance evaluation. In addition, although the methods in [13] and [38] are relatively old, they are still worth mentioning. Specifically, in [13], the authors proposed a method based on skin color detection and studied a suitable design for a color model that could be best used to construct an efficient face detector. Moreover, [38] discussed a method that works under uncontrolled illumination conditions; the authors presented a preprocessing chain that eliminated most of the effects of changing illumination, and introduced local ternary patterns and improved robustness by adding kernel principal component analysis-based feature extraction. More recent innovations include methods that are robust to pose and background variations [16, 42]. Preprocessing of facial images (particularly the alignment) is critical for face recognition and the importance of such techniques have been studied in [32, 37] and, [45]. Furthermore, although face alignment was not explicitly considered in [32], the study concluded that adding additional face alignments during the testing stage may further increase recognition accuracy.

Nevertheless, many existing PRSs (and their associated face recognition methods) are limited by their ability to recognize only a relatively small number of people under controlled environments. In other words, the performances of current PRSs are severely affected by factors such as illumination variations, changes in hair styles and facial expressions, makeup, motion artifacts, and occlusions [10, 12, 29, 40, 43]. Such inevitable practical challenges make PRSs one of the most challenging computer vision applications, for which no robust solutions are available. However, this is a most vital application, and hence, much research is being conducted in this direction. The present study is one such effort in the direction of real-time person recognition from video sequences, through partially mitigating the aforementioned problems using the joint prowess of appropriate facial alignment and deep-learning techniques. Furthermore, PRSs that use other soft biometric information as well as facial features have recently been studied [14], and notably, deep learning-based activity recognition [8, 25–28] is a parallel development, but is extremely different to PRS.

3 Face detection and landmark identification

The first and foremost step in the PRS design is the identification of the faces present (if any) in a given frame. This face identification is then followed by the identification of landmarks in each of the detected faces. These two procedures are not just crucial for the ensuing deep learning-based PRS, but also are instrumental in performing certain preprocessing techniques, such as significant face filtering and frame analysis. The overall procedures involved

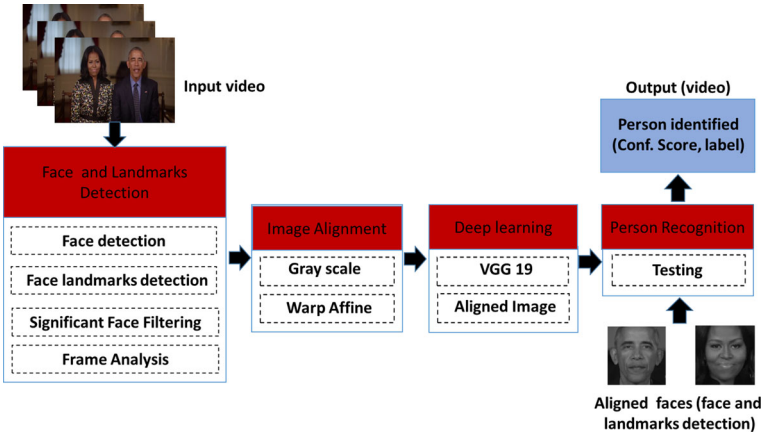

Fig. 1 An illustration of the entire proposed PRS approach for recognizing people in videos

in the proposed framework are depicted in Fig. 1. The details pertaining to face detection and landmark identification are discussed in the following two sections (3.1 and 3.2), respectively. Furthermore, the preprocessing techniques aimed at improving the recognition speed and accuracy are presented later in Section 3.3.

3.1 Face detection

In this study, we employed a customized version of the standard histogram of oriented gradients (HoG)-based face detection scheme [9, 11]. Specifically, for each frame, we first extracted the HoG features. Then, we used overlapping sliding windows that collect a set of HoG features in each window region over the entire frame. Next, a trained SVM classifier was used to determine whether the estimated features in a given window region accounted for a face or not. We fine-tuned the classification threshold to ensure reliable detection of faces and avoid false alarms, and we also refrained from using image pyramids for enhanced efficiency. Although such customizations resulted in failure to identify small faces in the current frame, it did not affect the performance, because we intended to recognize only faces of significant sizes and orientation. This is further discussed in Section 3.3. Furthermore, although other interesting face detection methods exist, such as [2] and [36], the performance of the aforementioned HoG-based approach was empirically found to be apt for the proposed PRS, and hence it was used in this study. Notably, we employed this face detection methodology (and the subsequent steps) for both the training data (as input for the deep-learning network c.f. Section 5) and for the testing phase. Once the faces were identified in the current frame, the next task was to identify the appropriate landmark positions pertaining to the respective faces.

