Hindawi
Security and Communication Networks
Volume 2021, Article ID 6673992, 12 pages
https://doi.org/10.1155/2021/6673992

Research Article

Fusion of Machine Learning and Privacy Preserving for Secure Facial Expression Recognition

Asad Ullah,[1] Jing Wang,[2] M. Shahid Anwar,[2] Arshad Ahmad,[3] Shah Nazir,[4] Habib Ullah Khan,[5] and Zesong Fei[2]

[1]*Department of Computer Science & IT, Sarhad University of Science and Information Technology, Peshawar, Pakistan*
[2]*School of Information and Electronics, Beijing Institute of Technology, Beijing, China*
[3]*Department of IT & Computer Science, Pak-Austria Fachhoschule: Institute of Applied Sciences and Technology, Haripur, Pakistan*
[4]*Department of Computer Science, University of Swabi, Swabi, Pakistan*
[5]*Department of Accounting & Information Systems, College of Business & Economics, Qatar University, Doha, Qatar*

Correspondence should be addressed to Jing Wang; wangjing@bit.edu.cn

Received 9 October 2020; Revised 8 December 2020; Accepted 18 January 2021; Published 30 January 2021

Academic Editor: Amir Anees

The interest in Facial Expression Recognition (FER) is increasing day by day due to its practical and potential applications, such as human physiological interaction diagnosis and mental disease detection. This area has received much attention from the research community in recent years and achieved remarkable results; however, a significant improvement is required in spatial problems. This research work presents a novel framework and proposes an effective and robust solution for FER under an unconstrained environment; it also helps us to classify facial images in the client/server model along with preserving privacy. There are a lot of cryptography techniques available but they are computationally expensive; on the other side, we have implemented a lightweight method capable of ensuring secure communication with the help of randomization. Initially, we perform preprocessing techniques to encounter the unconstrained environment. Face detection is performed for the removal of excessive background and it detects the face in the real-world environment. Data augmentation is for the insufficient data regime. A dual-enhanced capsule network is used to handle the spatial problem. The traditional capsule networks are unable to sufficiently extract the features, as the distance varies greatly between facial features. Therefore, the proposed network is capable of spatial transformation due to the action unit aware mechanism and thus forwards the most desiring features for dynamic routing between capsules. The squashing function is used for classification purposes. Simple classification is performed through a single party, whereas we also implemented the client/server model with privacy measurements. Both parties do not trust each other, as they do not know the input of each other. We have elaborated that the effectiveness of our method remains unchanged by preserving privacy by validating the results on four popular and versatile databases that outperform all the homomorphic cryptographic techniques.

1. Introduction

Facial expressions contain the most important nonverbal and rich emotional information in social communication [1]. People communicate with each other through verbal and nonverbal communications [2]. Nonverbal communication involves facial gestures, eye to eye contact, facial expressions, and paralanguage [3]. According to an earlier research, while communicating, 50 percent of the information is conveyed through facial expression, 40 percent through voice, and 8 percent through language. Apart from that, due to the rapid progression in technology, we spend most of the time on electronic devices that carry a variety of software interfaces that are tense, primitive, and nonverbal. Therefore, facial expression recognition should further improve to have a more natural and intelligent human–machine interaction.

Facial expression recognition is used in various domains like Intelligent Tutoring System (ITS), psychology, human-machine interaction, behavioral science, intelligent transportation, and interactive games [4]. It can help monitor the

abnormal expressions in the crowd at public places to avoid any crime. It can also be helpful in the service industry to timely capture the feedback of customers and it can provide timely treatment of patients by looking at the real-time expressions of the patient at the hospital. According to Ekman and Friesman [5], there are six basic expressions: happiness, surprise, disgust, fear, sadness, and anger (some researchers have termed neutral expression as the seventh expression). These expressions are conveyed almost among all species.

Facial expression recognition is widely studied by various researchers. Despite the available research, robust FER is yet an open and challenging task [6, 7] However, most of the recognition algorithms do not consider inter-class variations caused by the differences in facial attributes of the same individual. Hence, mostly, expression classification is done through facial expression information along with identity-related information [8, 9]. The main drawback it carries is that it affects the overall generalization capability of FER systems, thus resulting in degradation of performance on unseen identities [10]. An efficient FER system plays a vital role in the treatment of patients by observing their variable behavior patterns. Happiness expression depicts a healthy and positive mental state while sad and angry demonstrate an unhealthy mental state. Different mental diseases like autism or anxiety are detected due to the emotional conflicts of a particular patient. An important application of FER is E-health care; nowadays, almost 0.3 billion people are suffering from depression, which can also lead to suicidal tendencies if they are not treated timely and effectively [11]. In general, mental health treatment faces a lot of barriers like financial cost, social stigma, and shortage of accessible options. Normally, the clinical staff interviews a patient for identifying symptoms of depression via verbal and nonverbal indicators. Patients are asked to fill a questionnaire for the measurement of depression severity [12]. For timely detection of depression symptoms, an AI-based system will help in entrenching barriers for timely and effective treatment.

In this paper, we use a combination of different techniques to develop a robust model. Initially, we implement different preprocessing techniques to fine-tune and remove highly uncorrelated information in the images. Face detection is performed using facial attributes due to the following reasons: (1) The human face has a unique structure, with the most important local facial parts, such as eyes, mouth, nose, helping us to detect the face in an unconstrained environment. So, the partness map or the response map of five different parts is used in the method. (2) The face adheres to spatial arrangements like the hair being above the eyes and lips below the nose. Hence, the faceness score has been derived from the response configuration. (3) The face hypothesis is performed for the estimation of more accurate face locations. Our contribution is to introduce special attributes supervision to discover facial part responses. We adapt Deep Convolutional Generative Adversarial Network (DCGAN) for data augmentation. It helps us in the demonstration of realistic data augmentation and improvement in the generalization performance in the low-data regime.

