

Article

Facial Expression Recognition of Nonlinear Facial Variations Using Deep Locality De-Expression Residue Learning in the Wild

Asad Ullah ^{1,2,*} , Jing Wang ^{1,*} , M. Shahid Anwar ¹, Usman Ahmad ³ and Uzair Saeed ³ and Zesong Fei ¹

¹ School of Information and Electronics, Beijing Institute of Technology, Haidian District, Beijing 100081, China; shahidanwar786@gmail.com (M.S.A.); feizesong@bit.edu.cn (Z.F.)

² Riphah International University, Faisalabad Campus, Faisalabad 38000, Pakistan

³ School of Computer Science and Technology, Beijing Institute of Technology, Haidian District, Beijing 100081, China; usmanahmad@bit.edu.cn (U.A.); uzairsaeed@bit.edu.cn (U.S.)

* Correspondence: engr.asadihsan@gmail.com (A.U.); wangjing@bit.edu.cn (J.W.)

Received: 20 November 2019; Accepted: 2 December 2019; Published: 6 December 2019



Abstract: Automatic facial expression recognition is an emerging field. Moreover, the interest has been increased with the transition from laboratory-controlled conditions to in the wild scenarios. Most of the research has been done over nonoccluded faces under the constrained environment, while automatic facial expression is less understood/implemented for partial occlusion in the real world conditions. Apart from that, our research aims to tackle the issues of overfitting (caused by the shortage of adequate training data) and to alleviate the expression-unrelated/intraclass/nonlinear facial variations, such as head pose estimation, eye gaze estimation, intensity and microexpressions. In our research, we control the magnitude of each Action Unit (AU) and combine several of the Action Unit combinations to leverage learning from the generative and discriminative representations for automatic FER. We have also addressed the problem of diversification of expressions from lab controlled to real-world scenarios from our cross-database study and proposed a model for enhancement of the discriminative power of deep features while increasing the interclass scatters, by preserving the locality closeness. Furthermore, facial expression consists of an expressive component as well as neutral component, so we proposed a generative model which is capable of generating neutral expression from an input image using cGAN. The expressive component is filtered and passed to the intermediate layers and the process is called De-expression Residue Learning. The residue in the intermediate/middle layers is very important for learning through expressive components. Finally, we validate the effectiveness of our method (DLP-DeRL) through qualitative and quantitative experimental results using four databases. Our method is more accurate and robust, and outperforms all the existing methods (hand crafted features and deep learning) while dealing the images in the wild.

Keywords: facial variations; deep locality preserving; de-expression residue; conditional graphical adversarial network

1. Introduction

In recent years, the interest in facial expression analysis has been tremendously increased. It is considered to be one of the most powerful and essential aspects in recognition of emotional and mental states [1,2]. Facial expression analysis has a massive range of practical and potential applications, such as social robotics, intelligent tutoring systems, medical treatment, personalized service provision, driver fatigue surveillance, and various other augmented and virtual reality systems [3]. Still, automatically

learning the facial expression analysis in wild and real-time situations is quite difficult. Millions of images from daily life events are uploaded over network mostly on social media. To learn different aspects of face images automatically, large annotated databases are needed. Additionally, proper Action Unit (AU) coding [4] needs particular expertise to precisely learn and be perfect which may take months, therefore alternative solutions must be provided.

The role of cultural differences in facial emotion's is also a problem in defining definite prototypical AUs. Therefore, it is necessary to study the effective and emotional states from judgments based on majority common population. Meanwhile, facial landmark detection, head pose orientation and motion, and eye gaze estimation play a vital role in facial behavior analysis, because it is important to know about emotional attributes, attentiveness, and mental health [5,6]. Partial occlusion is another major obstacle in facial behavior analysis while considering unconstrained, in the wild and real-time scenarios. There are lot of chances that face may be obstructed by a hat, sunglasses, hairs, mustache, hand moving over mouth, etc. It can deteriorate the overall performance and effectiveness of the system. The presence of the occlusion will affect the extraction of discriminative features because of imprecise feature location, error in face alignment and inaccurate face registration. The recognition of microexpressions (MEs) is also important for [7] incremental face alignment in real time and proper understanding of deceitful human behaviors. MEs are brief; they indicate the underlying and genuine emotions rapidly.

