IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 May 2024, accepted 26 May 2024, date of publication 29 May 2024, date of current version 5 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3406911

RESEARCH ARTICLE

LS-SIFT: Enhancing the Robustness of SIFT During Pose-Invariant Face Recognition by Learning Facial Landmark Specific Mappings

SHINFENG D. LIN, (Senior Member, IEEE), AND PAULO E. LINARES OTOYA, (Member, IEEE)
Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan
Corresponding author: Shinfeng D. Lin (david@gms.ndhu.edu.tw)

ABSTRACT The proper functioning of many real-world applications in biometrics and surveillance depends on the robustness of face recognition systems against pose, and illumination variations. In this work, we employ ensemble systems in conjunction with local descriptors to address pose-invariant face recognition (PIFR). Facial landmarks are detected during the first step with a two fold usage. The landmark locations are employed to perform head pose classification (HPC). HPC allows to select only the visible landmarks for further processing. Then, local descriptors are extracted from the selected landmarks within a face image. We are proposing a novel learned descriptor (LS-SIFT) to overcome the robustness limitations of SIFT against large viewpoint variability during face recognition. Second, the extracted descriptors are used to train the base learners comprising an ensemble system for each subject in a face database (one ensemble per subject, one base learner per landmark). A novel GMM-based base learner model, named Mahalanobis Similarity (MS), is introduced in this work. Finally, face recognition is performed based on the ensemble systems' outputs from all the subjects in a face database. During the experimental trials, SIFT and LS-SIFT are employed individually for local feature extraction, whereas GMM and MS are used to build the ensemble systems, in an independent manner, for further comparison. The whole PIFR system is tested on CMU-PIE, Multi-PIE, and FERET databases, outperforming most of the state-of-the-art works regarding images with pose angles in the range of $\pm 90^o$.

INDEX TERMS Facial landmarks, local feature extraction, head pose description, ensemble learning, face recognition.

I. INTRODUCTION

The role of face recognition (FR) in real-world applications, including biometric authentication, surveillance, entertainment, and human-computer interaction, has expanded significantly in recent years [1], [2]. Significant progress has been made in automatic face recognition, with promising results achieved in both controlled and uncontrolled environments. However, face recognition still faces challenges due to wide variations in pose, illumination, and expression commonly encountered in real-world images [3], [4]. Face recognition can be approached as either an identification or a verification problem. Face identification refers to the 1:N matching problem involving the comparison of an unknown face with all the faces in a known identity database and making a decision based on the comparisons. This task is considered closed-set if the person is known to be in the database. On the other hand, face verification is the 1:1 matching problem, where the identity of a query face is accepted or rejected by comparing it with the face data of the claimed identity in the database [1].

In recent years, authors have shifted their attention towards using a holistic approach (a face image is cropped and processed as a whole) leveraging the power of deep CNNs (DCNN) [5], [6], [7]. This approach comprises three well-defined components. The first one involves the design of

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

a DCNN capable of extracting a discriminative (invariant) face embedding, which can be used later for face verification or identification. During the training of these DCNNs, a loss function (e.g. contrastive loss, triplet loss) must be defined such that the inter-class similarity is minimized while maximizing the intra-class similarity. FaceNet [5], ResNet-100, and ResNet-50 are some examples of these DCNNs with promising results. The second component entails the design of a loss function specifically tailored to achieve a high accuracy on face recognition. Some renowned works on this component include CosFace [6], and ArcFace [7]. The last component involves the manner face embeddings are employed to perform face recognition (e.g. classification, feature matching).

In the present work, we propose to address pose-invariant face recognition (PIFR) by using an ensemble learning approach and local descriptors. In this approach, the descriptors are computed from facial landmarks which are considered as keypoints. A novel local descriptor, named LS-SIFT, is introduced in this work. The main advantage of this descriptor over SIFT is its higher robustness against large viewpoint variations. On the other hand, a novel unsupervised base learner model, named Mahalanobis Similarity, is developed in this work. Its superior performance over other clustering base learner models (like GMM) is demonstrated experimentally.

Face recognition is carried out as follows. First, facial landmarks are detected in the input face image, and head pose classification (HPC) is performed from the landmarks' locations. This technique was introduced in our previous works and its relevance in achieving a high recognition performance was demonstrated. Second, local descriptor extraction is performed over the detected landmarks. During the training stage of the ensemble systems, several points surrounding a specific landmark are employed for feature extraction. The extracted descriptors (SIFT, LS-SIFT), associated with a landmark, are used to train the base learner corresponding with the same landmark (i.e. we are proposing a linkage between a landmark and a base learner). Thus, the feature descriptors are used to train the base learners comprising an ensemble system for each subject in a face database.

During the testing stage, feature extraction is performed only on the visible landmarks, and the obtained descriptors are fed to their corresponding base learners for all the subjects. Lastly, the ensemble systems' outputs (representing the similarity value between the input face image and the subject the system was trained for) are gathered, and the identity of the input face is obtained by choosing the ensemble system whose output value is the highest. For the purpose of performance assessment, we are employing CMU-PIE, Multi-PIE and FERET as the benchmarks. The results are compared with state-of-the-art methods to show a surpassing performance over most of them, especially for face images depicting extreme poses (close to $\pm 90^o$). In summary, the contributions of this paper are listed as follows:

- The potential of extracting local features from facial landmarks during face recognition is showcased in the present work. Indeed, we propose a novel landmark-centered ensemble learning framework where a base learner is exclusively trained with features obtained from the regions surrounding its corresponding facial landmark.
- We propose an enhanced version of the conventional Scale Invariant Feature Transform (SIFT), tailored for face recognition. The improvement of this local descriptor consists of learning landmark-specific mappings (MLPs are employed) which aim to reduce the variability between features extracted from the same landmark but at different view-points (pose-robust).
- Two base learner models are proposed to construct the PIFR ensemble systems. The first model is GMM. The novelty lies in the way GMM is trained (it is used to cluster features from the same class) not in the algorithm itself. The second one is a novel GMM-based model called Mahalanobis Similarity (MS). Its performance during PIFR trials showed to be considerably superior than GMM.
- Experimental trials on both face identification and verification are conducted on CMU-PIE, Multi-PIE and FERET. Conversely to adopting only the Rank-1 accuracy (recognition rate), as it is done in previous works using these datasets, we adopted the TAR@FAR and Rank-N accuracy metrics (used in the latest works on FR).

The subsequent sections of this paper are organized as follows. In Section II, we present the related work on head pose description using facial landmarks, PIFR from a holistic approach, and the use of local features to improve the performance during face recognition. In Section III, the proposed methodology is detailed. The experimental trials, results, and comparisons with state-of-the-art methods are detailed in Section IV. Finally, the paper is concluded in Section V.

II. RELATED WORK

A. WORK ON HEAD POSE DESCRIPTION USING FACIAL LANDMARKS

As it was mentioned in the introduction, the proposed PIFR framework performs Head Pose Classification (HPC), a simplified version of HPE at a coarse level, to select which facial landmarks are not occluded in the image due to pose variations. Furthermore, the proposed HPC model requires a head pose representation (i.e. a descriptor vector). Thus, it is worth reviewing some current works employing facial landmarks to compute a head pose descriptor during HPE. Abate et al. [8] provides a brief taxonomy on HPE where HPE works are categorized into those using RGB or RGB-D images, and whether the technique requires a training stage or not. According to [8], 2D landmark locations are processed to compute a quadtree-based pose descriptor, and

HPE is performed upon this descriptor without a training stage. The proposed quadtree-based descriptor is a binary vector representing whether a facial landmark is located within a box (i.e. a square region withing the image) or not. Even though the number of boxes generated during the quadtree computation might differ among different face images, the binary descriptor has a fixed length. HPE is performed in two stages. During the first stage, a synthetic face dataset (entailing face images at different poses with a precise ground truth annotation) is generated with pose ranges $\pm 45^o$ yaw, $\pm 30^o$ pitch, and $\pm 20^o$ roll, in steps of 5^o (2223 images in total). Furthermore, one pose descriptor is computed and stored for each generated image (associated with a head pose annotation) in the face database. The outcome of the first stage is a gallery of head pose values and their corresponding descriptors. In the second stage, an input face image is processed to obtain its binary pose descriptor, and HPE is performed by comparing the descriptor with all the descriptors in the gallery, choosing the pose whose associated descriptor has the highest similarity. BIWI and AFLW datasets were employed as benchmarks, achieving a mean MAE of 5.69^o and 7.45^o respectively.

Another landmark-based head pose descriptor is proposed in [9]. The elements of this descriptor consist of the landmark locations after a two-step normalization against scale and translation. The first normalization is performed by scaling the raw landmark locations by the inter-ocular distance, and regarding the center between the eyes as the origin of the 2D coordinate system. The second normalization step is performed by subtracting the mean of the descriptor, obtained after the first normalization, and scaling it by the standard deviation inverse. HPE is performed by training a Support Vector Regression (SVR) model with the vectors obtained after applying the proposed descriptor to a HPE database. In order to validate the proposed approach, the authors developed a synthetic HPE dataset (SyLaHP), covering pose ranges of $\pm 90^o$ yaw, $\pm 70^o$ pitch, $\pm 55^o$ roll. The Area under the Curve (AUC) of the HPE accuracy curve was employed for performance assessment instead of pose MAE. The experimental results of using different landmark detection models on BIWI and SyLaHP evidenced two facts. First, the proposed method is suitable for HPE even under extreme poses. Second, the method is very sensitive to the landmark detection model's accuracy.

B. WORK ON PIFR FROM A HOLISTIC APPROACH

Currently, PIFR is being addressed from a holistic insight with the aid of deep CNNs. Even though the method presented in the current article adopts a local approach, it is worth to summarize the latest progress done on PIFR from a holistic point of view.