For an accurate and robust FER, feature representation of the facial images is the most important step. A considerable amount of research has been done over local and global feature extraction [13]. Fan et. al [14] suggested a model, i.e., MRE-CNN, which aimed to enhance the learning power of the convolutional neural network by considering both the local and global features. Li and Deng [15] introduced the DLP-CNN framework in which the discrimination power of deep features is enhanced while maximizing the interclass scatter and by preserving the locality closeness. Still, they are unable to find the relative relationship between the local features. A face is composed of a certain structure where every part has a relative relationship with the other parts. To address this issue, we propose a method that is capable of spatial transformation due to the action unit aware mechanism and thus forwards the most desiring features for dynamic routing between capsules. Finally, the squashing function is used for classification purposes. We also faced the challenge of achieving classification while having client/server as mutually distrustful of disclosure of the private contents of the facial images and without presenting the result to the server. There are many practical and potential applications, but the main focus is to capture useful and discriminative features. The better feature representations can help to improve the overall efficiency of the system. An appropriate, flexible, and effective facial expression recognition system will add benefit to the industry. There are a lot of standard cryptophic techniques just like secure multiple-parties communication and homomorphic encryption, but they were computationally way too expensive. Thus, we have provided a practical solution to the aforementioned problems. We assess the effectiveness and performance of the introduced model on the Extended Cohn-Kanade, MMI, Oulu-CASIA, and Real-world Affective Faces (RAF) databases. Figure 1 shows some sample images from the CK + database.

The main contributions of this paper are as follows:

(1) We propose a network, which is capable of finding the active relationship between the features from different local regions. Spatial information is also introduced by having prior knowledge of the probability of an object's existence.

(2) Implementation of a simple FER without using the cryptographic techniques having high computational complexity.

(3) Simultaneously achieving the same classification accuracy as that of a conventional algorithm (non-privacy-preserving).

The organization of the next sections is as follows. In Section 2, we provide the problems with the existing methods. In Section 3, we elaborate on our novel architecture with the underlying information. Section 4 comprises the results and analysis. Finally, we provide the conclusion of our research and explain the direction for future work in the last section.

Figure 1: Different facial expressions' samples taken from CK + database.

2. Related Work

The main goal of FER is to capture the meaningful features that are discriminative and descriptive, and invariant to facial variations such as occlusion, illumination, pose, and other identity-related details. There are two main methods available for feature extraction: (1) handcrafted method and (2)deep-learning-based method. Nowadays, deep learning methods are gaining remarkable results. However, earlier, mostly facial expression recognitions were based on handcrafted/human-engineered features such as Histograms of Oriented Gradients (HOG) [16], *n*-dimensional scale-invariant feature transform (n: sift) [17], and Local Phase Quantization (LPQ) [18]. These methods are used for the extraction of global as well as detailed information of an individual face. However, the information obtained is from the overall facial region, and it ignores the expression changes in the local regions, which contain the eyes, nose, and mouth. These methods perform pretty well in a lab-controlled environment where subjects pose expressions under constant illumination, stable eye gaze, and head pose movement. Existing handcrafted approaches demonstrate comparatively less recognition accuracy. Efforts are exerted for manually extracting the desired discriminating features that are linked to expression changes. Considering in-the-wild scenarios deep learning methods for the robustness of facial expression recognition have been implemented [19–22]. However, deep representation is affected just because every facial attribute of a particular subject carries a hefty number of variations such as gender, ethnicity, and age of the particular posing expressions. It holds a very big disadvantage, i.e., the generalization capability for any model is highly and negatively affected; as a result of unseen objects, the performance of facial expression recognition is degraded. Although quite a lot of work has been done toward improving the performance of FER, alleviating the influence of inter-subject variations is still a challenge and an open area of research.

Several techniques have been implemented by reducing intra-class variations and by increasing the interclass differences, which further increases the discriminating property of the features extracted for FER in the real-time scenario [23]. Identity-Aware CNN (IACNN) proposed that by reducing the influence of identity-related information with the use of expression and identity-sensitive contrastive losses, the facial expression recognition performance can be enhanced [24]. The island loss has been proposed for extracting the effective discriminative features for FER [25]. Moreover, in [26], with the use of residue learning the person-independent expression representation has been learned. However, this technique was computationally costly, and due to the same intermediate representation used for the generation of neutral images for the same identities, it also was unable to disentangle the expression information from identity information. However, in [24], due to large data expansion caused by the compilation of training data in image pair forms, the effectiveness of contrastive loss is heavily affected [25]. Similarly, in [27], a fixed identity has been proposed for the transfer of facial expressions to fix the influence of identity relative information. The problem persists with the methods as the efficiency of FER depends on the expression transfer procedure. In short, it has been noticed that FER based on the deep learning methods has outperformed the traditional handcrafted methods. However, there is still a gap in deep learning because very few studies have employed facial depth images in the deep networks as an input. Compared with the existing models, the main goal is to design a network that can be fully adopted for the decomposition of the facial region, easy to implement, and is robust.

Different researchers have implemented different methods to ensure privacy. In [28], the privacy-preserving data classification was done with the use of Principal Component Analysis for feature extraction, and for classification, the nearest neighbor was used. However, it failed to perform in the presence of nonlinear facial variations. Fisher Linear Discriminant Analysis has been proposed in [29] and it had less error rate compared to PCA. However, it did not work well for maintaining the privacy of the discriminative features of a specific class in the multimodal class. Hence, LFDA was proposed to overcome the deficiencies in the FLDA. The work in [30] meets the privacy requirements by hiding the test image and achieving results using the Paillier Homo-morphic encryption [31]. In the research work in

[32], the author proposed EPOM that achieves the goal of secure integer number processing without resulting in privacy leakage of data to unauthorized parties. In [33], it has been proposed that subprotocols can dramatically reduce the number of messages exchanged during the iterative approximation process based on the coordinate rotation digital computer algorithm. Due to the large keys for the encryption as well as decryption, it involves computationally intensive operations such as a large number of exponents. Meanwhile, it also has a limited number of operations during the classification of data, which makes the client/server communicate even a lot more with each other. Hence, our proposed method can achieve true recognition rate even in the presence of the privacy protocol, which uses randomization and is capable of intense multiplication and addition.