In our work, we propose a common perception of expressions by a reliable unconstrained crowdsourcing approach. Automatic facial expression recognition require fine tuned facial images and for that purpose we need to preprocess the images. The preprocessing step involves face detection, data augmentation, and illumination normalization to provide as much as semantic visual information contained in the image. As most of the images have excessive background so it is useless to process whole image so face detection is required to process the meaningful data. Although most of the datasets have almost frontal images, as moving in to wild imagery requires better face detection, in this paper, state of the are method is used for face detection [8]. Data augmentation is the next step which is equally important as most of the datasets dont have enough images so you need sufficient training data for validation of results, we used on-the-fly augmentation [9]. Illumination normalization is also important just because even a same person is having same expressions still they can be varied in the existence of unconstrained environment resulting in huge intraclass variances. Histogram equalization and linear mapping has been used together to improve results. We bridge the gap like head pose orientation and motion, eye gaze estimation, occlusion, and microexpressions, which are nonlinearly tied with facial expressions. Although there are a lot of tools available still there is a room for proper landmark detection, we have used CE-CLM for both landmark detection and head pose estimation. For proper estimation of eye gaze our model has used CLNF. We also address the large intraclass variability and learning effective representations of expressions. Our proposed method is able of handling such situations with the support of a new and optimized face detector and landmark detector. The pipeline of our action units recognition is explained in Figure 1. Apart from variable intensity, poses, spontaneous responses and occlusions faces are also affected by two problems. First, expressions may be having different action unit combinations so for that purpose proper classification model is proposed for modeling multi-modality distribution of every facial expression in feature space. Second, due to crowdsourcing faces may be affected by multiple or compound emotions. So conventional methods do not work well over such images although they perform well over lab controlled environment. We proposed a model which is capable of handling such problems. Our method outperforms all other methods and is successful in predicting the precise feature descriptors that work best with sparse representation. The method is robust and achieves the best accuracy for facial expressions under wild conditions.

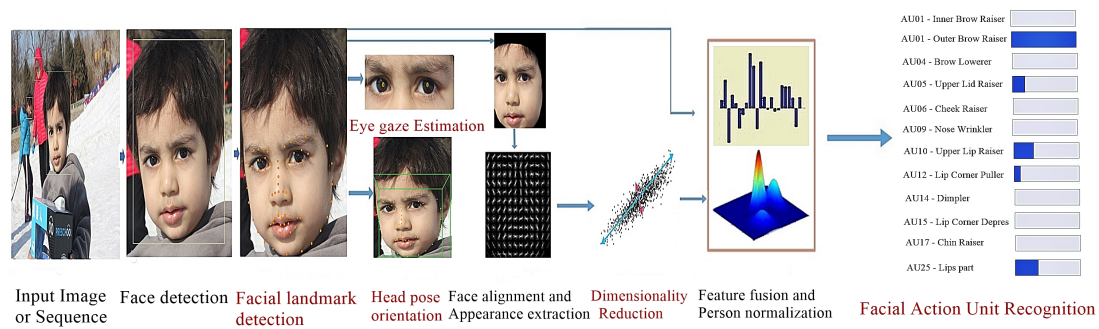


Figure 1. Pipeline of our proposed method.

2. Related Work

Facial expression representation is quite essential for evaluating the efficiency of FER systems, more precisely, in the existence of occlusion. Furthermore, facial expressions are generally associated with two key aspects i.e., the alike jobs (facial expression synthesis with occluded facial components) and associated methods (attention mechanism). Occlusion is considered to be a key obstacle targeting real-world facial expression synthesis problems, e.g., facial expression analysis, estimation of age, classification of gender, etc. Previously facial occlusions have been addressed by two main approaches: holistic-based approaches or addressed by part-based approaches. The former method's take the face as a complete, instead of further dividing it into mini-regions.