As it was mentioned above, applying face normalization to convert a pose-view face into a near-frontal one is a well-known pre-processing technique during FR. However, frontalizing a face image depicting extreme poses usually generate image artifacts or non-acceptable face images. To address this issue, An et al. [10] proposed a novel pose-specific face normalization (Adaptive Pose Alignment or APA), which works in conjunction with a DCNN-based face representation (SENet50 and LResNet100-IR were adopted as the backbones) to improve PIFR performance under challenging conditions. The authors justified the suitability of their adaptive face normalization technique by arguing that a proper face alignment technique may reduce the intra-class variability during FR, and also accelerate the training of the DCNN used for face feature extraction. Previous works on face alignment relied on a single canonical head template (in 2D or 3D) used to apply an affine transformation to the input pose-view image. The proposed APA method belongs to the template-based group, with the difference that 4 learned pose-specific face templates are employed instead of using a non-learned one. The whole face alignment method entailed three steps (head pose estimation, adaptive templates generation, face alignment). During face alignment, an input face image is matched to a face template according to its head pose, and a 2D affine transformation is applied. The CASIA-WebFace dataset was employed for learning the 4 face templates, while VGGFace2 and MS1M were employed to train the face representation DCNN, with SoftMax and ArcFace as the loss functions. Experimental trials on face verification and identification were conducted on 4 datasets (IARPA Janus Benchmarks IJB-A, IJB-C, LFW, CPLFW). The results, in terms of the TAR@FAR and Rank-N accuracy, revealed that the addition of the proposed alignment method did improve the performance on PIFR, achieving the best results when utilized with LResNet100-IR + ArcFace.

Another work using a holistic approach, in conjunction with face frontalization, was conducted in [2]. Petpairote et al. [2] stated that a strict condition for a PIFR algorithm to be used in real-world scenarios is to require only one frontal image per subject within the face gallery. To achieve this, a face frontalization technique is proposed as a pre-processing step. This technique consists of generating a face shape (represented by normalized landmark locations) database. Each entry of this database contains the shape of a frontal face image (0^o) plus 6 pose-view ($\pm 45^o$ in steps of 5^o) face shapes obtained from a specific subject (one entry per subject). Once the mentioned database is generated, face frontalization is conducted in three steps. First, the shape of an input face image is computed with the aid of face and facial landmark detection models. Second, the shape is matched with the pose-view shapes of each entry contained in the face shape gallery to find the closest one. Third, a piece-wise 2D warping (PAW) operation (landmark locations are regarded as vertices of a face mesh) is applied to the input image by employing the matched pose-view shape and its frontal counterpart.

A post-processing step was also proposed in [2] to enhance the results of the PAW by taking into account the level of self-occlusion due to pose variations. However, it requires a manual fine tuning of several parameters to

work properly, making it less practical. After the image has been frontalized, 3 different traditional global face representations (Local Gabor Binary Patterns, PCA, LBP) are applied independently, and their histograms are used as feature vectors. The whole methodology was tested on Multi-PIE and CMU-PIE databases, with recognition rates of 97.18% and 97.65% respectively. Even though the results outperformed some previous works, the images used for testing just covered the range of $[-45^o, +45^o]$. Besides, the performance is severely affected when tested with faces at extreme poses or with slight occlusion (e.g. use of glasses).

C. WORK ON PIFR AIDED BY LOCAL FEATURES

Some recent works on PIFR has demonstrated the potential of extracting local information from face images, instead of computing a unique global discriminative face feature. Indeed, the use of locally extracted data can outperform its holistic counterpart in scenarios depicting a high level of occlusion (e.g. masked face recognition) [11].

Lin et al. [11] stated that most of the DCNN-based face detection or recognition models are trained with non-occluded images. Thus, their performance is poor during masked face recognition. To address PIFR under this scenario, a learned face similarity score is proposed in [11]. This similarity score is obtained after training a DCNN (ResNet-101) to compute face embeddings at a global and local level. For the global part, a face angular classification loss (CosFace, ArcFace) is adopted. On the other hand, the authors proposed the addition of a patch-based local consistency loss to the total loss function, to give more emphasis to the features obtained from non-occluded patches. The whole FR framework was tested on both non-occluded datasets (VGG2-FP,AgeDB-30, CALFW, CFP-FF, CFP-FP, LFW) and a synthetically occluded database (masked LFW). The MS-Celeb-1M dataset was employed during the training stage. According to the experimental results, the use of ArcFace in conjunction with the proposed local consistency loss yielded the best results in all the regarded datasets. The authors concluded that the utilization of features extracted in a local manner leads to a better performance in PIFR under both masked and non-masked conditions.

One of the main application areas of face recognition is law enforcement. Indeed, Lai et al. [12] stated that global DCNN-based FR models cannot achieve a high precision during face verification when the images to be compared are taken from similar people (e.g. twins), or when the person of interest wears a mask. Besides, the authors pointed out that high-resolution face images can be obtained nowadays at a low cost. Thus, adopting a fully holistic approach for FR might lead to a considerable loss of relevant facial information present in this kind of images. With these considerations in mind, a local learned descriptor (PoreNet) is proposed in [12] to extract facial information from patches centered at face pores (e.g. wrinkles, pores, moles). In order to train PoreNet, a multi-scale pore detector based on Laplacian of Gaussian (LoG) is proposed. After localizing the pores, image patches are obtained from the original images (the patch size depends upon the pore's scale). The patches are concatenated with maps containing the (x, y) pixel positions within each patch, and then they are resized to a dimension of $5 \times 42 \times 42$. During the training stage, the Bosphorus dataset was utilized, considering patches obtained from 4 pose-view images/subject and 75 subjects. The experimental trials on FR were conducted on Bosphorus and Multi-PIE, adopting the Equal Error Rate (EER) as the performance metric. Again, the experimental results demonstrated that local descriptors can attain a better performance in FR under both occluded and non-occluded scenarios compared to global descriptors.

III. PROPOSED METHODS

A. HEAD POSE DESCRIPTION AND CLASSIFICATION

In our previous works [13], [14] we introduced a head pose descriptor called Face Angle Vector (FAV). The elements comprising the FAV are the angles between 12 mouth landmarks and the eye centers, totaling 24 elements. In order to compute this vector, 24 out of 68 facial landmarks (landmarks in the eyes and mouth) are detected within a face image by using the Facial Alignment Network (FAN) [15]. This vector is employed later to classify an input face image according to its pose angle into N_{pose} classes such that only the visible landmarks, corresponding to an specific pose class, are processed in the remaining steps of the PIFR system. In order to perform this classification task, a SVC model is utilized, and the number of pose classes is set to $N_{\text{pose}} = 5$. The performance of this classifier achieves an accuracy of 0.988, while the average of its F-1 scores is 0.986.

B. LEARNED MAPPING FOR ENHANCING THE ROBUSTNESS OF SIFT AGAINST POSE AND ILLUMINATION VARIABILITY (LS-SIFT)

SIFT has been widely used for performing PIFR. However, SIFT is not invariant to all kinds of affine transformations [16]. Indeed, SIFT it is not invariant to large viewpoint variations. Additionally, it is robust, yet not invariant, to illumination changes. These drawbacks lead to getting a limited accuracy when developing a PIFR system relying on SIFT. Considering these drawbacks, a non-linear vector transformation (i.e. mapping) for SIFT vectors extracted from a specific facial landmark is proposed in this work to improve the SIFT robustness against head pose and illumination variations. A more robust local descriptor obtained from a SIFT vector is proposed in this work, called Landmark-specific SIFT (LS-SIFT).

The network architecture depicted in Figure 1 is employed to learn LS-SIFT from SIFT descriptors obtained from a specific landmark. The working principle of this network is explained as follows. The network is actually trained to perform a classification task given SIFT descriptors obtained from the l^{th} landmark (considering a 24-landmark scheme) as input. In this classification environment, the universe of

TABLE 1. Description of the parameters employed in the LS-SIFT network architecture.

Layer	Type	Input Dropout	Output size	Activation function	Input Bias	Kernel Regularizer	Kernel Constraint
Input	Input layer	-	128	-	-	-	-
FC1	Fully connected	✗	256	ReLu	✓	✗	✗
FC2	Fully connected	0.2	512	ReLu	✓	L1 (10^{-3})	✗
FC3	Fully connected	0.3	1024	Swish	✓	L2 (10^{-3})	✗
FC4	Fully connected	0.3	512	ReLu	✓	L2 (10^{-3})	✗
FC5	Fully connected	0.3	200	ReLu	✓	L2 (10^{-3})	✗
Output	Fully connected	✗	$\lvert\mathcal{S}_{\text{LM}}\rvert$	Softmax	✗	✗	MinMaxNorm ($[0.8, 1.5]$)

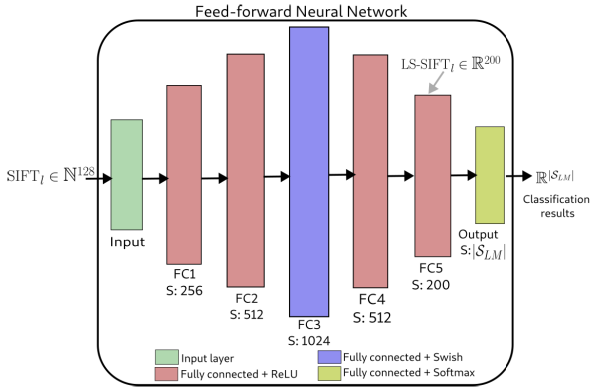

FIGURE 1. Network architecture used to compute LS-SIFT.

subjects comprising the training dataset is denoted as $\mathcal{S}_{LM}$, where $s \in \mathcal{S}_{LM}$ represents the collection of sample images corresponding to a specific subject. Furthermore, the output of the proposed network is a vector in $\mathbb{R}^{|\mathcal{S}_{LM}|}$ containing the classification results (i.e. the probabilities of the input vector SIFT$_l$ to match with a specific subject in $\mathcal{S}_{LM}$). However, the main purpose of this network is not to obtain the classification results, but to learn the non-linear transformation from the input layer to the second-to-last layer (FC5 in Fig. 1) which maximizes the accuracy of the classification results (i.e. the classification accuracy is just the objective function to be maximized during the network training).