3.2 Landmark identification

We next aimed to identify the 68 standard landmark positions in the identified faces. For this purpose, a state-of-the-art regression-based landmark detection algorithm [19] was employed. For the sake of completeness, the procedure involved in the landmark identification process is briefly explained as follows. The core of the landmark identification

algorithm is a cascade of intelligently formulated regression functions, which are aptly trained using a gradient-boosting procedure. Unlike other regression-based methods, [19] employed regression functions that estimated the predefined shapes from corresponding initial estimates and the associated subsets of facial pixels that were identified in relation to the initial estimates; furthermore, the reliable features corresponding to the predefined shapes were iteratively obtained to ensure the accuracy of the located predefined shapes. These factors were incorporated into the design of an appropriate loss function, which was then optimized to yield minimum loss both during training and evaluation. Such a formulation resulted in the efficient and precise location of landmarks irrespective of the illumination changes and other inevitable imaging artifacts, thereby making the method more suitable for our proposed *in-the-wild* PRS. The landmarks were used to locate the facial regions, and the accuracy of these landmark positions played a vital role because they were used for the subsequent alignment process (c.f. Section 4). Figure 2 presents samples of different facial images, their corresponding face ROIs, and their respective facial landmark positions.

3.3 Preprocessing (filtering) schemes

Once the face region and the associated landmarks were identified, we performed some filtering to ignore the *weak* faces. This procedure ensured enhanced performance accuracy of the final PRS. First, we calculated the facial dominance score (FDS), which is a function of the number of pixels in each identified facial image and the total number of pixels in the current frame. Mathematically, the FDS of a face can be defined as:

$$\text{FDS} = \frac{\text{No. of Facial Pixels}}{\text{Frame resolution}} \times 100. \tag{1}$$

In this study, if the FDS for a face image was greater than 15%, then we retained the face for recognition, otherwise we discard the current face because it would lead to ambiguous recognition. Furthermore, the orientation of the face is a factor that affects recognition accuracy, because faces with higher orientation are highly likely to be misidentified. To locate faces with reasonable orientations, we counted the number of facial landmarks obtained for a given face. Faces with at least 51 landmarks (up to a maximum of 68 landmarks) were considered suitable for the PRS, whereas the other faces (with lesser landmarks) were ignored as faces with severe orientations. These two procedures were used to filter out weak facial images and retain the reasonable ones for the recognition task.

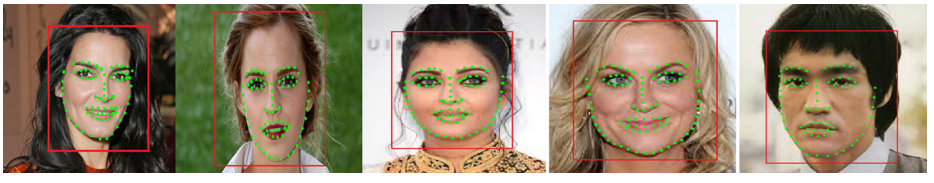

Fig. 2 Samples of facial images, associated facial regions of interest (red rectangles), and identified landmark positions (green points)

4 Landmark-based image alignment

Now we discuss the mandatory alignment techniques that are pivotal to the accuracy of the PRS. Notably, the proposed PRS was designed to identify people (in this case celebrities) in the wild (i.e., irrespective of the background). Furthermore, the recognition was performed using an appropriately trained multiclass deep-learning network. Hence, the alignment process was aimed at providing a reasonable facial image (irrespective of the orientation effects in the original frame). The alignment procedure for each facial image was based on the respective landmarks estimated (c.f Section 3.2), and the detailed procedure includes the following steps:

1. First, convert the RGB image to grayscale and perform the alignment in the grayscale image. This is just for the sake of efficacy because the ensuing deep-learning network will be completely trained and tested using grayscale images.
2. The obtained image may have illumination issues owing to factors such as shadowing effects and excessive lighting. To circumvent this, use contrast-limited adaptive histogram equalization (CLAHE) [35]. CLAHE is a conventional preprocessing technique that has been frequently used in many image classification applications to nonlinearly enhance the quality of a given image.
3. Based on the estimated landmark positions for a given face, the left and right eye centers are obtained, and the corresponding angles with respect to the horizontal axis are also calculated. If (x_l, y_l) and (x_r, y_r) are the center coordinates of the left and right eye, respectively, then the angle (θ) is estimated as

$$\theta = \arctan \frac{y_l - y_r}{x_l - x_r} \times 180. \tag{2}$$

4. According to the estimated angle, the rotation matrix is computed and affine warping (transformation) is applied to the facial image to account for the implicit angular orientation.