2.1. Holistic-Based Approaches

One holistic approach is by making a generative model which will be capable of reconstructing a whole face from the occluded face. These methods depend on the data provided for training with varying conditions of occlusion. Especially, Kotsia et al. examined that how Facial expression recognition performance is disturbed by partial occlusions and determined that occlusion over mouth causes a massive drop-off in Facial expression recognition performance than the corresponding occlusion over eyes [10]. To solve the occlusion-related problems, the strength of the features is being usually improved through designated regularization, i.e., L1-norm [11]. This notion is considered to be appropriate for non-facial occlusions, e.g., Osherov and Lindenbaum suggested to jointly work out for the re-weight L1 regularization to tackle with arbitrary occlusions in the end-to-end framework [12].

2.2. Part-Based Approaches

Conversely, the face is split into numerous overlapped/nonoverlapped segmentations by part-based methods. Looking into the patches over the face, available research either attains patches present around the facial landmarks, split the face into a number of uniform parts, or attain patches by some sampling set of procedures, or perceive occluders explicitly [13]. Next, the part-based approaches perceive and work on compensating the missing portion [14,15], the removal of occluded areas [16], or assigning the weights to both the occluded and nonoccluded patches separately [14,17].

ACNN is different from holistic based and part-based because it need not explicitly tackle/detect occlusions and secondly it unifies representation learning and patterns of occlusions in an endwise CNN. Past work on person re-identification [18], gender classification [19], and Facial expression recognition [20] has also implemented approaches of combining local and global features. The gACNN varies from afore mentioned approaches by inserting an endwise trainable Gate Unit (GU). GU not just only know different occlusion pattern's from the data which is provided and express them with weights of the model in code, but it is also capable of weighing various patches for image keeping in mind there is no occlusion. Thus it is estimated to attain much better and consistent Facial expression recognition output on both affected and nonaffected face images.

Human beings have the capability to move quickly towards prominent entities in a visual scene (cluttered) [21], i.e., we position our gaze rapidly towards the interested entities in scene. This sort of mechanism has been effectively implemented in various computer vision and pattern recognition problems, which consists of image caption [22], fine-grained recognition of images [23], person re-identification [24], and visual query answering [25]. An RNN/LSTM approach has been taken to foresee the subsequent attention region which relies on present attention region position having visual features. This framework has been used for visual query answering as well as the image caption. Moreover, Zhao et al. assessed multiple two-dimensional attention maps, having convolutional feature maps of same spatial size in contrast to the weight [23]. That is a simple approach but the problem is it does not consider occlusion patterns. Juefei-Xu et al. train sample images of several blurred levels in order to impose the shift in attention throughout the process of learning [26]. For gender classification, this model is more effective to occlusions. First, a face image is split into blocks; then, through reinforcement learning, the spatial attention regulating mechanism over the blocks is erudite [27]. Attention models permit significant features to come at the front as required. This is quite helpful in a case where there are occlusions presented in an image. Compared with the existing models, we have to adopt facial landmarks for decomposing the region, which is quite effective and easy to implement. Patch based CNN with attention has been implemented with the attention mechanism for occluded images [28]. Highly discriminative feature vectors are extracted using gentle boost decision trees for every basic expression around distinct facial landmark points [29]. Partial occlusion has been addressed using both texture and geometric feature based approaches for the better classification of facial expressions [30]. Privileged information has been taken from the non occluded faces for the recognition of facial expressions under occlusion. The nonoccluded facial images are just used for training purpose [31].

3. Proposed Method

3.1. Preprocessing

A preprocessing step is important to provide meaningful features for further processing of facial images, i.e., illuminations, backgrounds, and head poses are irrelevant and are present in unconstrained scenarios. Therefore, the meaningful features are required ahead of training deep neural network. This step is needed to detect and normalize the visual information conveyed by the face.

3.1.1. Face Detection

Face detection is very important just because it is needed before any facial expression analysis process. Excessive background in an image which is highly uncorrelated to expression recognition generally exists even if image is taken from few benchmarking datasets. Meanwhile, the majority of standard datasets contain good quality and high-resolution images with almost near-frontal head poses and very little partial occlusions. Therefore, the Viola and Jones face detection algorithm [32] has proven to be very efficient in most of the cases.