The images employed to train the proposed network are collected from the CMU-PIE and Multi-PIE databases. Indeed, 20 subjects from the CMU-PIE are randomly selected and their images with neutral expression, 9 poses, and ambient illumination (9 images per subject) are added to the image training dataset. On the other hand, 80 subjects from the Multi-PIE are randomly selected, and their images with neutral expression, 13 poses, and 4 different illumination conditions ($13 \times 4 = 52$ images per subject) are adjoined to the image training dataset. Thus, a total of $|\mathcal{S}_{LM}| = 100$ subjects are considered for training the proposed network, with a total of $\sum_{s \in \mathcal{S}_{LM}} \langle s \rangle = 4340$ images.[1] Once the training dataset of images has been defined, the next step involves processing the dataset to compute the SIFT descriptors extracted from the detected facial landmarks, as illustrated in Algorithm 1 (lines 8-19). The face images $\mathcal{I}$ of the subject s are obtained with the function `getImages()`. Each image $I \in \mathcal{I}$ is processed to locate the facial landmarks. If the l facial landmark is visible (i.e. it is not self-occluded by pose variations), then its location is regarded, and the SIFT feature description algorithm is applied to all its surrounding points p' in the image $\mathbf{I}$ within a radius r (see lines 11-14 in Algorithm 1). The resulting SIFT vectors $\vec{f}$ are added to the training data $\mathbf{X}$, and the function `toCategorical()` is utilized to create binary vectors required to train the proposed network with the Categorical Cross-entropy loss function [17]. The model $\mathcal{M}_l$ is trained, and added to the set of landmark-specific SIFT transformation models $\mathscr{M}$ (lines 20-21). The process is repeated for all the 24 landmarks considered in this work.

As can be seen in Fig. 1, the network proposed to compute LS-SIFT comprises 6 fully-connected layers (including the output layer). The parameters of each layer are detailed in the Table 1. The layers' output size increases until the layer FC3 (with an output size of 1024). These initial layers of the network, with an increased output size, act as feature extractors, detecting low-level patterns and structures in the SIFT descriptors obtained under different conditions of pose and illumination. It is worth mentioning that, the Swish activation function is employed for biggest layer (FC3), instead of the ReLu function employed in the other layers. The decision of using the Swish function is based on empirical trials, where the classification accuracy was higher than the one obtained with ReLu. The following layers (FC4, FC5) aim to perform dimensionality reduction, which helps to create a more robust and compact representation of the input data. Kernel regularization is included in some of the hidden layers to promote sparsity in the layer weights, and thus improving the generalization of the models. L1 regularization promotes sparsity in the kernel weights by forcing some weights to become exactly zero. L2 regularization, on the other hand, promotes smaller weights, preventing a single weight from dominating the whole learning process [18], [19].

In Table 1, the regularization methods are expressed as L1 or L2 followed by the penalty value in parentheses. Finally, the network is forced to concentrate all the relevant

[1]The operator $\langle . \rangle$ represents the quantity of images associated with a specific subject.

Algorithm 1 LS-SIFT Learned Mappings' Training

Input: An image dataset DB_{LM} corresponding to the subject set $\mathcal{S}_{LM}$

Result: A set of trained network models $\mathscr{M}$ (one model per facial landmark).

$\mathscr{M} \leftarrow \{\}$
for $(l = 0; l < 24; l++)$ **do**
  $\mathcal{M}_l \leftarrow$ *createMLPmodel()* `/* Create an instance of the network depicted in Figure 1 */`
  $\mathbf{X} \leftarrow [\,]$
  $\mathbf{y} \leftarrow [\,]$
  for $s \in \mathcal{S}_{LM}$ **do**
    id = *getID*(s) `// the id is an integer between` $[0, |\mathcal{S}_{LM}| - 1]$
    $\mathcal{I} \leftarrow$ *getImages*(s, DB_{LM})
    for $I \in \mathcal{I}$ **do**
      if *isVisible*$(\mathbf{I}, l)$ **then**
        $p =$ `getLandmarkLocations`$(\mathbf{I}, \{l\})$
        $P' =$ `surroundingPoints`(p, r)
        foreach $p' \in P'$ **do**
          $\vec{f} =$ `ComputeSIFT`$(\mathbf{I}, p')$
          $\mathbf{X}$.append$(\vec{f})$
          $\mathbf{y}$.append(`toCategorical`(id,$|\mathcal{S}_{LM}|$)
        end
      end
    end
    $\mathcal{M}_l$.*train*$(\mathbf{X}, \mathbf{y})$
    $\mathscr{M} \leftarrow \mathscr{M} \cup \{\mathcal{M}_l\}$
  end
end
Return $\mathscr{M}$;

information involved in the classification accuracy towards the FC5 layer. This is accomplished by removing the input bias from the Output layer, and establishing the MinMaxNorm kernel constraint. This constraint establishes a lower bound and an upper bound to the norm of the weights incident to each hidden unit. In practice, it has been observed that setting a norm boundary of [0.8, 1.5] leads to improved results when utilizing the suggested SIFT mappings for PIFR. In order to train the model, the Adam optimizer is selected, with the Categorical Cross-entropy as the loss function to be minimized, and the accuracy as the metric to monitor the model performance (it must be maximized).

C. FACIAL LANDMARK-CENTERED LOCAL FEATURE EXTRACTION

The procedure carried out for describing the facial information surrounding a specific landmark on a face image is defined as facial landmark-center feature extraction. This procedure is based on the feature extraction of a generic point (i.e. a pixel comprising its intensity and location) within a digital image. In this work, two local descriptors are considered for feature extraction. The first local descriptor is SIFT. In this work, we do not employ the SIFT keypoint detector. But, a generic point is converted into a SIFT keypoint by specifying the point location and its diameter [20]. This process is carried out by the function `ComputeSIFT()`, defined in Algorithm 2. The second descriptor is LS-SIFT. As it was previously described, LS-SIFT is a variant of SIFT, tailored for face recognition, which aims to perform a non-linear transformation to the SIFT descriptor obtained from a specific facial landmark such that the resulting descriptor vector has an enhanced robustness against viewpoint, and illumination variations. The computation of LS-SIFT is detailed in the function `ComputeLS-SIFT()` of Algorithm 2. Additionally to the specification of the location of a pixel position (i.e. landmark location) p in an image I, its landmark index l (according to the 24-landmark scheme) must be also specified, so that the landmark specific non-linear transformation $\mathcal{M}_l$ is applied to the SIFT descriptor $\mathbf{f} \in \mathbb{N}^{128}$ to obtain the LS-SIFT descriptor $\mathbf{f}_l \in \mathbb{R}^{200}$.

D. FACE RECOGNITION USING ENSEMBLE SYSTEMS

A generic ensemble system used for classification has three essential components. The first component of an ensemble system is the set of base learners, which are individual weak classifiers. Base learners receive input data and compute classification decision values. The output can be either discrete or continuous, depending on the employed classification model (e.g. Support Vector Regression, Artificial Neural Networks, Naive Bayes). The second component focuses on base learner training methods. The selection of a specific training method depends on the underlying approach used to obtain the final classification result from the ensemble system. Common training methods include Bagging, Boosting, and AdaBoost [21], [22]. These methods determine how the base learners are trained individually (or collectively), and how the output of each base learner contributes to the overall ensemble's classification performance. The third and final component is the combination rule (e.g. mean rule, product rule), which determines how the outputs from the base learner set are integrated to obtain the ensemble support value.

The conventional way of using ensemble learning states that the base learners comprising an ensemble system must be trained with data obtained from samples corresponding to different classes (e.g. subjects) under a classification environment. Furthermore, an ensemble system can be built at 4 different levels (combination level, classifier level, feature level, data level) [22]. When the feature level is chosen, the base learners are trained with different feature subsets from the feature space. We are proposing to implement an ensemble system at a feature level (each base learner is trained exclusively with features extracted from a specific facial landmark). According to the conventional ensemble learning approach, the base learners still need to

Algorithm 2 Facial Landmark-Centered Local Feature Extraction From a Face Image

Input: A gray scale face image $\mathbf{I}$, set of non-occluded landmark indexes $\mathscr{L}$
Result: Set of features extracted upon facial landmarks $\mathcal{F}$.
Data: Descriptor parameters:
- SIFT: σ_{sift}, $\emptyset_{\text{sift}}$
- LS-SIFT: Facial landmark index l

$\mathbf{P}$ = `getLandmarkLocations`$(\mathbf{I}, \mathscr{L})$ // $p_l \in \mathbb{N}^2; \forall p_l \in \mathbf{P}$
$\mathcal{F} \leftarrow \{\}$ // The set of extracted local descriptors is initialized as an empty set.
for $l \in \mathscr{L}$ **do**
  f_l = `computeDescriptor`$(\mathbf{I}, p_l \in \mathbf{P}, l)$
  $\mathcal{F} = \mathcal{F} \cup \{f_l\}$
end
Return $\mathcal{F}$
/* The following functions are employed for feature extraction. The function computeDescriptor() is replaced by any of these ones. */
Function `ComputeSIFT`$(\mathbf{I},p,l)$**:**
  kp = `getKeypoint`$(p,\emptyset_{\text{sift}})$ // see [20]
  $\mathbf{f}$ = `getSIFTDescriptor`$(\mathbf{I},kp,\sigma_{\text{sift}})$
  Return $\mathbf{f} \in \mathbb{N}^{128}$
End Function

Function `ComputeLS-SIFT`$(\mathbf{I},p,l)$**:**
  $\mathbf{f}$ = `ComputeSIFT`$(\mathbf{I},p)$
  $\mathbf{f}_l = \mathcal{M}_l(\mathbf{f})$ // Apply landmark-specific feature mapping
  Return $\mathbf{f}_l \in \mathbb{R}^{200}$
End Function

be trained with features extracted from different classes (i.e. subjects), and their outputs should be vectors in $\mathbb{R}^C$ (C is the number of classes), which are combined later to compute the final classification result.

A different base leaner training method, called *"Input Decimation Ensembles"* is proposed in [23]. This method generates different subsets of training features for each classifier within its ensemble. This strategy involves training each base classifier (one for each class) with a distinct feature subset, effectively reducing correlations among the classifiers. Input decimation reduces the classification results correlation between the classifiers by selectively training them with features that are highly correlated with a particular class (i.e. the base learner associated with a class is primarily trained with data acquired specifically from that particular class). During classification, the output of each base learner is not a vector but a value between $[0, 1]$, and the classification is performed by choosing the class associated with the base learner with the highest output value.