3. Proposed Method

3.1. Preprocessing. Preprocessing is very important as it aims to capture the meaningful features, align, and normalize the most needed visual information conveyed by the facial image. Every real-time image is affected by nonlinear facial variations, i.e., varying illumination, the difference in the contrast between the foreground and background, and irrelevant head poses. Therefore, to get the maximum possible semantic meanings of the features for further training the deep neural network, we need to perform some preprocessing techniques. This step is used for the elimination of highly uncorrelated data in the image.

3.1.1. Face Detection. Face detection is one of the vital steps in the FER because of the excessive background, and there is still highly uncorrelated information in the image even taken from a few benchmark datasets. Most of the datasets have an almost frontal view and high-resolution images. So Viola and Jones algorithm [34] is used in most scenarios.

Faceness-net has been used in this paper. A full image is provided as an input image to the convolutional neural network for generation of partness map. The partness map is generated for different facial parts like eyes, nose, mouth, etc. Facial attributes are further categorized to distinguish it from other parts, just like how hair can be blond, black, wavy, straight, etc. Therefore, in the next stage, face proposals are much more refined, so that the usefulness of facial attributes are explored for learning an optimized and robust face detection. A CNN is trained over uncropped images and is used for obtaining face part detectors without any explicit part supervision. The faceness score is evaluated based on the face part responses and considering the spatial arrangements associated with them. After the generation of face proposals, a strong face detector is trained and it outperforms all other methods.

In Figure 2, the face is divided into five important parts, where eyes, nose, and hair are much more effective as compared to mouth and beard, which can be partially occluded. Therefore, the combination of facial parts gives much better results compared to individual facial parts.

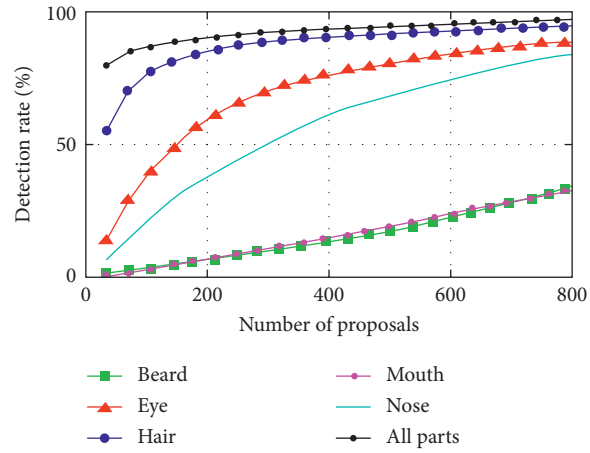


Figure 2: Impact of different facial parts to face proposal (individual or as a combination).

3.1.2. Data Augmentation. As far as the deep neural network is chosen for FER, data augmentation is used to produce much better results by providing a large amount of data. It is effective in the generalization capability of the model as many of the publicly available datasets are not large enough to validate the results more efficiently. Large training data yield to a well-trained model.

There are some standard methods of data augmentation like skewing, rotating, shifting, changing the color scheme, resizing the image, and enhancement of image noise [22]. To automatically learn the augmented data in the low-data setting, we have used Deep Convolutional Generative Adversarial Network (DCGAN). It is used for the alleviation of the overfitting problem over the on-the-fly data. The samples provided as input are randomly cropped from all the four sides and then a horizontal flip is performed for making a dataset ten times bigger than the original one.

3.2. Dual Enhanced Capsule Network (DE-Capsnet). The entire network has been shown in Figure 3, where the model is divided into portions. Firstly, we have to preprocess the images to avoid the uncorrelated information linked to the facial image. Then, we have two modules for further processing. In the first part, the box with the purple dashed line is attention aware of action units and consists of deep convolutional layers for the extraction of the enhanced features maps, and this has been termed as enhancement module 1. In the later part, with the use of dynamic routing, those enhanced feature maps are encoded between capsules, and the process of decoding is done by the fully connected layers (the process has been shown in the green dashed lines). At the end, the squashing function is used for the recognition of facial expressions.

VGG19 is used in enhancement module 1 because it is very much robust in object classification besides having a simple architecture. For a better understanding of the description, each stage is having multiple convolutional layers followed by a max-pooling layer. In the first 2 stages, each stage is having 2 convolutional layers. Whereas in the last 3

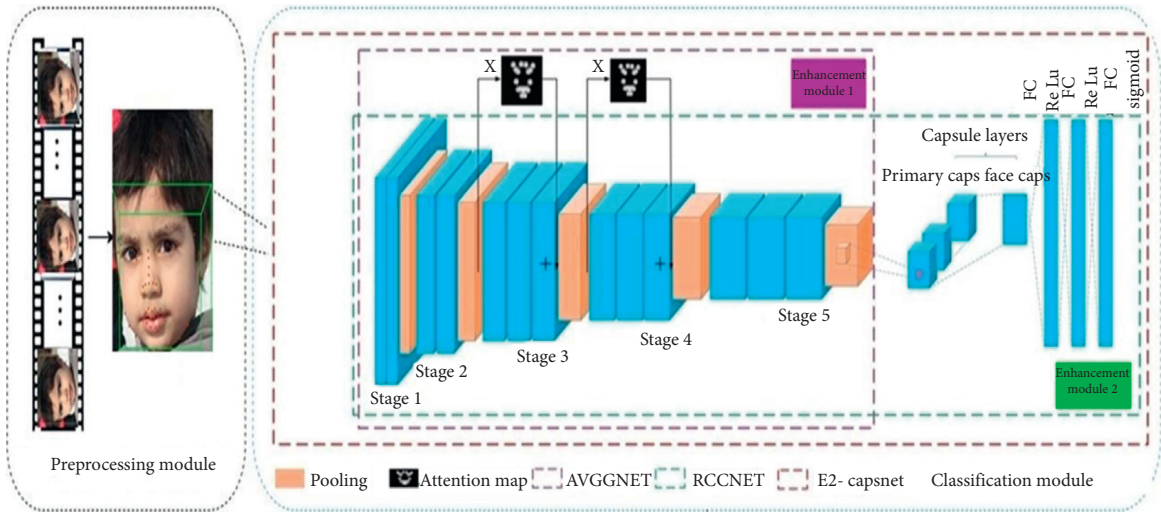


Figure 3: Overview of the proposed method.

stages, each stage is having 3 convolutional layers, respectively. We do not retain the last 3 layers as we have to get the feature maps.