However, the better face detection algorithm is highly required when we move to the wild imagery. As practical applications not necessarily have the frontal view or can have multiple faces in an image, so V & J algorithm has been failed in such circumstances. For attaining high performance of FER, state-of-the-art algorithm is required. We have used Cascade Deformable Part Models (CDPM) which is multi-view face detector and can detect in the wild under the existence of occlusion and spanning of head poses [8].

3.1.2. Data Augmentation

Most of the publicly available datasets lack a large number of images. Therefore, data augmentation is quite essential for validating the results as sufficient training data is required for deep

facial expression recognition. Data augmentation consists of two types: on-the-fly data augmentation and offline data augmentation.

For alleviation of overfitting the on-the-fly data, augmentation has been used. The process is manipulated by randomly cropping the input samples from four sides and after that a horizontal flip is done which results in making a dataset which is 10 times bigger than the actual one. There are many offline data augmentation operations like rotation, shifting, skewing, scaling, color jittering, contrast, etc. Combinations of different operations results in the unseen training samples give a more robust and effective performance. A more detailed affine transformation matrix has been used which is capable of generating random images which vary in terms of skew, rotation, and scale [9].

3.1.3. Illumination Normalization

Illumination normalization is also very important because even if the same person is having same expressions still images can be varied in the existence of unconstrained environment resulting in huge intraclass variances. From the studies, it has been noted that combination of illumination normalization with histogram equalization yields to great facial recognition performance than individual illumination normalization. However, merely applying histogram equalization is good when the brightness of foreground and back ground are similar, but it may overemphasize local contrast. A weighted summation approach has been used in our work for combining histogram equalization and linear mapping to improve results.

3.2. Facial Landmark Detection and Tracking

This is also a key step to perform for FER. Variety of facial landmark detection tools available but there is always room to suggest a state of the art method. Convolutional Experts Constrained Local Model (CE-CLM) has been proposed for detection and tracking of facial landmarks. This method consists of a series of optimizations to improve overall performance. Firstly, the model size has been reduced from 180,000 parameters to 90,000 parameters having 68 landmarks for each, which increased the speed by 1.5 times and there was very minimal change in the accuracy, still, the accuracy outperformed the other existing models. Secondly, multiple initialization hypothesis has been used for profile and in the wild images at different orientations. Looking for the best likelihood out of the 4 scales, if one of the scales is above threshold so other hypotheses are not evaluated. If we do not have any scale above the threshold, three best hypotheses with maximum likelihood is evaluated and the best one is picked and in result, the performance is improved 4-fold.

3.3. Head Pose Estimation

Despite facial landmark detection, the head pose (translation and orientation) is also implemented by our model. With the use of orthographic camera projection, CE-CLM by internally using a three-dimensional representation of landmarks projects head pose estimation to the image. Meanwhile, the camera calibration parameters give more robustness to more accurate and precise head pose estimation which helps in better overall performance.

3.4. Eye Gaze Estimation

The constrained local neural field (CLNF) is used in our model for detecting eyelids, iris, and pupil for the proper estimation of eye gaze. The SynthesEyes dataset is used for validating the training purpose. The eye gaze vector is computed by detecting the eye location and the pupil by looking into the coordinates of the eyeball. The results are quite promising and accurate. Landmark detection, head pose estimation and eye gaze estimation is shown in Figure 2.

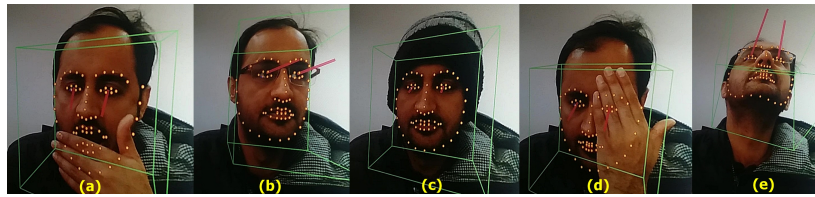


Figure 2. Landmark detection, head pose estimation and eye gaze estimation of partial occluded facial images in real-time (from left to right: (a) hand over mouth; (b) wearing glasses; (c) wearing hat; (d) hand obstructed the face; (e) looking at the top).