In our previous work [14] we adopted a similar ensemble learning insight as in [23] to perform PIFR. Indeed, in the present work we are employing an ensemble learning framework (see Fig. 2) similar to the one adopted in our previous work [14]. In this framework, the input face image undergoes processing to determine the face bounding box and facial landmark locations. Subsequently, the input face image and landmark locations are separately processed by two distinct blocks. The first block performs Head Pose Classification (HPC), as discussed earlier. Based on the computed face pose class from the HPC block, the Base Learner Selection (BLS) block determines which base learners will be utilized in the ensemble system. In fact, the base learners are selected according to the computed pose class (see Fig. 4) given that some of them might be occluded due to the head pose. Each base learner in this study is associated exclusively with a specific facial landmark, following a 24-landmark scheme (i.e. an ensemble system comprises 24 base learners). Consequently, once the selected base learner subset is determined, the BLS block also specifies the landmarks that should be described in the face image. The information regarding the selected base learners and corresponding landmarks is then passed to the block in charge of performing feature extraction. This block utilizes the received information and computes the descriptors for each selected landmark from the input face image.

For the purpose of this work, a base learner $\beta_{p,l}$ is a model expert in performing face recognition in a not very accurate way (i.e. a weak face recognition model). This base learner processes the feature descriptor vector $\vec{f_l}$ obtained from the l^{th} facial landmark of the p^{th} subject, on a face database (i.e. a base learner is linked to a facial landmark), to compute its decision support $d_{p,l}$, representing the "likelihood" that $\vec{f_l}$ corresponds to the subject p. In this work, two different models are proposed as base learners to be used independently for further comparisons. The first model is Gaussian Mixture Models (GMM). GMM is a parametric probability density function (PDF) represented as a weighted sum of Gaussian component PDFs as defined in (2). Where $\mathcal{N}(\vec{x}|\lambda_i)$ is the normal PDF (defined in (1)) with a mean vector $\vec{\mu}_i \in \mathbb{R}^D$, and a covariance matrix $\Sigma_i \in \mathbb{R}^{D\times D}$. The main goal of GMM is to cluster a collection of training vectors $\mathbf{X}$ (i.e. assign a vector to a specific Gaussian component PDF), such that the likelihood[2] of observing $\mathbf{X}$, given a specific set of GMM parameter values λ, is maximized. Given that the likelihood, defined in (3), is a product of F scalars within the interval $[0, 1]$, its value converges quickly to zero. Therefore the log-likelihood $L(\mathbf{X}|\lambda) \in (-\infty, 0]$ is preferred to be used during the estimation of the parameters λ. The optimal parameters λ^* are obtained by using the Expectation-Maximization algorithm with the log-likelihood as the objective function to maximize.

[2]The likelihood measures how well $\mathbf{X}$ fits the GMM distribution.

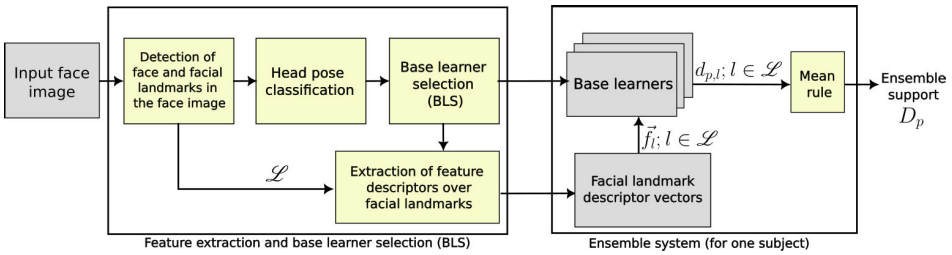

FIGURE 2. Workflow of the proposed ensemble learning based PIFR framework. ***Data or objects are represented as gray blocks while operations are depicted in yellow.***

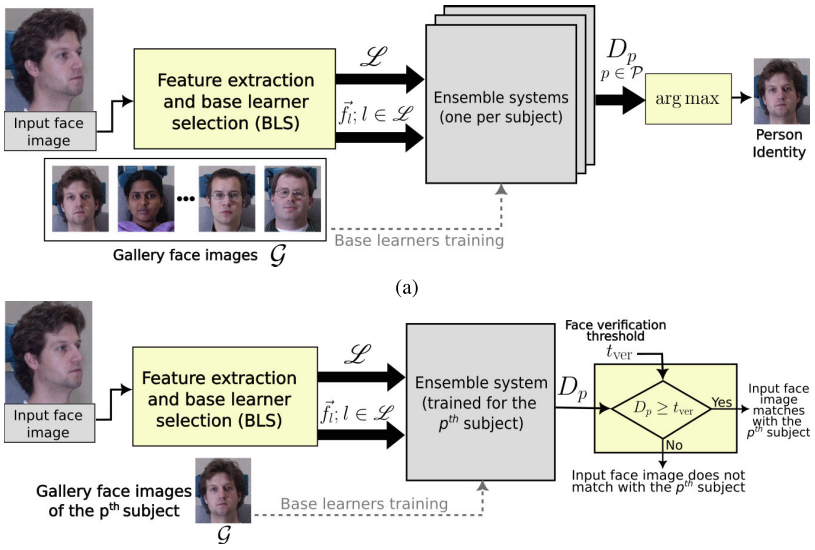

FIGURE 3. Usage of the proposed PIFR framework: (a) During 1:N face identification (b) During 1:1 face verification.

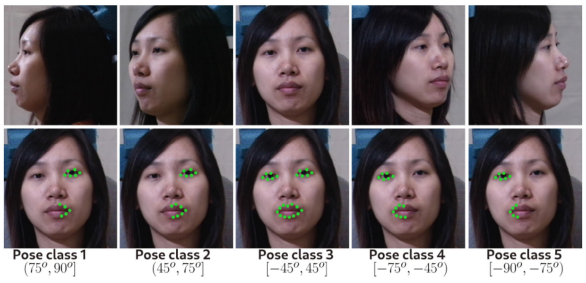

FIGURE 4. Graphical representation of BLS for different pose classes. ***The selected base learners (visible or non-occluded ones) are depicted as green circles for each pose class.***

For the purpose of this work, a GMM model is trained as detailed in Algorithm 3.[3] Once the landmark l is detected in the training face image (gallery image $I \in \mathcal{G}_p$), feature extraction is performed over the points surrounding the landmark location within a radius r (lines 6-15). The extracted features (i.e. descriptors) are arranged in a matrix $\mathbf{X}$, and the GMM model is fitted with $\mathbf{X}$ as the training data (line 16). Finally, the fitted GMM is stored in the base learner object corresponding to the landmark l (line 17). After training a base learner with a GMM model, its decision support $\delta_{\text{GMM}} \in (-\infty, 0]$, defined in (4), is the log-likelihood that the landmark descriptor vector $\vec{f}_l$ obtained from the l^{th} landmark of an input face image corresponds to the subject the GMM model was trained for. Where λ^* are the GMM model's parameters (weights, covariance matrices, mean vectors) after fitting it to $\mathbf{X}$. In this work, the GMM model comprises $n_{\text{comp}} = 3$ Gaussian components, and all the components share the same covariance matrix[4]

$$\mathcal{N}(\vec{x}|\vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}\|\Sigma\|^{1/2}} \exp\{-\frac{1}{2}(\vec{x}-\vec{\mu})^{\intercal}\Sigma^{-1}(\vec{x}-\vec{\mu})\} \tag{1}$$

$$p(\vec{x}|\lambda) = \sum_{c=1}^{n_{\text{comp}}} w_c \mathcal{N}(\vec{x}|\vec{\mu}_c, \Sigma_c);$$

[3]Pseudocode is employed to describe base learner training, instead of figures, to clarify the mathematical notations used along the paper, avoiding misconceptions.

[4]According to our experimental trials, using one covariance matrix per component slightly improves the performance on PIFR than using a shared covariance matrix. Moreover, the testing time increases by a factor around 4 when using one covariance matrix per component.

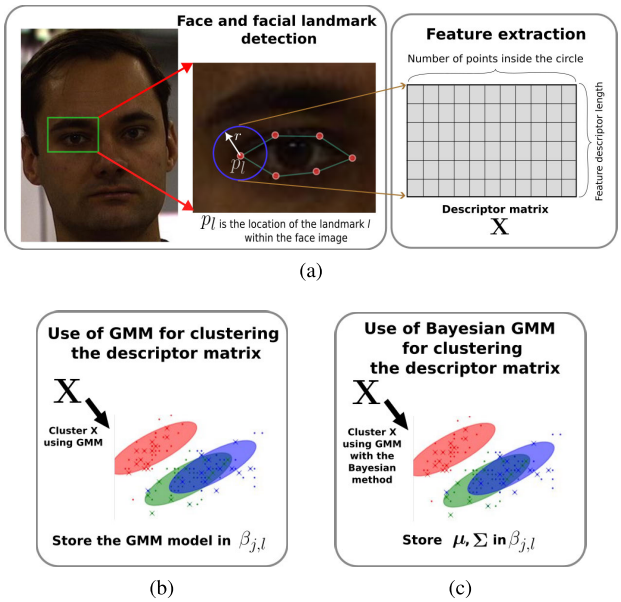

FIGURE 5. Graphical representation of the training process for the proposed base learners: (a) Descriptor matrix **X** computation; (b) GMM model training; (c) MS model training. *A circle in blue indicates the surrounding points of the l landmark within a radius r.*

$$\text{s.t. } \lambda = \{w_c, \vec{\mu}_c, \Sigma_c\}\ \forall c \in \{1, .., n_{\text{comp}}\} \tag{2}$$

$$p(\mathbf{X}|\lambda) = \prod_{f=1}^{F} p(\vec{x}_f|\lambda);\ \vec{x}_f \in \mathbf{X} \tag{3}$$

$$\delta_{\text{GMM}}(\vec{f}_l) = \ln p(\vec{f}_l|\lambda^*) \tag{4}$$

The second proposed base learner is called Mahalanobis Similarity (MS). It combines the advantages of GMM for data clustering, with the superiority of the Mahanalobis distance over other vector similarity metrics. In this case the Bayesian variant of GMM, called Bayesian GMM, is employed for data clustering due to the way the parameters λ^* are obtained. Under a Bayesian environment, the posterior likelihood, defined in (5), implies that a prior probability distribution must be specified for the GMM parameters λ. This prior introduces constraints or penalties to prevent overfitting and improve the clustering performance of the GMM model. In summary, the goal of the MAP[5] estimation is to find the parameter values λ^* that maximize the posterior probability $p(\lambda|\mathbf{X})$ given the observed data $\mathbf{X}$ and the prior distribution $p(\lambda)$. A full explanation of the Bayesian GMM underlying mathematical background is well explained in [24]. The training procedure, detailed in Algorithm 4, is very similar as for GMM. However, the fitted Bayesian GMM distribution is not stored in the base learner object as a whole. Instead, only the computed covariance matrices Σ_c, and means $\vec{\mu}_c$ are stored (see line 17 of Algorithm 4). Furthermore, the log-likelihood is not employed as the decision support for a trained Mahalanobis Similarity base learner. Instead, Σ_c and $\vec{\mu}_c$ are employed to compute the Mahalanobis distance of an input descriptor vector $\vec{f}_l$ to the closest component

[5]Maximum a-posterior estimation.