To achieve the attention map, we have used the generation method by Li et al. [35]. Furthermore, we have made appropriate adjustments to the datasets used in our work for getting the key facial landmarks. Figure 3 shows the facial image with blue facial landmarks along with the attention map. Action unit's centres are obtained with key facial points by using scaled distance. To make sure that the scales must be the same among all the facial images, the facial images are resized. Hence, for making the shifting distance among the images as much adaptive as possible, the measurement reference is used to indicate the shift in the distance. To locate the action unit, centres of the inner corner distance have been used as scaled distance. For each action unit, the 7 pixels in the nearby area have been taken in the experiments, as a result, size of each action unit area is 15×15. $\mathbf{H_w}$ is assigned as the higher weight, which is the closest point to the action unit centre.

$$\mathbf{H_w} = 1 - 0.07\mathbf{m_d}. \tag{1}$$

The Manhattan distance is termed as $\mathbf{m_d}$ to action unit centre. Hence, those areas that are having higher values in the attention map correspond to the active areas of action units in facial images, and an attention map will further enhance them.

After the generation of attention maps, the maps are further forwarded to stage 3 and stage 4, as shown in Figure 3. The feature maps which are generated after the pooling layer of the second stage are multiplied with the attention map of the first stage, and after that being parallel with the convolutional layers of the third stage. Hence, the results obtained after the convolution are added element by element and then forwarded to the max-pooling layer of the current stage as an input. A similar operation is done at the fourth stage by jointly combining the convolutional layers with the attention map. Here, we explain the reason behind using attention maps; it is just because all areas are not equally important for facial expression recognition.

After the enhancement module 1, we get $512 \times 7 \times 7$ feature maps. For the dynamic routing the feature maps are further fed between primary capsule layers and face capsule layers. Three fully connected layers are used for decoding and reconstructing the facial image. The nonlinear function, i.e., the squashing function is used for facial expression recognition, which is defined in equation (2) as follows:

$$\mathbf{u_k} = \frac{\|\mathbf{j_k}\|^2}{1 + \|\mathbf{j_k}\|^2} \frac{\mathbf{j_k}}{\|\mathbf{j_k}\|}, \tag{2}$$

where $\mathbf{k}$ is used for the capsule, and $\mathbf{u_k}$ and $\mathbf{j_k}$ are output and input vectors, respectively. $\mathbf{L_m}$ is the minimizing margin loss and $\mathbf{L_r}$ is the reconstructing loss used for updating the parameters in the network. Total loss is defined as $\mathbf{L_t}$. Loss function expressions are defined in the equations (3)–(5), respectively.

$$\mathbf{L_m} = \mathbf{I_{cc}}\,\mathbf{max}\left(0, \mathbf{b}^+ - \|\mathbf{u_{cc}}\|\right)^2 + \lambda\,(1 - \mathbf{I_{cc}})\mathbf{max}\left(0, \|\mathbf{u_{cc}}\| - \mathbf{b}^-\right)^2, \tag{3}$$

$$\mathbf{L_r} = (\mathbf{f_c} - \mathbf{f})^2, \tag{4}$$

$$\mathbf{L_T} = (\mathbf{L_m} + 0.0005\mathbf{L_r})^2, \tag{5}$$

where $\mathbf{cc}$ is termed as the classification category and for that particular category the indication function is denoted by $\mathbf{I_{cc}}$. The upper and lower boundaries are represented by $\mathbf{b}^+$ and $\mathbf{b}^-$, respectively. The f represents the original image, whereas $\mathbf{f_c}$ represents the reconstructed image. This classification is based on one party; the training and testing phase is done by that party. However, we propose a method through which the server will be in charge of training, and testing will be done by both parties collaboratively.

3.3. Information Security. A security algorithm is information-secure in the sense that its security springs purely from scientific theory. The thought of information secure communication was initiated by the applied scientist, Shannon; he further added that the one-time pad system records excellent security subject to the subsequent two conditions [36]:

(1) The key that randomizes the information ought to be random and will be used one time

(2) The length of the key had to be as long as the length of the information

Even if any rule randomizes its parameter and the above conditions are satisfied, it is still hard to unmask the parameters even if an adversary is having exceptional computation power; e.g., if the random pace is the same as the message space, and is adequate to 1024–bits, then prior and posterior probabilities are the same, i.e., there is no particular advantage to urging posterior probability than prior probability.

3.4. Privacy Preserving Security. The main theme is to ensure secure operation between the client and the server. Both of them want to communicate with each other, and for that purpose need to compute $u^T v + p$. Where u is a vector known to client and v is a vector known to server with p being a scalar. However, only client will know the outcome of $u^T v + p$.

Where u input to client is composed of integers and v input to server is composed of floating points. Since we tend to perform integer random numbers, the process of conversion into integers is achieved by scaling the elements of the vector v, and it is approximated to the nearest integer. We use the scaling factor s that is large in $(u^T sv + s * p)$. First client adds random numbers in the vector v and server does few operations and returns the result to client. So, the operation is made valid by first scaling the scalar p and vector v by scaling the factor s, then the outcome is divided by that scaling factor.

So one thing is for sure, i.e., the server won't know anything about the client input and the same will be the case with the client. The client will just get to know about the result without having any knowledge about server vector and scalar. Hence, the above process is called a two-party protocol, which is completely information secure. Figure 4 demonstrates that any unknown input face image of any identity can be applied to synthesize a realistic equivalent face image of any other image.

3.5. Facial Expression Recognition Based on Privacy Preserving. The first step of every procedure is to mark the basic requirements and then fulfill them accordingly. The 3 requirements meeting with this process are as follows:

Requirement 1: without using more sophisticated public encryption system.