3.5. Deep Preserving Feature Learning

Along with the in the wild challenges like variable intensity, poses, spontaneous responses, and occlusions faces are also affected by two challenges. First, real-time expressions may be linked with different action unit combinations, so for that purpose, a proper classification algorithm is required for multi-modality modeling distribution of every facial expression in the feature space. Next, due to crowdsourcing, a huge amount of real-time affective faces contains multiple or compound emotions. Therefore, conventional hand-crafted representations that work well on the lab controlled environment do not perform well on the facial expression recognition under the wild. Nevertheless, conventional DCNN has outperformed hand engineered features on a big scale in visual recognition problems. However, still, there is a gap for intraclass variation because real-time faces do have larger intraclass differences in the presence of nonlinear facial variations.

In this paper, we proposed Deep Locality Preserving De-Expression Residue Learning (DLP-DeRL). It consists of two learning processes: the first process is used to generate neutral face by Conditional Generative Adversarial Network's (cGANs), and the other process is used for learning from the middle layers provided in the generator. Two face images pairs are used as an input for training the cGANs. F_{input} is a facial image which shows any expression and considering that particular subject, F_{target} is used for getting a neutral facial image. Then after training while keeping the identity information same, the resultant neutral facial image for any input is reconstructed by the generator. Therefore, from an expressive-face image to neutral, the information related to expression is recorded in the middle layers. For the next process, while we fix the generator parameters and combine it with the output of middle layers and then input it to deep models for expression classification.

3.5.1. Neutral Face Regeneration

Given an expressive facial image, by implementing cGAN a neutral facial representation is generated. Two different players are used by cGAN. One is Generator (G) and other is Discriminator (D). Min-max game is played with the discriminator for training the Generator in order to recover training data distribution. Input face image and output face image both are provided to cGAN for training. Discriminator tries to distinguish between target image and output image, whereas the generator generates an image, which is close to target image.

The discriminator's objective is expressed as

$$L_{cGAN}(D) = \frac{1}{M} \sum_{i=1}^M \{ \log D(F_{input}, F_{target}) + \log(1 - D(F_{input}, G(F_{input}))) \} \quad (1)$$

where M is used for total number of training image pairs.

The Generator's objective is given below.

$$L_{cGAN}(G) = \frac{1}{M} \sum_{i=1}^M \{ \log D(F_{input}, G(F_{input})) + \theta_1 \cdot \|F_{target} - G(F_{input})\|_1 \} \quad (2)$$

Here, overblurring of the output image is inclined as one loss and other loss is used for the image similarity. The final objective is given below.

$$G e^* = \arg \min_G \max_D L_{cGAN}(D) + \theta_2 \cdot L_{cGAN}(G) \quad (3)$$

3.5.2. Facial Expressive Component Learning

Afterward, neutral face is regenerated through comparison of neutral face expression and query face expression takes place; the expression information can be analyzed either at pixel level or it can be at feature level. However, due to various facial variations, the pixel level change is inefficient even without changes in expression. Therefore, the difference between a neutral image and query image is noted in the middle layers and the expressive components are exploited in those layers to solve the problem.

$$F_{\text{exp}=\text{neutral}}^{\text{id}=\text{B}} = G \left(F_{\text{exp}=\text{E}}^{\text{id}=\text{B}} \right) \quad (4)$$

where $F_{\text{exp}=\text{neutral}}^{\text{id}=\text{B}}$ represents an image with facial expressions, the generative model takes it as an input and generates a neutral image, E represents six basic facial expressions, and G is used for generator. In the above equation, we can notice that the image of subject (B) containing an expression so it can become a neutral face of that particular subject (B). Therefore, the second learning process is for the recording of the unique expression information of every individual in the middle layers of the generator and this distinctive information is termed “de-expression residue”. Furthermore, to address multiple modality and ambiguity of real-time expressions, a new supervised layer is added to the design named Locality preserving Loss (LP loss), to advance the discrimination capability of deep features. The locality of each sample is preserved and making the local neighborhoods as compact as possible within each class.