Algorithm 3 Base Learner Training Algorithm for GMM

Input: The training face image gallery $\mathcal{G}_p$ for the subject p, a predefined list of landmarks $\mathscr{L}$.
Result: The set of available trained base learners B_p for the subject p.
Data: Landmark neighborhood radius r

1: $B_p \leftarrow \emptyset$
2: **for** $(l = 0; l < 24; l++)$ **do**
3: $\quad \Phi \leftarrow$ `initGMMmodel()`
4: $\quad \beta_{p,l} \leftarrow$ `initBLModel`$(p, l,$ *type: GMM*$)$
5: $\quad \mathbf{X} \leftarrow [\,]$ `// initialize an empty descriptor matrix`
6: $\quad$ **foreach** $I \in \mathcal{G}_p$ **do**
7: $\quad\quad$ **if** `isVisible`$(\mathbf{I}, l)$ **then**
8: $\quad\quad\quad pnt =$ `getLandmarkLocations`$(\mathbf{I}, \{l\})$
9: $\quad\quad\quad P' =$ `surroundingPoints`$(\mathbf{I}, pnt, r)$
10: $\quad\quad\quad$ **foreach** $pnt' \in P'$ **do**
11: $\quad\quad\quad\quad \vec{f} =$`computeDescriptor`$(\mathbf{I}, pnt')$
12: $\quad\quad\quad\quad \mathbf{X}$.append$(\vec{f})$
13: $\quad\quad\quad$ **end**
14: $\quad\quad$ **end**
15: $\quad$ **end**
16: $\quad \Phi.fit(\mathbf{X})$
17: $\quad \beta_{p,l}.save(\Phi)$
18: $\quad B_p \leftarrow B_p \cup \{\beta_{p,l}\}$
19: **end**
20: **Return** B_p

(i.e. Gaussian distribution). Then, this distance is converted into a similarity value within $(0, 1]$, as defined in (6), which becomes the decision support of the Mahalanobis Similarity base learner.

$$\begin{aligned} p(\lambda|\mathbf{X}) &= \frac{p(\lambda)p(\mathbf{X}|\lambda)}{p(\mathbf{X})} \\ p(\lambda|\mathbf{X}) &\propto p(\lambda)p(\mathbf{X}|\lambda) \\ \Delta_{l,c} &= \vec{f}_l - \vec{\mu}_c \end{aligned} \tag{5}$$

$$\delta_{\text{MS}}(\vec{f}_l) = \left[1 + \ln\left(1 + \min_c\left(\Delta_{l,c}^{\intercal}\Sigma_c^{-1}\Delta_{l,c}\right)\right)\right]^{-1} \tag{6}$$

A face recognition ensemble system distributes the computed feature vectors $\vec{f}_l$ to their corresponding base learners (selected by the BLS block), and combines the outputs $d_{p,l}$ of the selected base learners $\beta_{p,l}$ by using the combination rule to compute the ensemble decision support value D_p. The value of D_p indicates the degree of support[6] that the input image matches the p^{th} subject (i.e. person) for which the p^{th} ensemble system was trained. The mean rule is employed to compute D_p, as defined in (7). It is worth mentioning that,

[6]In some cases $D_p \in [0, 1]$ represents the likelihood value. On the other hand for GMM $D_p \in (-\infty, 0]$, represents the log-likelihood. In any case, a greater value suggests that it is more likely that the input vector correspond to the subject p.

other combination rules (e.g. employing gating networks, an algorithm to assign base learner weights by estimating the diversity among the learners, assign weights based on Discriminative Power Analysis) could be employed to improve the recognition performance as detailed in [21], [22], and [25]. Nevertheless, these other combination approaches might require additional processing or training steps each time a new subject is added to the face database. For the sake of simplicity, we adopted the mean rule due to the satisfactory outcomes obtained in our previous work [14].

$$D_p = \frac{1}{|\mathscr{L}|} \sum_{l \in \mathscr{L}} d_{p,l} \tag{7}$$

$$P_E = \arg\max_{p \in \mathcal{P}} D_p \tag{8}$$

The landmark set $\mathscr{L}$, comprises all the landmarks selected for a given head pose class (a subset $\mathscr{L}$ of the total 24 landmarks is selected by the BLS block for each pose class to ensure that a given landmark is not occluded in the image). Finally, the predicted identity P_E, defined in (8), is obtained by choosing the ensemble system with the highest ensemble decision support value.

According to the definition of a base learner for this work, a model is trained for a specific landmark on a subject from the database. This process is called base learner training, depicted in Figure 5. The steps followed during base learner training depend on the working principle of the model type used as a base learner. In case of adopting GMM as the base learner, the steps detailed in Algorithm 3 are followed. For the case of using MS, Algorithm 4 is followed.

Algorithm 4 Base Learner Training Algorithm for Mahalanobis Similarity (MS)

Input: The training face image gallery $\mathcal{G}_p$ for the subject p, a predefined list of landmarks $\mathscr{L}$.
Result: The set of available trained base learners B_p for the subject p.
Data: Landmark neighborhood radius r

$B_p \leftarrow \emptyset$
for $(l = 0; l < 24; l++)$ **do**
  $\Phi \leftarrow$ `initBayesianGMMmodel()`
  $\beta_{p,l} \leftarrow$ `initBLModel`$(p, l,$ *type: MS*$)$
  $\mathbf{X} \leftarrow [\,]$ `// initialize an empty descriptor matrix`
  foreach $I \in \mathcal{G}_p$ **do**
    if *isVisible*$(\mathbf{I}, l)$ **then**
      $pnt =$ `getLandmarkLocations`$(\mathbf{I}, \{l\})$
      $P' =$ `surroundingPoints`$(\mathbf{I}, pnt, r)$
      foreach $pnt' \in P'$ **do**
        $\vec{f} =$`computeDescriptor`$(\mathbf{I}, pnt')$
        $\mathbf{X}$.append$(\vec{f})$
      end
    end
  end
  $\boldsymbol{\mu}, \Sigma \leftarrow \Phi.fit(\mathbf{X})$ `// Fit the BGMM model to` $\mathbf{X}$ `and return the means and covariance matrices`
  $\beta_{p,l}.save(\boldsymbol{\mu}, \Sigma)$
  $B_p \leftarrow B_p \cup \{\beta_{p,l}\}$
end
Return B_p

IV. EXPERIMENTAL RESULTS

The proposed method is implemented in Python 3 language on an Ubuntu 20.04 PC with a Core™ i7-8700H CPU, 16.00 GB of RAM, and a NVIDIA GeForce RTX 4070 Ti graphics card. For face detection, the Google MediaPipe model is employed. Whereas the pre-trained Face Alignment Network (FAN) [15] is used for facial landmark detection.

A. EMPLOYED FACE DATABASES

The CMU-PIE database [26] comprises over 40000 images of 68 subjects. This database has over 600 images from 13 poses (variation in the head yaw and pitch angles), with 43 different illuminations (the authors used a ''flash system''), and with 4 different expressions (neutral, talking, blinking, and smiling). In order to verify the effectiveness of the proposed PIFR method, only the images with ambient illumination, neutral expression and yaw angle variation are used. Thus, in this work 9 images per subject are employed with a total of 612 images.

The CMU Multi-PIE face database [27], developed between October 2004 and March 2005, was designed to facilitate the development of algorithms for face recognition across various challenging conditions, including pose, illumination, and expression variations. The database comprises a total of 337 subjects, with 129 subjects appearing in all four sessions, and containing over 750,000 images. For the purpose of this work, PIFR is performed on each session independently, by utilizing the images with neutral expression and ambient illumination (i.e. only pose variation). In total, 11973 face images (13 images per subject) from the CMU Multi-PIE are employed in this work.

TABLE 2. Description of the databases employed during the face recognition experiments (*only the images with neutral expression, ambient illumination, and pose variation are taken into account*).