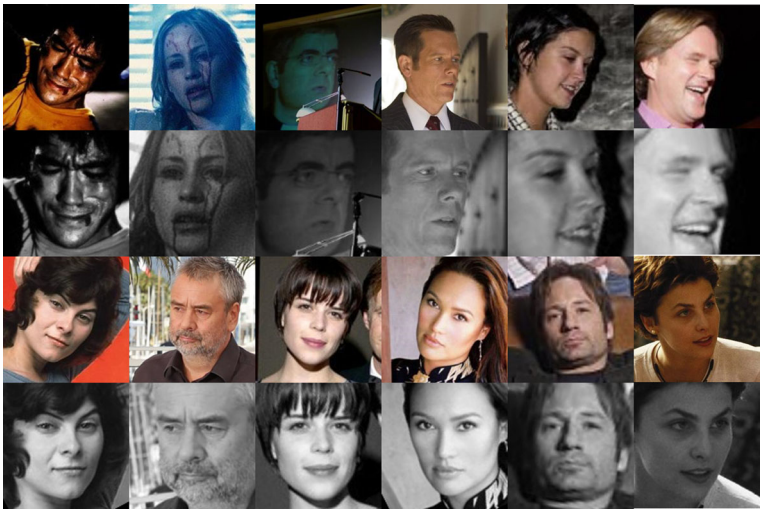

Fig. 3 Several illustrations of the output of the face alignment procedure (even rows) and their corresponding original input facial image (odd rows)

5. Finally, the image is resized according to the distance between the center of the eyes and the chin. This distance is maintained as constant for all facial images to ensure uniformity (see Fig. 3).

This procedure is applied only for facial images that pass the filtering criteria stipulated in Section 3.3. Some examples of the face alignment procedure are provided in Fig. 3, in which the alignment procedure can be readily seen to aid in providing an optimal facial images irrespective of the actual orientation of the respective original images (also shown in Fig. 3).

5 CNN for person recognition

The main module of the proposed PRS methodology is the deep-learning network that takes the aligned facial image as the input and estimates the person corresponding to the input facial image. Because the PRS is expected to identify numerous classes, we resorted to using a powerful deep-learning network that can effectively and efficiently handle the recognition process. For this purpose, we involved two of the largest available public databases, namely CASIA WebFace [41] and Labelled Faces in the Wild (LFW) [17], for the training and validation of the associated deep-learning network, respectively. In this section, we discuss the details pertaining to the involved CNN and the associated training, validation, and testing.

For more than two decades, CNNs have predominantly been used in various computer vision applications. In spite of their initial success, neural network-based methods suffer from two main intrinsic limitations. The first limitation pertains to the lack of sufficient depth in the network to learn highly diverse and specific features. This could be attributed to the inability of handling millions of weights (and updating them) during the training process. The unavailability of sufficiently large databases is another major constraint that has heavily dictated the limitations of CNNs.

The predecessor for the existing deep-learning frameworks dates back to the late 1980s, when [21] first proposed a deep-network architecture. However, at that time, powerful computational hardware and robust, effective, and efficient large-scale learning algorithms were not widely available. For decades, the development and utilization of deep-learning technology was almost at a standstill. In recent years, rapid advancements in GPU architecture have revived this fascinating field. Since 2012, a plethora of deep-learning models, network architectures, and learning algorithms have been proposed for various computer vision applications. The areas in which deep-learning solutions have made great contributions include object classification, regression, and identification (which refers to the combination of localization and classification). In most of these diverse applications, the obtained accuracy levels have been remarkably close to human accuracy; deep learning techniques have set new benchmarks in almost all relevant applications. These great successes can be attributed to not only innovative CNNs, but also to the existence of large databases and their public availability. Notably, deep-learning applications are data-centric approaches and are heavily dependent on the availability of large training, validation, and testing databases. The availability of such large databases encourages deep training/learning using deep-learning networks. Furthermore, such a large database could obviously result in over-fitting when training CNNs (also known as shallow networks).

The deep-learning CNN considered in this study is the VGG-19. The reason for choosing this network is because of its demonstrated superior performance for various classification tasks as reported by various research groups. We will briefly present the architectural details

as well as the subtle yet significant customizations made to the VGG-19 architecture to suit the requirements of the PRS. In this study, the input image was a cropped 114 × 114 grayscale image. Because augmentation has been shown to be a critical process in improving training and validation accuracies [4, 15], we employed randomized (within a reasonable range) interpolations to obtain 10 different facial ROIs for each input and used them to train the deep VGG network. Such augmentation aids in mitigating the effects of practical facial ROI variations, especially during the validation and testing phase. The convolutional kernel size was set as 3 in all convolutional layers. Max pooling was used to downsize the images across different layers. The most commonly used rectified linear unit (ReLu) activation function $f(x) = \max(0, x)$ was applied and the respective pooling layers were appropriately defined with a kernel size and stride of 2. The purpose for defining a small kernel size was to ensure the learning of fine details (features) pertaining to faces. Because CNNs are able to learn the required features by themselves after appropriate training, we used the largest publicly available CASIA WebFace database [41] for training. Further details regarding the training, convergence, and validation are discussed in Section 6. Figure 4 illustrates the VGG-19 CNN used for the PRS in this study.