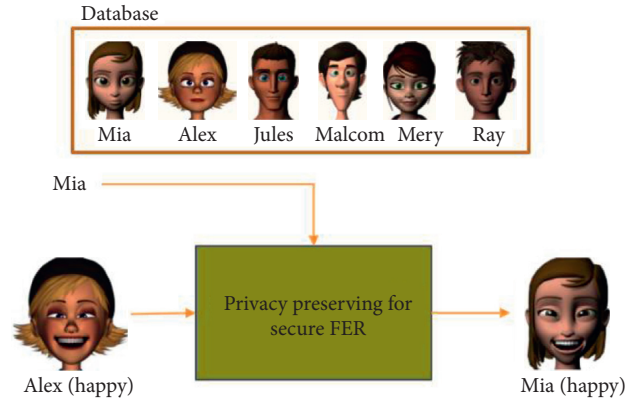


Figure 4: Application to an input face image of any identity to synthesize a realistic, expression-equivalent output face image of a target identity.

Requirement 2: hide a sample of client input data and server split result.

Requirement 3: hide server classification parameters and client will be unaware of means of database.

To explain this, let us break down the traditional assessment phase into four steps as follows:

Step 1. Find test image difference.

$$\text{test} = \widetilde{\text{test}} + \overline{b}, \tag{6}$$

where test is a difference test image, $\widetilde{\text{test}}$ is a testing image, and $\overline{b}$ is mean of database.

At the start, the client cannot send test image due to privacy issues. Therefore, the client only sends the image with the noise vector, $n \sim \in Z_n * 1$ having the same size as the test image. Since the server only receives the noise vector, it receives no information about the test image vector. So the difference noise vector is given as

$$n = \widetilde{n} + \overline{b}. \tag{7}$$

However, the difference between test image and noise image is just known to them. Let's represent it as

$$r = \widetilde{n} - \widetilde{\text{test}} = n - \text{test}. \tag{8}$$

Step 2. Illustration of the lower extremity difference.

$$a = B^T * \text{test}, \tag{9}$$

where B is the transformation matrix, a is a low dimensional vector corresponding to test. However, the server needs to project a low-dimensional vector with noise image given below:

$$\overline{a} = B^T * n. \tag{10}$$

Step 3. Euclidean distance calculation.

$$\overline{Ed_i} = \left\| \overline{a} - j_i \right\|_2^2, \tag{11}$$

where j_i is the training image (low dimensional) and $\overline{Ed_i}$ is the Euclidean distance and $i = 1, 2 \ldots..N$.

Step 4. Calculation of distance length to match the test image in the known section.

$$\overline{Ed_i} = Ed_i - m_i, \tag{12}$$

where the matching training image is denoted by m_i, but it is hard for the server to calculate the original distance Ed_i because the server doesn't know the r vector and the test image. So, in order to attain the matching image, the server will send all the $\overline{Ed_i}$, with a random number r_i for each $\overline{Ed_i}$ where

$$\widetilde{Ed_i} = \overline{Ed_i} + r_i = Ed_i + m_i + r_i. \tag{13}$$

Now the client can calculate the actual distance from the above equation as

$$Ed_i = \widetilde{Ed_i} - (m_i + r_i). \tag{14}$$

It is to make sure that only client knows the m_i; if the server gets to know about that, then he can calculate Euclidean distances between the provided test images and training images. Thus, the server will be able to find the expression corresponding with the test image and ultimately it will effect privacy.

3.6. Privacy Analysis. In this part of the research, we are interested in knowing whether our method is susceptible to any privacy leakage. Our method is based on the computation of both parties and therefore the only single possibility of privacy leakage can be the interaction between both the parties. To prove that our method does not leak unwanted information to a client or server, Goldreich's Privacy definition is used [37]:

3.7. Definition of Secured Privacy for Both Parties' Computation. The protocol we use for security should not disclose the hidden information to a third party (semi-honest) except the information that can be triggered by looking at the input and output of those parties.

Our primary purpose is to verify whether the proposed two-party calculation satisfies the definition of privacy or not. In the above four steps, it is clearly mentioned to the client and the server about their inputs and outputs. Therefore, we have to make sure that both of them don't infer other than the known inputs and outputs so that the proposed method would make sure that privacy is assured.

The client's ultimate goal is to make sure that the server is unaware of the test image and also just keep away the classification result. On the other side, the server had to keep the classification parameters away from the client. The client will just share the noise image initially, instead of sharing the true/original image; however, the size of both the images will remain the same. So, the server will know the size, and it will not be a privacy leakage. In return, the server also shares the random Euclidean distance obtained with the help of a random integer. Hence, information-theoretic security is achieved.

4. Results and Discussion

We have used four most popular databases for populating the results. These databases are CK+ [38], MMI [39], Oulu-CASIA [40], and RAF [23]. The RAF is used for large posed and real-world expressions, as the first three don't have large posed expressions. So to check the robustness of our method over large posed expressions, we have used the RAF data base.

4.1. Description of Databases. The Extended Cohn-Kanade database is the widest and the most popular database used in facial expression recognition. It contains 593 video sequences, which do vary from 10–60 frames with a shift from neutral to other expressions. There are a total of 123 subjects who performed different expressions, the ages of the subjects ranging from 18 to 30 years. Out of the 123 subjects, most of them are females. A total of 327 video sequences out of them are categorized into seven expressions. The core reason behind the algorithms not being uniform over CK+ is that it doesn't provide specific training, validation, and test sets.

The MMI database is laboratory-controlled and 75 subjects have performed 2900 expressions, both video sequences and static images with high resolution, out of which 326 video sequences are obtained from 32 subjects. The MMI database is different from CK+ as it uses onsets, offsets, and apex phases. In the sequences, the neutral expression is performed at the start of every sequence and reaches the peak and then returns back to the neutral expression. This database has very challenging conditions, i.e., it takes care of large inter-personal variations; every subject is performing different nonuniform expressions while wearing glasses, mustaches, etc.

The Oulu-CASIA database consists of 2880 images from 80 subjects for six expressions; most of them are males aged between 23 and 58 years. This database is specially designed to tackle the problem of illumination due to environmental changes. It consists of two different imaging systems; the first one is Near Infrared (NIR), whereas the second one is Visible Light (VIS). There are 3 different variable illumination scenarios: the first one is normal indoor illumination; the second one is used for weak illumination considering the scenario where just the computer display is on; and the third one is having all the lights off, i.e., dark illumination.