Database		Number of subjects	Testing images per subject	Total number of images for testing
CMU-PIE		68	9	612
Multi-PIE	Session 01	249	13	3237
	Session 02	203	13	2639
	Session 03	230	13	2990
	Session 04	239	13	3107
FERET		200	9	1800

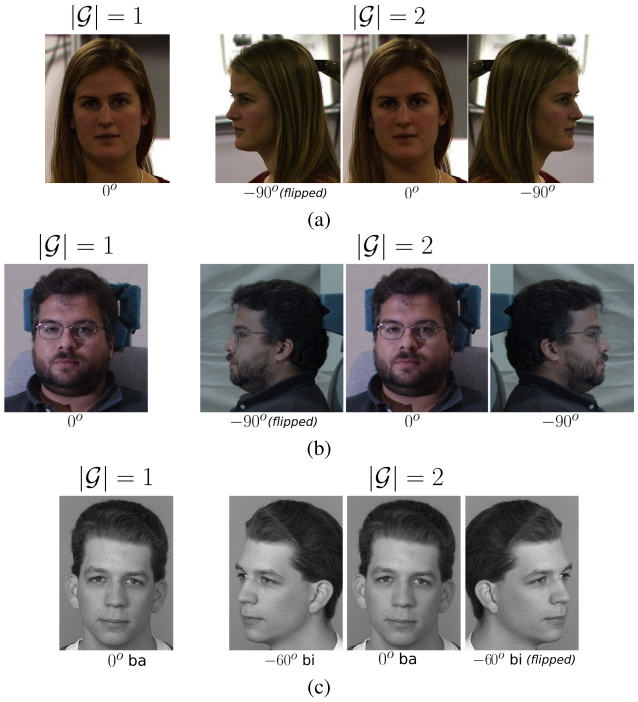

FIGURE 6. Gallery images employed during the training stage of the proposed ensemble systems: (a) CMU-PIE; (b) Multi-PIE; (c) FERET. *On the left: Images used for a gallery size of 1. On the right: Images used for a gallery size of 2.*

TABLE 3. Effect of utilizing different descriptors on the face recognition performance with different base learner models on the CMU-PIE Database ($|\mathcal{G}| = 1$).

Base learner	Feature descriptor	1:1 Verification TAR FAR 0.01	1:1 Verification TAR FAR 0.1	1:N Identification Rank-1	1:N Identification Rank-5
GMM	SIFT	0.576	0.715	0.777	0.877
	LS-SIFT	0.768	0.940	0.943	0.991
MS	SIFT	0.681	0.815	0.936	0.980
	LS-SIFT	**0.764**	0.906	**0.974**	**0.995**

The Facial Recognition Technology (FERET) Database [28] was developed from 1993 to 1997 with support from DARPA, USA. It aimed to create automatic face recognition capabilities for security, intelligence, and law enforcement applications. In this study, the Face Recognition Vendor Test 2000 (FRVT2000) protocol, detailed in the FERET database reports [28], is followed. This protocol aims to test the performance of a FR system to deal with different head orientations. It involves 200 subjects, with 9 testing images per subject. The gallery image consists of the frontal pose (yaw angle of 0^o), while the testing images include non-frontal poses with yaw variations of $\pm 15^o$, $\pm 25^o$, $\pm 40^o$, and $\pm 60^o$. A brief description (number of subjects, number of images) about the way these three databases are employed in this work is summarized in Table 2.

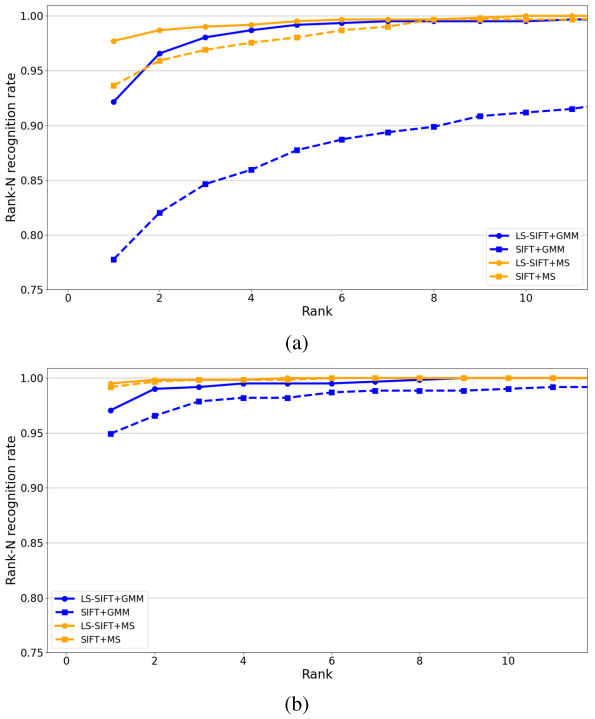

FIGURE 7. Cumulative Matching Characteristic (CMC) curve of the proposed methods on the CMU-PIE database: (a) CMC for CMU-PIE with $|\mathcal{G}| = 1$; (b) CMC for CMU-PIE with $|\mathcal{G}| = 2$.

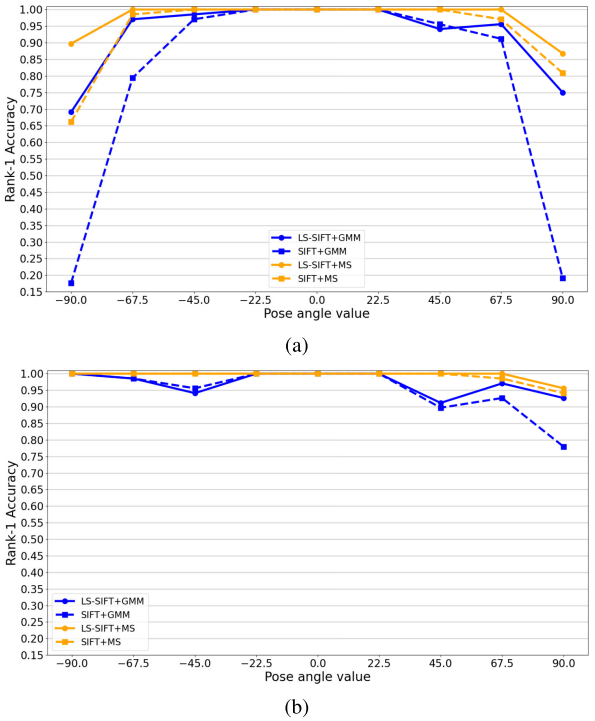

FIGURE 8. Rank-1 accuracy obtained with the proposed methods for different pose values on the CMU-PIE database: (a) Detailed Rank-1 accuracy with $|\mathcal{G}| = 1$; (b) Detailed Rank-1 accuracy with $|\mathcal{G}| = 2$.

B. EXPERIMENTAL RESULTS ON CMU-PIE, MULTI-PIE, AND FERET DATABASES

Most of results obtained in works utilizing CMU-PIE, Multi-PIE, and FERET databases are presented in terms

TABLE 4. Effect of using different descriptor types on the face recognition performance with images from the Multi-PIE database (*S stands for session number, MS is employed as the base learner model*).

S	Descriptor	Gallery size	1:1 Verification TAR	1:N Identification	
			FAR=0.01	Rank-1	Rank-5
01	SIFT	$\|\mathcal{G}\| = 1$	0.454	0.773	0.868
		$\|\mathcal{G}\| = 2$	0.732	0.912	0.969
	LS-SIFT	$\|\mathcal{G}\| = 1$	0.564	0.848	0.935
		$\|\mathcal{G}\| = 2$	**0.823**	**0.948**	**0.984**
02	SIFT	$\|\mathcal{G}\| = 2$	0.725	0.911	0.970
	LS-SIFT	$\|\mathcal{G}\| = 2$	**0.810**	**0.945**	**0.984**
03	SIFT	$\|\mathcal{G}\| = 2$	0.714	0.914	0.969
	LS-SIFT	$\|\mathcal{G}\| = 2$	**0.801**	**0.943**	**0.980**
04	SIFT	$\|\mathcal{G}\| = 2$	0.730	0.917	0.972
	LS-SIFT	$\|\mathcal{G}\| = 2$	**0.819**	**0.949**	**0.985**

TABLE 5. Effect of using different descriptor types on the face recognition performance with images (pose variation only) from the FERET database (*MS is employed as the base learner model*).

Feature descriptor	Gallery Size	1:1 Verification TAR	1:N Identification	
		FAR = 0.01	Rank-1	Rank-5
SIFT	$\|\mathcal{G}\| = 1$	0.560	0.876	0.942
	$\|\mathcal{G}\| = 2$	0.912	0.976	0.995
LS-SIFT	$\|\mathcal{G}\| = 1$	0.663	0.902	0.953
	$\|\mathcal{G}\| = 2$	**0.936**	**0.980**	**0.995**

of the face recognition rate (also called Rank-1 accuracy), which is an indicator of the face identification performance. Additional experiments on face verification and identification are conducted in this study, as it was done in our previous work [14]. The TAR@FAR metric is used to measure the performance on face verification, while the Rank-N accuracy is employed for face identification [29]. All the experimental trials are conducted for two different gallery image sizes. For the case of CMU-PIE and Multi-PIE, the frontal image (0^o pose) is employed for training when utilizing a gallery image size of $|\mathcal{G}| = 1$. Conversely, when the gallery image size is set to $|\mathcal{G}| = 2$, the images with 0^o, and -90^o are employed for training the FR models (see Fig. 6a, Fig. 6b). In order to improve the robustness of the method under $|\mathcal{G}| = 2$, the flipped version of the -90^o image is also employed for training, as can be appreciated in Fig. 6.