As mentioned earlier, deep learning is a data-centric approach. Hence, the amount and quality of data used in training the CNN should be pivotal. In this study, we focused on alignment and preprocessing to ensure that the data for training and testing were consistent during the training and testing procedures. Once the data were prepared for training, the next task was to proceed with the design and training of the CNN. The standard and well-known convolutional architecture for fast feature embedding (CAFFE) was used to design, train, and test the proposed VGG-19 CNN. The aligned data were first converted to the lighting memory-mapped database (LMDB) data format, which can be used by the CAFFE architecture for learning. We used the CASIA WebFace database for the training and independently used the LFW database for validation. For the purpose of validation, the deeply learned features were extracted from the images in the LFW database and an SVM was trained with them exclusively to verify the efficacy of our trained CNN. Further details regarding validation and testing are discussed in Section 6.

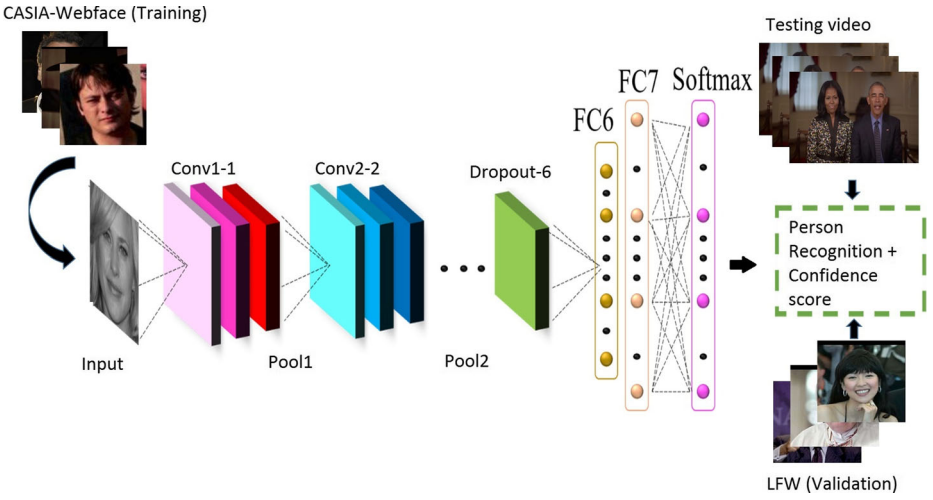

Fig. 4 Illustration of the VGG-19 convolutional neural network used in the person recognition system to recognize the people corresponding to the input facial image

The CASIA WebFace database is a vast collection of images of celebrities. Its specialty is that the images are taken in the wild, and hence, they can have very different and varying backgrounds. The next step in our process was to convert the aligned and preprocessed images into LMDB format. The model we designed was fed into the CAFFE architecture in the form of a text file. Once the network had been designed, the training procedure followed. The training parameters and other details are discussed in the subsequent section.

6 Performance evaluation and discussion

This section describes the experimental analysis of the proposed PRS's performance. The proposed method was implemented on a Windows PC equipped with an Intel Core i7 - 3.6 GHz CPU with 16 GB RAM and a GeForce GTX TITAN X GPU (NVIDIA). For training and testing the VGG-19 network, the well-known open source CAFFE framework [18] was employed. The multithreading environment was designed using Visual Studio 2013, and OpenCV and CUDA 7.5 were used as additional dependencies for the implementation.

The CASIA WebFace database contains approximately 0.4 million images with more than 9000 classes, and in this study, it was used to train the VGG-19 CNN. We discarded some erroneous classes and selected a total of 8994 classes for training. Each facial image was rescaled to 114×114 and considered as an input for the VGG deep-learning network (during both the training and testing phases). The 80-10-10 rule was followed, meaning that 80% of the data was used for training, 10% for validation (during training), and the remaining 10% was for testing. The images in CASIA WebFace were cropped according to the facial ROI and aligned as per the procedures explained in Section 4. Table 1 summarizes the design parameters for training the VGG-19 CNN. Furthermore, it can be inferred from the accuracy graph in Fig. 5 that the maximum accuracy obtained during the training was nearly 0.8. The convergence graph in Fig. 6 reveals superior convergence during the training process.