The Real-world Affective Faces Database is used, which consists of 29672 great, diverse real-world facial images. These images are downloaded from the Internet based on the approach of crowdsourcing; 40 annotators are used for independently labeling each image. This database consists of the large variability in different subjects' gender, age, ethnicity, varying lighting conditions, head pose, eye gaze, occlusions, and post-processing operations, which helps us to validate our network over versatile databases.

4.2. Implementation Details. The facial image is first preprocessed using face detection, data augmentation, and illumination normalization for fine-tuning of the image. The highly uncorrelated data are removed in order to process them further for a high-quality result. Then, the landmark detection is used to identify the key facial points. After that, VGG19 is used as a backbone of the network, where feature maps of $512 \times 7 \times 7$ are obtained after the 1st enhancement module. Then, $256 \times 6 \times 6$ feature maps are obtained from 2×2 convolutional kernels having the stride value of one; those feature maps are further forwarded to primary capsule layers with an 8D capsule and 32 convolutional layers. There are 3 routing iterations which are then executed between the primary capsule layers and the Face Capsule layers. Every expression is having 16D Capsules, where all the lower capsules forward information to the above capsule. Then with 3 fully connected layers, we use the squashing function for further classification. Adam optimizer is used for learning with a rate of 0.0001. The value of $\mathbf{b}^+$ is 0.9 and $\mathbf{b}^-$ is 0.1. Furthermore, the batch size is set to 16 and the maximum iteration is set to 300. Our whole network training is end to end.

In the Extended Cohn-Kanade database, we take the last frame to three frames and consider the first frame as a neutral expression for data selection. The subjects have been divided into a group of 10, and a 10-fold cross-validation is performed. Table 1 shows the average accuracy rates compared with other existing state-of-the-art methods. Our image-based method achieves the highest accuracy of 98.95 percent against sequence-based techniques that extract the features from a sequence of images or videos.

In the MMI database, we take three frames from the middle of each sequence that is associated with peak information and develop a dataset consisting of 624 images. Afterward, the data augmentation is performed and then distributed among 10 sets. For experimentation, the 10 cross-fold person independent validation is performed using the first frame, i.e., neutral expression, and it takes three peak frames from every frontal sequence. Table 2 shows the dominance in the average accuracy rates compared with other existing methods.

In the Oulu-CASIA database for training and testing, we use the last three frames from every sequence. A 10-fold cross-validation is performed just like CK+ in which based on the subject, each fold is completely disjointed with all the remaining folds. Table 3 shows the average accuracy rates, which outperform all novel methods. It achieves the highest accuracy of 91.2 percent.

Just like other databases in the RAF database, we perform a 10-fold cross-validation too. Table 4 shows the average accuracy rates of our method on the RAF database. We first obtained the true positives, false positives, true negatives, and false negatives, and then over 10 folds we calculated the F1 score and precision per class. Figure 5(a) shows the per-class precision and Figure 5(b) shows the per-class F1 score on the following databases.

4.3. Threats to Validation. There are a few factors that can enhance the robustness of facial expression recognition. While validating our approach, there are some limitations to the existing publicly available novel databases. The recognition of the expression with a closed mouth is less accurate as compared with the expression with an open mouth. Considering the agreement of facial expressions by face angles, we noticed that perceived arousal from the frontal face is more than compared with the shift in face angle. The happiness, disgust with closed mouth, and surprise remains unaffected with the face turned away. Furthermore, the

Table 1: The performance comparison of different approaches on the CK + database.

Method	Accuracy
LBP TOP [41]	88.99
HOG 3D [16]	91.44
MSR [42]	91.40
STM-explet [43]	94.19
DTAGN [44]	97.27
3D-CNN-DAP [45]	92.4
NMF-SSCCA [46]	97.3
FER-MPI-SFL (baseline) [47]	98.2
(Ours)	98.95

Table 2: The performance comparison of different approaches on the MMI database.

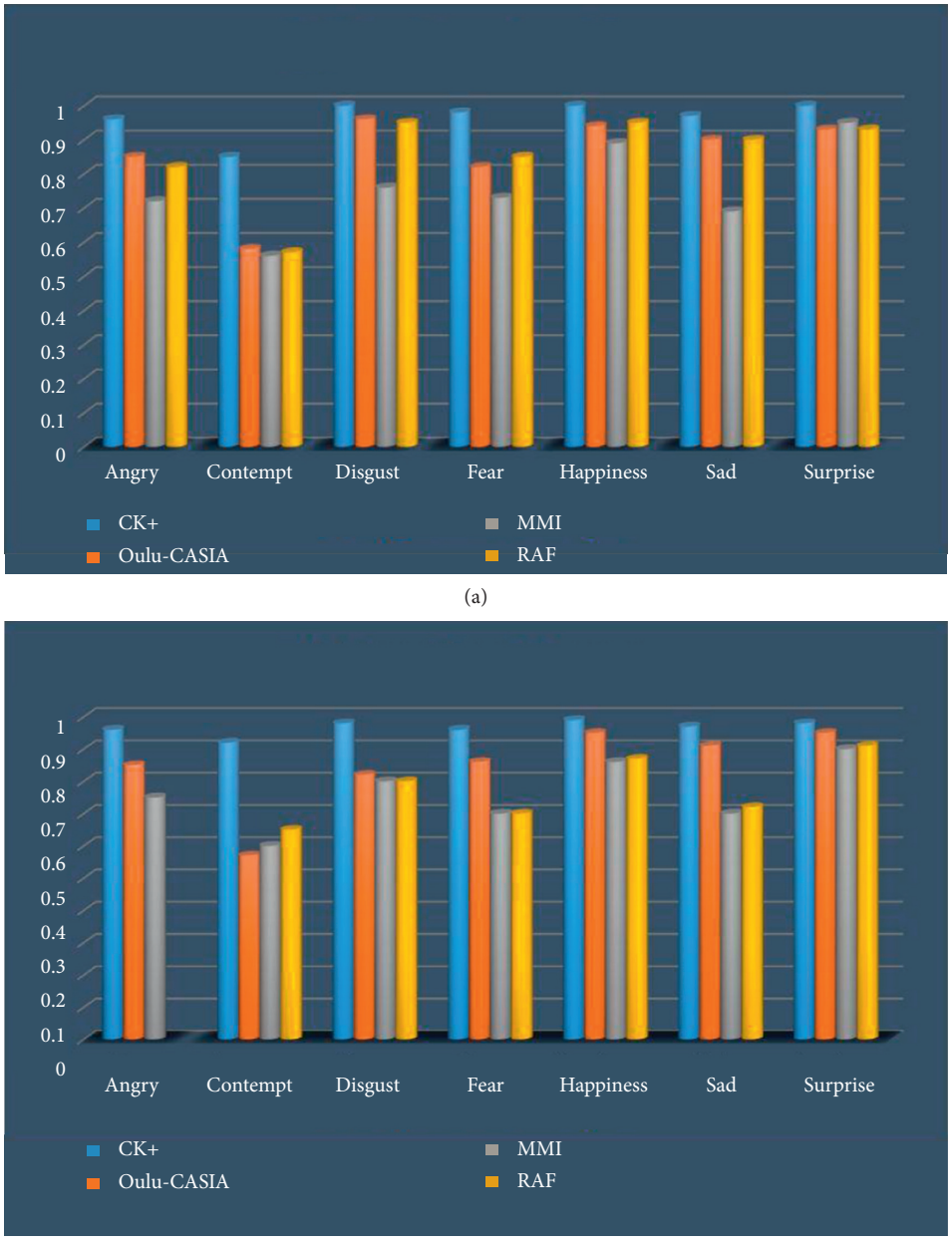
Method	Accuracy
LBP TOP [41]	59.51
HOG 3D [16]	60.89
CSPL [48]	73.53
STM-explet [43]	75.2
DTAGN-joint [44]	70.3
3D-CNN-DAP [45]	63.4
FER-MPI-SFL (baseline) [47]	83.1
(Ours)	89.31