The recognition performance achieved on CMU-PIE, for a gallery size of 1, can be better visualized with their Cumulative Matching Characteristic (CMC) curves, shown in Fig. 7a. From Fig. 7a it is evidenced that in general, utilizing LS-SIFT with any base learner yields better results than SIFT. Besides, the recognition rate is close to 100% at Rank-5 when utilizing LS-SIFT+MS. Additional experiments are conducted with a gallery size $|\mathcal{G}| = 2$, and their results are depicted in Fig. 7b. There is a clear improvement compared to using $|\mathcal{G}| = 1$. A detailed (per pose) Rank-1 accuracy for

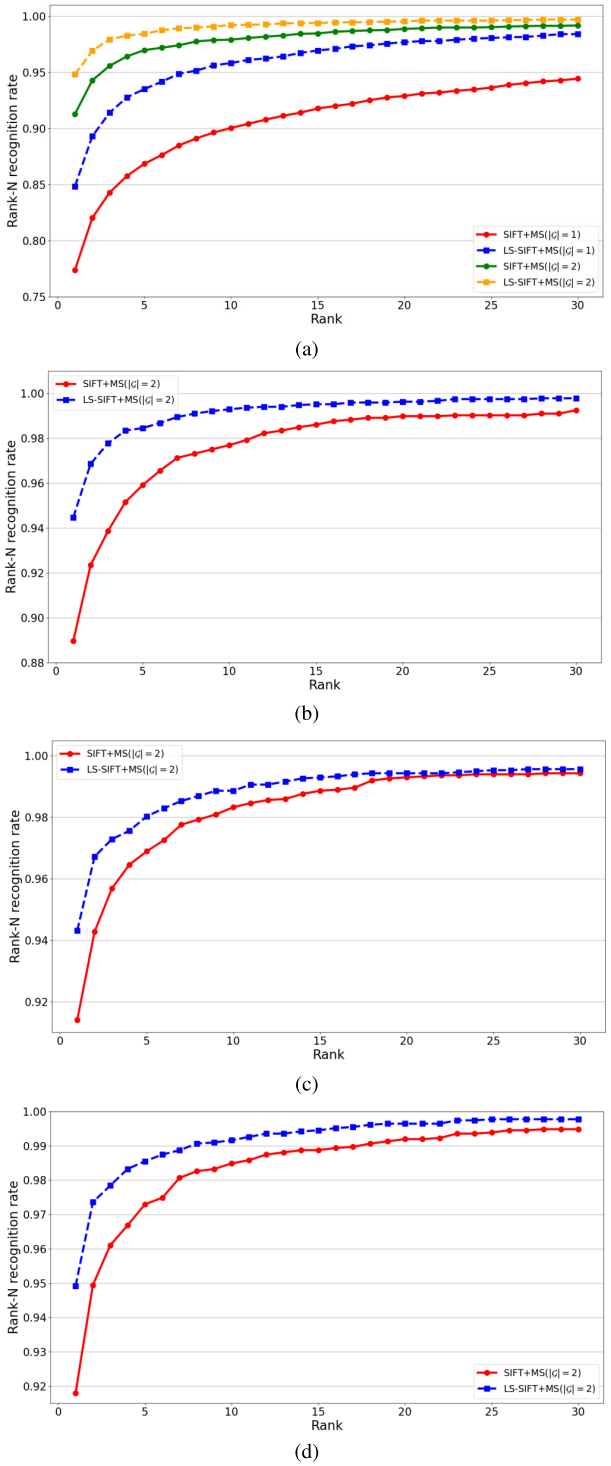

FIGURE 9. Cumulative Matching Curve for different methods tested on the Multi-PIE database: (a) CMC for Session 1; (b) CMC for Session 2; (c) CMC for Session 3; (d) CMC for Session 4.

the different combinations of descriptors and base learners is shown in Fig. 8a. Again, the use of $|\mathcal{G}| = 2$ improves the overall PIFR performance on images with large poses, as shown in Fig. 8b. The results of both face verification and identification on CMU-PIE are summarized in Table 3.

TABLE 6. Detailed performance comparison of the proposed method with state-of-the-art methods for PIFR on the CMU-PIE database.

Method	Rank-1 Accuracy *a.k.a.* Face recognition rate (%)									
	-90^o	-67.5^o	-45^o	-22.5^o	0^o	22.5^o	45^o	67.5^o	90^o	Overall
FLM + PFER-GEM [30]	88.70	100.00	100.00	100.00	100.00	100.00	100.00	98.38	91.90	98.24
Statistical classification + Local Gabor Features [31]	72.40	83.10	100.00	100.00	100.00	100.00	100.00	99.70	97.50	94.10
PBPR-MtFL [4]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	98.53	99.84
Divide-and-rule + LBP-Huffman [32]†	-	-	100.00	100.00	100.00	100.00	100.00	-	-	-
Face frontalization + LGBP [2]†	-	-	91.2	98.5	100.00	100.00	98.5	-	-	-
Ensemble SVM + SIFT [14]	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Ensemble MS + LS-SIFT ($\|\mathcal{G}\| = 1$)	89.71	100.00	100.00	100.00	100.00	100.00	100.00	100	86.76	97.38
Ensemble MS + SIFT ($\|\mathcal{G}\| = 2$)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	98.53	94.12	99.18
Ensemble MS + LS-SIFT ($\|\mathcal{G}\| = 2$)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	95.59	99.51

† Only face images with pose angles between $\pm 45^o$ were considered for testing.

TABLE 7. Detailed Rank-1 accuracy comparison of the proposed method with state-of-the-art methods for PIFR on the Multi-PIE database (***Session 1, pose variation only***).

Method	Rank-1 Accuracy *a.k.a.* Face recognition rate (%)													
	-90^o	-75^o	-60^o	-45^o	-30^o	-15^o	0^o	15^o	30^o	45^o	60^o	75^o	90^o	Overall
SFC-GAN [33]	-	-	-	-	-	-	98.7	97.7	96.3	91.0	**85.8**	78.4	62.4	-
Stacked OPR [34] †	-	-	-	86.00	95.00	96.70	98.20	97.30	96.70	90.70	-	-	-	-
Ensemble MS + LS-SIFT ($\|\mathcal{G}\| = 1$)	48.58	70.68	87.95	98.80	100.0	100.0	100.0	100.0	99.60	96.39	82.73	69.88	47.56	84.84
Ensemble MS + SIFT ($\|\mathcal{G}\| = 2$)	100.0	88.35	87.55	99.20	100.0	100.0	100.0	100.0	99.60	97.19	71.49	66.67	76.42	91.27
Ensemble MS + LS-SIFT ($\|\mathcal{G}\| = 2$)	**100.0**	**95.98**	**95.58**	**99.20**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**97.19**	84.34	**81.53**	**78.46**	**94.80**

† Only face images with pose angles between $\pm 45^o$ were considered for testing.

TABLE 8. Detailed performance comparison of the proposed method with state-of-the-art methods for PIFR on the FERET database.

Method	Rank-1 Accuracy *a.k.a.* Face recognition rate (%)									
	-60^o	-40^o	-25^o	-15^o	0^o	15^o	25^o	40^o	60^o	Overall
RR+Gabor [35]†	78.00	91.00	96.00	96.00	-	98.00	99.00	96.00	87.00	92.63
PBPR-MtFL [4]†	**100.0**	**99.00**	**100.0**	**100.0**	-	**100.0**	**100.0**	**100.0**	**98.00**	**99.63**
Pose normalization + PCA Gabor Features [36]	83.5	94.5	98.5	100.0	-	100.0	99.00	94.50	80.50	93.81
Ensemble MS + SIFT ($\|\mathcal{G}\| = 1$)	64.50	88.00	94.00	97.50	100.00	98.00	96.50	91.50	59.00	87.67
Ensemble MS + SIFT ($\|\mathcal{G}\| = 2$)	100.00	98.50	99.00	98.00	100.00	99.00	99.50	96.50	88.50	97.67
Ensemble MS + LS-SIFT ($\|\mathcal{G}\| = 1$)	76.00	91.00	95.50	96.00	100.00	97.50	96.50	91.50	68.50	90.28
Ensemble MS + LS-SIFT ($\|\mathcal{G}\| = 2$)	100.00	99.00	99.00	98.00	100.00	99.50	99.50	98.00	89.50	98.05

† Only 100 out of 200 subjects were considered for testing.

The experimental results on Multi-PIE are summarized in Table 4. For Session 01, experiments with $|\mathcal{G}| = 1$ and $|\mathcal{G}| = 2$ are conducted. The performance superiority, in terms of Rank-1 accuracy, of using LS-SIFT over SIFT is very noticeable when using $|\mathcal{G}| = 1$. Indeed, there is an accuracy difference of over 7.0% (around 227 faces images). This performance gap is reduced when using $|\mathcal{G}| = 2$ images. However, the superiority of using LS-SIFT remains, with a difference of 3.6% in terms of the Rank-1 accuracy. The results on face verification, confirms the higher level of robustness of LS-SIFT compared to SIFT. The results on the remaining sessions of Multi-PIE evidence again that the best results are obtained with LS-SIFT. In order to visualize the degree of enhancement offered by using LS-SIFT instead of SIFT, the CMC curves of each session from the Multi-PIE database are presented in Fig. 9.

The third dataset employed for testing the proposed PIFR framework is FERET. As was mentioned above, the FRVT2000 test is employed. This test involves performing PIFR on 200 subjects with 9 images per subject. In this work experimental results are obtained with $|\mathcal{G}| = 1$, and $|\mathcal{G}| = 2$. The frontal image is employed for the trials with

$|\mathcal{G}| = 1$, while the -60^o image and its flipped counterpart are included when $|\mathcal{G}| = 2$. After performing PIFR with the proposed methods, the obtained experimental results are summarized in Table 5. It can be seen that the results LS-SIFT are slightly better than the ones with SIFT on the FERET database. Indeed, when $|\mathcal{G}| = 1$ is employed, LS-SIFT shows an improvement of 2.6% (equivalent to 47 images) over SIFT on the Rank-1 accuracy. The same situation occurs for face verification.

C. PERFORMANCE COMPARISON ON FACE RECOGNITION WITH STATE-OF-THE-ART WORKS

The CMU-PIE database has served as a benchmark for evaluating various state-of-the-art methods in PIFR. The obtained experimental results are compared with those prior works that have utilized this database [2], [32], [37] in Table 6. Notably, some of these works only considered a pose angle range of $\pm 45^o$ in their experiments. In contrast, the proposed approach incorporates pose angles ranging from $\pm 90^o$, achieving a remarkable recognition rate of 100% on almost all the pose angles. Moreover, it is evident that some works that also considered $\pm 90^o$ pose-view images experienced a significant decline in performance beyond the $\pm 45^o$ range. It should be pointed out that, in our previously published work [14], an ensemble learning approach was used with SVM as the base learner model, SIFT descriptors, and a gallery size of $|\mathcal{G}| = 3$. However, experimental trials on Multi-PIE showcased that the method in [14] is suitable to be used in low-scale face databases (around 50 subjects). In large-scale databases (more than 200 subjects), there is a significant recognition rate drop due to the way the base learners are trained.