This is because the CASIA WebFace database is large and cumbersome with some classes having very limited number of images, contributing to the long-tail effect. It should be emphasized that the obtained training (convergence) accuracy of approximately 80% was fairly close to the current standard in the literature. Nevertheless, such a convergence was sufficient enough for tracking the required features for person classification. To demonstrate this, we used the LFW database and utilized the trained CASIA WebFace model to extract the features for the LFW images and compared the same with the corresponding ground-truth features for the multiple classes in the LFW database using its View 2 protocol. Such

Table 1 Design parameters for training the VGG-19 convolutional neural network

Parameter	Values
Batch Size	32
No. of Test Iterations	250
No. of Test Interval	2500
Base Learning Rate	0.01
Momentum and Decay	0.9 and 0.0005
Step Size	300,000
Max. No. of Iterations	500,000

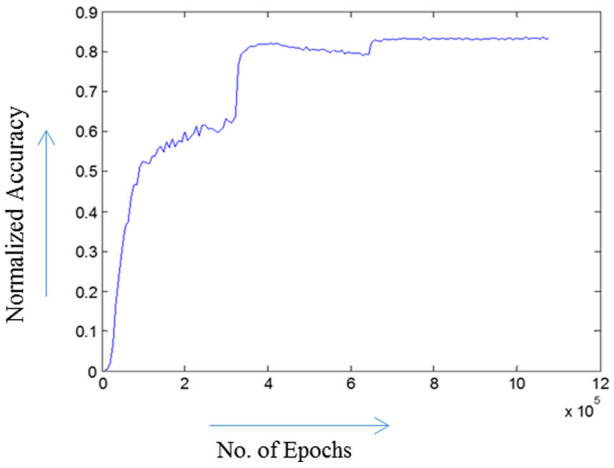

Fig. 5 Accuracy curve for the training of the CASIA WebFace database using the VGG-19 convolutional neural network

an analysis is common for evaluating the classification accuracy and we followed the evaluation protocol dictated in [37]. We trained an SVM classifier using the CASIA WebFace extracted features and used it to calculate the recognition accuracy for the LFW database (95.86%). To further enhance this accuracy, we investigated several possibilities and found that the CASIA WebFace image dataset suffers from long-tail effect [44]. The long-tail effect can be described as a process by which a nonuniform distribution of samples occurs in each class. In other words, some classes of the database could have fewer than 10 images, whereas other classes could have more than 300 images. The classes with more than 300 images could potentially induce bias over the classes with fewer than 10 images because of a greater number of samples. Hence, we scrutinized and modified the CASIA WebFace database. We set a threshold (for example, 40 images) and discarded the classes that contained fewer images than this threshold. Through this procedure, we discarded nearly 3000 classes, and a total of approximately 5000 classes were selected to train the PRS model. The obtained model did not exhibit significant deviations in accuracy and convergence, which indicates that the long-tail effect is not a serious factor in the CASIA WebFace database.

Table 2 Classification accuracies obtained from applying different methods to the LFW database

Classification approach	Accuracy
Joint Bayesian [6]	92.42%
Tom Pete Classifier [3]	93.30%
High Dim. LBP [7]	95.17%
TL Joint Bayesian [5]	96.33%
DeepFace [37]	97.15%
DeepID [34]	97.45%
FaceNet [32]	99.63%
Our Method -I	95.86%
Our Method -II	96.83%
Human Accuracy	97.53%

However, the obtained model performed marginally more effectively when tested on the LFW database, yielding an SVM classification accuracy of 96.83% with a standard deviation of 0.4 on a 10-fold cross validation. The SVM results (i.e., classification accuracies) are summarized in Table 2 along with the accuracies exhibited by other related approaches. In Table 2, it can be instantly observed that the proposed method has an accuracy very close to the human level.

Moreover, Table 2 shows the accuracy obtained using our methodologies. Method I with all classes had an accuracy of 95.86% and Method II with minor classes removed had an accuracy of 96.83%, which are superior to most contemporary methods and closer to state-of-the-art methods, such as DeepID [34], DeepFace [37], and FaceNet [32]. Although the performance of DeepID [34] is higher, it should be mentioned that DeepID employs an extensive database and different protocols for the training process. Evidently in the case of deep learning, the amount of data is more crucial than the algorithm itself. Accordingly, the performance of deep learning would increase in proportion to the amount of data. The mild drop in accuracy of our method compared with DeepFace may be because of the smaller training data size, as well as the rigorous preprocessing (filtering) methods we employed to discard faces that failed the filtering criteria. As previously stated, such filtering methods are mandatory to minimize misclassification errors, especially among celebrities with similar appearances. Finally, although FaceNet exhibits the highest accuracy of all the other methods, unlike the proposed effective and efficient methodology of this study, FaceNet used a different triplet loss with a triplet selection strategy and employed computationally expensive Inception models, which are trained with an extensive private database to achieve the highest accuracy.