Table 3: The performance comparison of different approaches on Oulu-CASIA database.

Method	Accuracy
LBP TOP [41]	68.1
HOG 3D [16]	70.6
STM-explet [43]	74.59
Atlases [49]	75.52
DTAGN-joint [44]	81.46
FN2EN [50]	87.71
PPDN [51]	84.59
FER-MPI-SFL (baseline) [47]	87.39
(Ours)	91.2

Table 4: The performance comparison of different approaches on RAF database.

Method	Accuracy
E2E-FC	23.99
AIR [52]	67.37
NAL [53]	84.22
IPA2LT (EM [54] + CNN)	85.30
IPA2(LTNET) (baseline) [55]	86.77
(Ours)	97.15



(a)

(b)

Figure 5: Performance metrics on four databases. (a) Per-class precision on four databases. (b) Per-class F1 score on four databases.

effective valence near the frontal is conveyed more by the full left-side profile rather than the full right side profile. It is because of this reason that the left hemiface observes a more spontaneous response than the right hemiface. The facial expression analysis can be enhanced by the facial motion information if the image is subtle or degraded. The dynamic neutral expression with the blinking of eyes or chewing is also a threat. Moreover, the dwell time is also a key factor; it takes more time over eyes than the mouth. However, the dwell time over the mouth of happy expression is relatively high. With an increase in the intensity, it can also be noticed that the accuracy is also increased, whereas the dwell time and round trip is decreased. Overall, the response time of females is faster than males even in a low-intensity environment. In the end, it was also concluded that the dwell time of the female eye is more than that of the male.

5. Conclusions

In this paper, we have introduced a state-of-the-art architecture that is robust and effective. A facial image is first preprocessed using different techniques to counter the problems of the excessive background, limitation of data, varying illumination, pose-variation, and occlusion. The facial image is fine-tuned and then forwarded to a dual enhanced capsule network that is capable of handling the

spatial transformation. It uses action units aware mechanism, which helps to locate the active areas, which can help in better facial expression recognition. The feature representation ability is enhanced due to multiple convolutional layers and it helps to capture the key information present in the particular structure of the face. We performed the privacy preservation with the help of a randomization technique, which added the benefit of less computationally expensive. It also performs secure communication between the two untrusted parties.

Different databases have different sets of pictures under varying conditions. As a result, class imbalance occurs due to the inconsistency in expression annotations. So a cost-sensitive layer can be enhanced for training the deep neural networks. Meanwhile, a powerful, deep neural network can be designed to have prior knowledge of the change in the local environment, which is capable of predicting specific parameters and inherently handling and recovering facial occlusions without any intervention. Furthermore, to improve the robustness of the FER, it can be fused with other models. The incorporation with other modalities like depth information from three-dimensional face models, neurosciences, cognitive sciences, infrared images, and physiological data can be a good future research direction.

Data Availability

All the data are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] N. Samadiani, G. Huang, B. Cai et al., "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors*, vol. 19, no. 8, p. 1863, 2019.

[2] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, "Facial expression analysis under partial occlusion: a survey," *ACM Computing Surveys (CSUR)*, vol. 51, p. 25, 2018.

[3] A. K. Vail, T. Baltrušaitis, L. Pennant, E. Liebson, J. Baker, and L. P. Morency, "Visual attention in schizophrenia: eye contact and gaze aversion during clinical interactions," in *Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 490–497, IEEE, San Antonio, TX, USA, October 2017.

[4] A. Ullah, J. Wang, M. S. Anwar, U. Ahmad, U. Saeed, and Z. Fei, "Facial expression recognition of nonlinear facial variations using deep locality de-expression residue learning in the wild," *Electronics*, vol. 8, no. 12, p. 1487, 2019.

[5] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, p. 124, 1971.

[6] J. Kumari, R. Rajesh, and K. M. Pooja, "Facial expression recognition: a survey," *Procedia Computer Science*, vol. 58, pp. 486–491, 2015.