The results obtained on session 1 of Multi-PIE are compared with the results of other state-of-the-art methods assessing the FR performance on images with pose variation exclusively. This comparison is presented in Table 7. Most of the results achieved in this work outperform the ones obtained in the works selected for comparison. It is noticeable that, the best results are obtained by combining MS with LS-SIFT, and a gallery size of $|\mathcal{G}| = 2$. Moreover, the results obtained with MS + LS-SIFT and $|\mathcal{G}| = 1$ outperforms [34] for any pose value.

In Table 8, the results obtained on FERET are compared with state-of-the-art works. The comparison shows that the performance of the proposed method with $|\mathcal{G}| = 1$ is inferior than other works utilizing the FERET database. If $|\mathcal{G}| = 2$ is employed, the obtained results are close to the best result on FERET. However, the testing protocol on FERET establishes that only the frontal face image should be used as the gallery image.

V. CONCLUSION

The presented work addressed the problem of face recognition under pose variations (pose-invariant face recognition) by combining the ability of local feature descriptors in representing facial information at specific face regions (facial landmarks), and the power of ensemble systems in combining several weak classifiers (base learners) to achieve a high recognition accuracy. In order to perform face recognition, an input face image is processed to detect its facial landmark locations. Then, these landmark locations are processed to classify the input image according to its head pose class. According to its pose class, facial landarks are selected and feature extraction is performed over them. The feature vectors are input to their corresponding base learners, and the ensemble decision is computed from the base learners' outputs. Finally, the face identity is computed by choosing the ensemble system whose ensemble decision support is the highest.

Most of the works on PIFR focus merely on the face identification task with the Rank-1 accuracy (recognition rate) as the main indicator. In this work, the results on both face verification, and identification were included. As above mentioned, the proposed PIFR framework includes a local feature extraction stage. A novel local feature descriptor, called Landmark-specific SIFT (LS-SIFT) was proposed in this work. LS-SIFT is obtained by applying a learned non-linear vector transformation (mapping) to a SIFT feature obtained from a specific facial landmark, improving its robustness against pose variations. On the other hand, two novel base learner models were proposed (GMM, Mahalanobis similarity). In case of GMM, the training methodology is the novelty. The PIFR performance obtained by using these models were compared during the experimental trials on CMU-PIE, showing that the Mahalanobis similarity (MS) model performs the best (close to 100% with a gallery size of 2). Furthermore, experiments on CMU-PIE showed a significant performance improvement when LS-SIFT was employed as the feature descriptor, compared to SIFT.

The FR experiments on Multi-PIE were conducted on each of its 4 sessions. MS was selected as the base learner model, and the experiments aimed to show the superiority of LS-SIFT over SIFT. This superiority for FR were evidenced on every session. Indeed, the use of LS-SIFT yielded a Rank-1 accuracy over 94% on any session. On the other hand, an accuracy around 90% is obtained when using SIFT. This superiority is also appreciable on the face verification results, where the TAR@FAR metric was employed, especially for low FAR values (say 0.001). For the case of FERET, the obtained results were close to the best results on this database under the FRVT2000 protocol. In summary, the current work has evidenced the performance improvement during FR when LS-SIFT is employed instead of the conventional SIFT. Furthermore, the Mahalanobis Similarity (MS) base learner model, introduced in this work, showed to perform remarkably better than the conventional GMM during PIFR.

REFERENCES

[1] M. Taskiran, N. Kahraman, and C. E. Erdem, "Face recognition: Past, present and future (a review)," *Digital Signal Process.*, vol. 106, 2020.

[2] C. Petpairote, S. Madarasmi, and K. Chamnongthai, "2D pose-invariant face recognition using single frontal-view face database," *Wireless Pers. Commun.*, vol. 118, pp. 2015–2031, 2021.
[3] C. Ding, J. Choi, D. Tao, and L. S. Davis, "Multi-directional multi-level dual-cross patterns for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 518–531, Mar. 2016.
[4] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 980–993, Mar. 2015.
[5] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[6] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.
[7] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022.
[8] A. F. Abate, P. Barra, C. Bisogni, M. Nappi, and S. Ricciardi, "Near real-time three axis head pose estimation without training," *IEEE Access*, vol. 7, pp. 64256–64265, 2019.
[9] P. Werner, F. Saxen, and A. Al-Hamadi, "Landmark based head pose estimation benchmark and method," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3909–3913.
[10] Z. An, W. Deng, J. Hu, Y. Zhong, and Y. Zhao, "APA: Adaptive pose alignment for pose-invariant face recognition," *IEEE Access*, vol. 7, pp. 14653–14670, 2019.
[11] D. Lin, Y. Li, Y. Cheng, S. Prasad, and A. Guo, "Masked face recognition via self-attention based local consistency regularization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 436–440.
[12] S.-C. Lai, M. Kong, K.-M. Lam, and D. Li, "High-resolution face recognition via deep pore-feature matching," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3477–3481.
[13] S. D. Lin and P. L. Otoya, "Large pose detection and facial landmark description for pose-invariant face recognition," in *Proc. IEEE 5th Int. Conf. Knowl. Innov. Invention. (ICKII)*, Jul. 2022, pp. 143–148.
[14] S. D. Lin and P. E. L. Otoya, "Pose-invariant face recognition via facial landmark based ensemble learning," *IEEE Access*, vol. 11, pp. 44221–44233, 2023.
[15] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.
[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, Nov. 2004.
[17] Google. (Jul. 2023). *TensorFlow API Documentation*. Accessed: Jun. 13, 2023. [Online]. Available: https://www.tensorflow.org/api_docs
[18] *Keras API Documentation*, Keras, Jun. 2023. Accessed: Jun. 13, 2023. [Online]. Available: https://keras.io/api/
[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.
[20] *Open Source Computer Vision (OpenCV)*, document 4.5.5, Intel, Santa Clara, CA, USA, Dec. 2023. [Online]. Available: https://docs.opencv.org/4.5.5/
[21] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, 3rd Quart., 2006.
[22] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, 2nd ed. Hoboken, NJ, USA: Wiley, 2014, ch. 3.
[23] N. C. Oza and K. Tumer, "Input decimation ensembles: Decorrelation through dimensionality reduction," in *Multiple Classifier Systems*. Berlin, Germany: Springer, 2001, pp. 238–247.
[24] J. Lu, "A survey on Bayesian inference for Gaussian mixture model," 2021, *arXiv:2108.11753*.
[25] L. Leng, M. Li, C. Kim, and X. Bi, "Dual-source discrimination power analysis for multi-instance contactless palmprint recognition," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 333–354, 2017.
[26] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.
[27] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–8.
[28] P. Phillips, H. Moon, P. Rauss, and S. Rizvi, "The FERET evaluation methodology for face-recognition algorithms," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 1997, pp. 137–143.
[29] Z. Cheng, X. Zhu, and S. Gong, "Surveillance face recognition challenge," 2018, *arXiv:1804.09691*.
[30] A. Moeini and H. Moeini, "Real-world and rapid face recognition toward pose and expression variations via feature library matrix," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 969–984, May 2015.
[31] L. A. Cament, F. J. Galdames, K. W. Bowyer, and C. A. Perez, "Face recognition under pose variation with local Gabor features enhanced by active shape and statistical models," *Pattern Recognit.*, vol. 48, no. 11, pp. 3371–3384, 2015.
[32] L.-F. Zhou, Y.-W. Du, W.-S. Li, J.-X. Mi, and X. Luan, "Pose-robust face recognition with Huffman-LBP enhanced by divide-and-rule strategy," *Pattern Recognit.*, vol. 78, pp. 43–55, Jun. 2018.
[33] H. Lin, H. Ma, W. Gong, and C. Wang, "Non-frontal face recognition method with a side-face-correction generative adversarial networks," in *Proc. 3rd Int. Conf. Comput. Vis., Image Deep Learn. Int. Conf. Comput. Eng. Appl. (CVIDL ICCEA)*, May 2022, pp. 563–567.
[34] Y. Tai, J. Yang, Y. Zhang, L. Luo, J. Qian, and Y. Chen, "Face recognition with pose variations and misalignment via orthogonal Procrustes regression," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2673–2683, Jun. 2016.
[35] A. Li, S. Shan, and W. Gao, "Coupled bias–variance tradeoff for cross-pose face recognition," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 305–315, Jan. 2012.
[36] J. Yan, Y. Mei, X. Liu, C. Dai, and T. Yu, "Patch-wise normalization for pose-invariant face recognition from single sample," in *Proc. IEEE Int. Conf. Internet Things (iThings), IEEE Green Comput. Commun. (GreenCom), IEEE Cyber, Phys. Social Comput. (CPSCom), IEEE Smart Data (SmartData)*, Jul. 2018, pp. 712–715.
[37] E. A. Mostafa and A. A. Farag, "Dynamic weighting of facial features for automatic pose-invariant face recognition," in *Proc. 9th Conf. Comput. Robot Vis.*, May 2012, pp. 411–416.

SHINFENG D. LIN (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Mississippi State University, in 1991. He was the Director of the Bureau of Education, Hualien, Taiwan, from January 2002 to September 2003. He is currently a Professor with the Department of Computer Science and Information Engineering, National Dong Hwa University (NDHU), Taiwan. He is also the Director of the Artificial Intelligence Office, NDHU. He has published over 150 journals and conference papers. His research interests include signal/image processing, machine learning, pattern recognition, and information security. He is a fellow of IET. He won the Gold Medal Award from the 2005 International Trade Fair "Ideas-Inventions-New Products" (IENA), Nuremberg, Germany.

PAULO E. LINARES OTOYA (Member, IEEE) was born in Trujillo, La Libertad, Peru, in 1993. He received the B.S. degree in electronic engineering from Universidad Privada Antenor Orrego (UPAO), Trujillo, in 2019, and the M.Sc. degree in computer science and information engineering from National Dong Hwa University (NDHU), Hualien, Taiwan. He is currently pursuing the Ph.D. degree with National Taiwan University. From 2019 to 2021, he was a Research Assistant with the UPAO Multidisciplinary Research Laboratory (LABINM-UPAO), conducting research on computer vision applied to precision agriculture. His research interests include computer vision, artificial intelligence, face recognition, head pose estimation, and human action recognition.