Our deep-learning approach is a classification-based approach. In other words, the VGG-19 CNN network is trained with thousands of classifiers, each corresponding to a person in the CASIA WebFace and LFW databases. Figure 6 presents the convergence curve for the training of CASIA WebFace database using VGG-19. Even humans can easily make recognition errors when images are unclear as well as when a person's makeup and hairstyle differ greatly to their trademark characteristics. Hence, it is natural to expect the PRS to make erroneous recognitions of people involved. Therefore, we proposed introducing a simple measure called the confidence score, which serves as measure for the reliability of the estimated output. The confidence score of an input image is a function of the significant

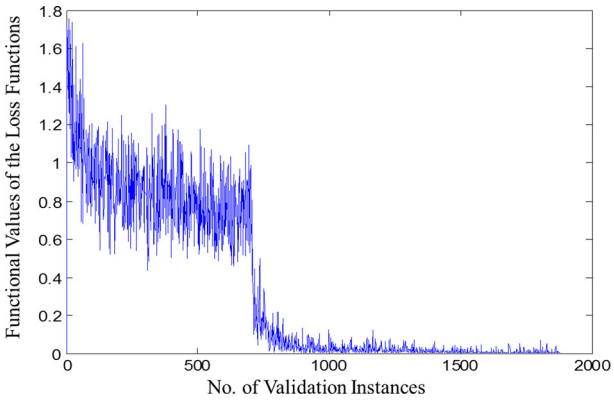

Fig. 6 Convergence curve for training the CASIA WebFace database using the VGG-19 convolutional neural network

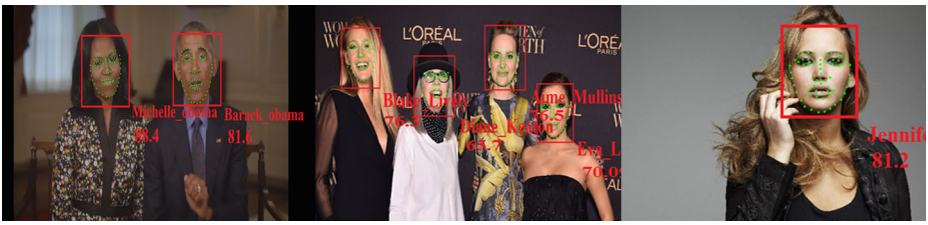

Fig. 7 Screenshot frames of the person recognition system's output from three random test video sequences

softmax output scores from the trained CNN. Mathematically, this can be expressed as:

$$confidence = s_1 - \sum_{i|s_i>0.05} s_i \times 100, \tag{3}$$

where s_i is the ith softmax score, which is normalized to be in the range of [0, 1]. It is trivial to observe the rationale behind (3). If the topmost softmax score is significantly larger than the other scores, then the confidence measure is high, whereas if the topmost softmax score is close to the consecutive scores, then it is highly possible that the obtained recognition estimate could be erroneous, and hence has a lower confidence value.

After the validity of the trained deep-learning CNN was verified, it could be used in the PRS to recognize the facial image in each frame. To ensure a fair evaluation, we randomly selected YouTube videos of celebrities in the CASIA WebFace database and tested the PRS on them. During the testing phase, alternate frames of the input video were extracted and the identified faces (post-alignment) were input into the trained VGG-19 CNN to recognize the person or people in a given frame. The proposed PRS offered robust recognition performance for most facial regions that had reasonable facial orientations and sizes, with an average recognition time per person of approximately 370 milliseconds. Figure 7 presents samples of the obtained recognition results for different video inputs. Most of the obtained recognition results revealed a higher confidence measure, thereby demonstrating the superior efficacy of our PRS.

7 Conclusion and future directions

PRSs are one of the most challenging but critical applications of computer vision. In this study, we successfully developed a deep learning-based multiclass PRS for the real-time recognition of people (using nearly 9000 celebrity classes) from videos. The main aspects of this study were face alignment, filtering (i.e., the elimination of weak faces), and training and testing of the CNN exclusively for this PRS. The ROI cropped and aligned CASIA WebFace database was used for training the VGG-19 CNN, and the accuracy of the trained CNN was verified using the LFW database. The obtained accuracy was remarkably close to state-of-the-art classification methods. The real-time PRS was tested on random celebrity videos, and for most cases, the recognition accuracy and confidence scores were significant except for cases with severe orientation issues.

This study is one of the first of its type, and hence, it has limitations. Therefore, much scope exists for further improvement. The first and foremost extension of this study will be to include a greater number of classes, including common people. This must be coupled with appropriate training and preprocessing techniques. Furthermore, the image alignment

process must be extended for even smaller faces in videos, which is ignored in the current PRS. Moreover, designing novel weight- or voting-based recognition schemes could be instrumental in enhancing recognition accuracy [23]. Combining several existing databases instead of relying on a single large database (i.e., CASIA WebFace) is currently under investigation. Such a combined training mechanism would lead to the recognition of a greater number of classes (people) compared with the training methodology employed in the present work. We hope this study will foster more research in this direction, thereby aiding in achieving a global real-time PRS that can recognize numerous classes with high accuracy.