[7] A. Ullah, J. Wang, M. S. Anwar, U. Ahmad, U. Saeed, and J. Wang, "Feature extraction based on canonical correlation analysis using FMEDA and DPA for facial expression recognition with RNN," in *Proceedings of the 2018 14th IEEE International Conference on Signal Processing (ICSP)*, pp. 418–423, IEEE, Beijing, China, August 2018.

[8] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Identity-free facial expression recognition using conditional generative adversarial network," 2019, http://arxiv.org/abs/1903.08051.

[9] H. Yang, Z. Zhang, and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 294–301, IEEE, Xi'an, China, May 2018.

[10] N. Van Quang, J. Chun, and T. Tokuyama, "CapsuleNet for micro-expression recognition," in *Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–7, IEEE, Lille, France, May 2019.

[11] W. H. Organization and others, "Depression: key facts," 2018.

[12] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[13] M. Zhu, D. Shi, and J. Gao, "Branched convolutional neural networks incorporated with jacobian deep regression for facial landmark detection," *Neural Networks*, vol. 118, pp. 127–139, 2019.

[14] Y. Fan, J. C. Lam, and V. O. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *Proceedings of the 27th International Conference on Artificial Neural Networks*, pp. 84–94, Springer, Rhodes, Greece, October 2018.

[15] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, pp. 356–370, 2018.

[16] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *Proceedings of the British Machine Vision Conference 2008*, Leeds, UK, September 2008.

[17] W. Cheung and G. Hamarneh, "*n*-SIFT: *n*-Dimensional scale invariant feature transform," *IEEE Transactions on Image Processing*, vol. 18, no. 9, pp. 2012–2021, 2009.

[18] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proceedings of the Face and Gesture 2011*, pp. 314–321, IEEE, Santa Barbara, CA, USA, March 2011.

[19] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Proceedings of the Asian conference on computer vision*, pp. 143–157, Springer, Singapore, November 2014.

[20] B. K. Kim, H. Lee, J. Roh, and S. Y. Lee, "Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 427–434, ACM, Seattle, WA, USA, November 2015.

[21] H. W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 443–449, ACM, Seattle, WA, USA, November 2015.

[22] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 435–442, ACM, Seattle, WA, USA, November 2015.

[23] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2852–2861, Honolulu, HI, USA, July 2017.

[24] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 558–565, IEEE, Washington, DC, USA, June 2017.

[25] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 302–309, IEEE, Xi'an, China, May 2018.

[26] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2168–2177, Salt Lake City, UT, USA, June 2018.

[27] K. Ali, I. Isler, and C. Hughes, "Facial expression recognition using human to animated-character expression translation," 2019, http://arxiv.org/abs/1910.05595.

[28] Z. Erkin, M. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft, "Privacy-preserving face recognition," in *Proceedings of the 9th International Symposium on Privacy Enhancing Technologies, PETS'09*, pp. 235–253, Seattle, WA, USA, August 2009.

[29] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[30] Y. Rahulamathavan, R. Phan, J. Chambers, and D. Parish, "Facial expression recognition in the encrypted domain based on local Fisher discriminant analysis," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 83–92, 2012.

[31] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Advances in Cryptology EUROCRYPT 99*, pp. 223–238, Springer, Berlin, Germany, 1999.

[32] X. Liu, H. Robert, Deng, Kim-Kwang, R. Choo, and J. Weng, "An efficient privacy-preserving outsourced calculation toolkit with multiple keys," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 11, 2016.

[33] Y. Liu, Z. Ma, X. Liu, S. Ma, and K. Ren, "Privacy-preserving object detection for medical images with faster R-CNN," in *Proceedings of the IEEE Transactions on Information Forensics and Security*, Barcelona, Spain, October 2019.

[34] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI, USA, June 2003.

[35] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "EAC-net: deep nets with enhancing and cropping for facial action unit detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 11, pp. 2583–2596, 2018.

[36] C. E. Shannon, "Communication theory of secrecy systems*," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.

[37] O. Goldreich, "Secure multiparty computation," 1998, http://www.wisdom.weizmann.ac.il/oded/pp.html.

[38] P. Lucey, J. F. Cohn, T. Kanade, and J. Saragih, "The Extended Cohn-Kanade Dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *Proceedings of the Computer Vision and Pattern Recognition Workshops*, San Francisc, CA, USA, June 2010.

[39] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Amsterdam, Netherlands, July 2005.

[40] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[41] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[42] R. Ptucha and A. Savakis, "Manifold based sparse representation for facial understanding in natural images," *Image Vision Computing*, vol. 31, no. 5, pp. 365–378, 2013.

[43] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014.

[44] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, CL, USA, December 2015.

[45] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Proceedings of the ACCV*, Singapore, November 2014.

[46] A. Ullah, J. Wang, M. S. Anwar, U. Ahmad, J. Wang, and U. Saeed, "Nonlinear manifold feature extraction based on spectral supervised canonical correlation analysis for facial expression recognition with RRNN," in *Proceedings of the 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–6, Beijing, China, October 2018.

[47] W. Wu, Y. Yin, Y. Wang, X. Wang, and D. Xu, "Facial expression recognition for different pose faces based on special landmark detection," in *Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 1524–1529, IEEE, Beijing, China, August 2018.

[48] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2562–2569, IEEE, Providence, RI, USA, May 2012.

[49] Y. Guo, G. Zhao, and M. Pietikäinen, "Dynamic facial expression recognition using longitudinal facial expression atlases," in *Proceedings of the European Conference on Computer Vision*, pp. 631–644, Springer, Florence, Italy, October 2012.

[50] H. Ding, S. K. Zhou, and R. Chellappa, "FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition," 2016, http://arxiv.org/abs/1609.06591.

[51] X. Zhao, X. Liang, L. Liu et al., "Peak-piloted deep network for facial expression recognition," in *Proceedings of the European Conference on Computer Vision*, Amsterdam, Netherlands, October 2016.

[52] S. Azadi, J. Feng, S. Jegelka, and T. Darrell, "Auxiliary image regularization for deep cnns with noisy labels," 2015, http://arxiv.org/abs/1511.07069.

[53] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," 2016.

[54] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 20–28, 1979.

[55] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 222–237, Munich, Germany, September 2018.