Acknowledgements This work was supported in part by the Ministry of Science and Technology, Taiwan, Grants MOST 105-2221-E-305-006-MY3, MOST 105-2622-E-305-001-CC3 and MOST 106-2622-E-305-003-CC3, and by the Orbit Technology Incorporation. The Authors would like to thank the providers of CASIA WebFace and LFW database. We would like to acknowledge that the videos used in testing the proposed PRS are chosen from YouTube.

References

1. Ahonen T, Hadid A, Pietikäinen M (2004) Face recognition with local binary patterns. In: Proceedings of ECCV, pp 469–481
2. Amos B, Ludwiczuk B, Satyanarayanan M (2016) Openface: a general-purpose face recognition library with mobile applications. Tech. rep., CMU-CS-16-118, CMU School of Computer Science
3. Berg T, Belhumeur PN (2012) Tom-vs-pete classifiers and identitypreserving alignment for face verification. In: Proceedings of BMVC, vol 2, p 7
4. Bloice MD, Stocker C, Holzinger A (2017) Augmentor: an image augmentation library for machine learning. arXiv preprint arXiv:1708.04680
5. Cao X, Wipf D, Wen F, Duan G, Sun J (2013) A practical transfer learning algorithm for face verification. In: Proceedings of ICCV, pp 3208–3215
6. Chen D, Cao X, Wang L, Wen F, Sun J (2012) Bayesian face revisited: a joint formulation. In: Proceedings of ECCV, pp 566–579
7. Chen D, Cao X, Wen F, Sun J (2013) Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification. In: Proceedings CVPR, pp 3025–3032
8. Cui J, Liu Y, Xu Y, Zhao H, Zha H (2013) Tracking generic human motion via fusion of low-and high-dimensional approaches. IEEE Trans Syst Man Cybern Syst Hum 43(4):996–1002
9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceeding of CVPR, vol 1, pp 886–893
10. Ding C, Tao D (2018) Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE Trans Pattern Anal Mach Intell 40(4):1002–1014
11. Felzenszwalb P, Girshick R, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645
12. Fontaine X, Achanta R, Süsstrunk S (2017) Face recognition in real-world images. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1482–1486
13. Gonzalez C, Jose M (2010) Detecting skin in face recognition systems: a color spaces study. Digital Signal Process 20(3):806–823
14. Gonzalez-Sosa E, Fierrez J, Vera-Rodriguez R, Alonso-Fernandez F (2018) Facial soft biometrics for recognition in the wild: recent works, annotation, and cots evaluation. IEEE Trans Inf Forensics Secur 13(8):2001–2014
15. Hauberg S, Freifeld O, Boesen A, Larsen L, Fisher JW, Hansen LK (2016) Dreaming more data: class-dependent distributions over dieomorphisms for learned data augmentation. In: Proceedings of 19th international conference on artificial intelligence and statistics
16. Hu L, Kan M, Shan S, Song X, Chen X (2017) Ldf-net: learning a displacement field network for face recognition across pose. In: 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017). IEEE, pp 9–16
17. Huang GB, Mattar M, Lee H, Learned-Miller E (2012) Learning to align from scratch. In: Advances in neural information processing systems, pp 764–772

18. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In: ACM international conference on multimedia, pp 675–678
19. Kazemi V, Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. In: Proceedings CVPR, pp 1867–1874
20. Keren D, Osadchy M, Gotsman C (2001) Antifaces: a novel fast method for image detection. IEEE Trans Pattern Anal Mach Intell 23(7):747–761
21. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1(4):747–761
22. LeCun Y, Boser B, Denker JS, Howard RE, Habbard W, Jackel LD, Henderson D (1990) Handwritten digit recognition with a back-propagation network. In: Advances in neural information processing systems, pp 396–404
23. Lee KC, Ho J, Yang M, Kriegman D (2003) Video-based face recognition using probabilistic appearance manifolds. In: Proceedings of CVPR, vol 1
24. Li Y, Gong S, Liddell H (2000) Support vector regression and classification based multi-view face detection and recognition. In: Proceedings of international conference on automatic face and gesture recognition
25. Liu Y, Nie L, Han L, Zhang L, Rosenblum DS (2015) Action2activity: recognizing complex activities from sensor data. In: IJCAI, pp 1617–1623
26. Liu L, Cheng L, Liu Y, Jia Y, Rosenblum DS (2016) Recognizing complex activities by a probabilistic interval-based model. In: AAAI, vol 30, pp 1266–1272
27. Liu Y, Nie L, Liu L, Rosenblum DS (2016) From action to activity: sensor-based activity recognition. Neurocomputing 181:108–115
28. Lu Y, Wei Y, Liu L, Zhong J, Sun L, Liu Y (2017) Towards unsupervised physical activity recognition using smartphone accelerometers. Multimed Tools Appl 76(8):10,701–10,719
29. Masi I, Chang FJ, Choi J, Harel S, Kim J, Kim K, Leksut J, Rawls S, Wu Y, Hassner T et al (2018) Learning pose-aware models for pose-invariant face recognition in the wild. IEEE Trans Pattern Anal Mach Intell
30. Omkar P, Vedaldi A, Zisserman A (2015) Deep face recognition. In: Proceedings of British machine vision conference, vol 1, p 6
31. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis (IJCV) 115(3):211–252
32. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of CVPR, pp 815–823
33. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. CoRR arXiv:1409.1556
34. Sun Y, Wang X, Tang X (2014) Deep learning face representation from predicting 10,000 classes. In: Proceedings of CVPR, pp 1891–1898
35. Sund T, Moystad A (2006) Sliding window adaptive histogram equalization of intra-oral radiographs: effect on diagnostic quality. J Dentomaxillofac Radiol 35(3):133–138
36. Tadas B, Robinson P, Morency LP (2013) Constrained local neural fields for robust facial landmark detection in the wild. In: Proceedings IEEE international conference on computer vision workshops, pp 354–361
37. Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: closing the gap to human-level performance in face verification. In: Proceedings of CVPR, pp 1701–1708
38. Tan X, Triggs B (2010) Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans Image Process 19(6):1635–1650
39. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of CVPR
40. Yang J, Ren P, Zhang D, Chen D, Wen F, Li H, Hua G (2017) Neural aggregation network for video face recognition. In: International conference on computer vision and pattern recognition, vol 4, p 7
41. Yi D, Lei Z, Liao S, Li SZ (2014) Learning face representation from scratch. arXiv:1411.7923
42. Yin X, Yu X, Sohn K, Liu X, Chandraker M (2017) Towards large-pose face frontalization in the wild. In: Proceedings of the international conference on computer vision, pp 1–10
43. Zhao J, Cheng Y, Xu Y, Xiong L, Li J, Zhao F, Jayashree K, Pranata S, Shen S, Xing J et al (2018) Towards pose invariant face recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2207–2216
44. Zhou E, Cao Z, Yin Q (2015) Naive-deep face recognition: touching the limit of lfw benchmark or not? arXiv preprint arXiv:1501.04690
45. Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: Proceedings of CVPR, pp 2879–2886

Niraimathi Puhalanthi has received the B.E degree in computer science and engineering from Anna University, Chennai, India, in 2008, the Master degree in Information and communication engineering from National Taipei University, New Taipei City , Taiwan, in 2017. Niraimathi was the recipient of Gold medal for academic excellence in her B.E program, and was also the recipient of Phi Tau Phi award in her Masters. Her research interests include Deep learning, Computer vision, and Optical Inspection.

Daw-Tung Lin received the B.S. degree in control engineering from National Chiao-Tung University, Hsin Chu, Taiwan, in 1985, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland at College Park, MD, U.S.A., in 1990 and 1994, respectively. He served as the Acting President and Vice President for Academics and Research of National Taipei University from 2016 to 2017 and from 2015 to 2017, respectively. From 2011 to 2015, he served as the Dean of Academic Affairs of National Taipei University, Taiwan. In this role, he collaborated with faculty, deans and University administrators to lead the development of University strategies for engagement in faculty teaching excellence and students learning assurance. While helping to steward academic programs already under way, he created/supported a growing number of industrial partnerships and education opportunities. He earned and ran the University Teaching Excellence Program, which awards over NT$60,000,000 each year to provide faculty and students in key positions with teaching improvement, assurance of learning, innovation and entrepreneurship integration, global career navigation, interdisciplinary professional elite developing, holistic thinking and action cultivation, and global mobility and collaborations.

Previously he served as the Director of the Computer Center with Chung Hua University from 2001 to 2003, the Dean of the College of Engineering with Chung Hua University from 2003 to 2005, the Chairperson of the Department of Computer Science and Information Engineering with National Taipei University from 2006 to 2009, the Director of the Graduate Institute of Communication Engineering with National Taipei University from 2010 to 2011, and the Dean of Academic Affairs of National Taipei University from 2011 to 2015. From 1995 to 2004, he was an Associate Professor of the Department of Computer Science and Information Engineering at the Chung Hua University, Taiwan. He has been with the National Taipei University since 2005, where he became a tenured Professor of computer science and information engineering in 2009. He is a senior member of IEEE. He has been a regular contributor to the literature in computer vision and image processing. His research interests include image processing, computer vision, pattern recognition, and intelligent surveillance.

Affiliations

Niraimathi Puhalanthi[1] · Daw-Tung Lin[1]

Niraimathi Puhalanthi
niraingt@gmail.com

[1] Department of Computer Science and Information Engineering, National Taipei University, New Taipei City, Taiwan