As shown in Figure 3, learning is done in the middle layers of the generator from de-expression residue, filters are fixed for those layers, and all the same size layers are concatenated and are given as an input to a local CNN model for expression recognition. Cost function of every local CNN model is recorded and concatenated with the Locality preserving loss. We further combine and concatenate all those fully linked layers of every local CNN model with the final layer. The total loss function is then represented as

$$T - \text{loss} = \lambda_1 l_1 + \lambda_2 l_2 + \lambda_3 l_3 + \lambda_4 l_4 + \lambda_5 l_5 \quad (5)$$

where weights are arranged properly to achieve highest recognition rates on the datasets.

People are seemed to be non-expressive often in natural interactions. Considering the long video sequence of an individual we can easily assume that the lowest intensity most of the time should be zero. However, almost all the available Action Unit predictors sometime under-estimate or overestimate the values of action units for a particular person. To avoid prediction errors. Lowest nth percentile of the predictions has been taken on a particular person and is subtracted from all predictions. This approach is termed as person calibration. It can be easily implemented with having histogram of previous predictions.

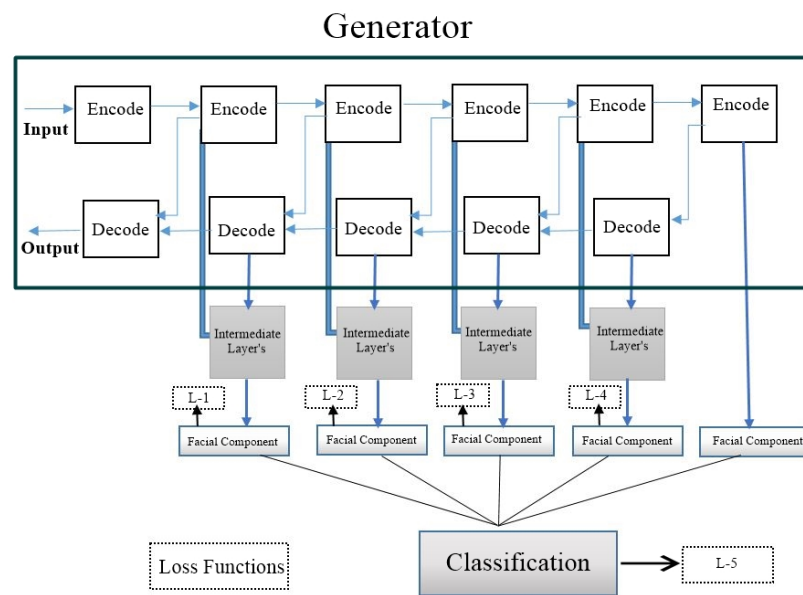


Figure 3. Framework of our proposed model DLP-DeRL.

4. Experimental Results

In this paper, we implemented a model that is able to immediately predict the expression and transfer result. We evaluated our results with four popular and widely used databases, i.e., Extensive Cohn Kanade [33], Oulu-CASIA [34], MMI [35] and BU-3DFE [36]. The results are explained with the help of confusion matrix because while dealing with multi-class classification we need to know which class is more dominant and towards which class the model is biased. The confusion matrix that is dominantly diagonal is termed as a good classifier [37].

We have used Cascade Deformable Part Models for detection of a face in the real world condition to remove unwanted data from the image. It is a multi-view face detector and performs well under the presence of occlusion and spanning of head poses. Then to alleviate the overfitting, data augmentation is done. Next, the weighted summation approach of combining histogram equalization and linear mapping is performed to enhance illumination. CE-CLM is used for facial landmark detection and tracking, it contains 68 landmarks for each image. Along with facial landmark detection, we have also performed head pose estimation and eye gaze estimation to achieve more robust and effective results.

The generative model is initially trained over BU-4DFE, in which 60,600 images are taken from 101 subjects. Each subject is composed of six sequences and each sequence further contains six basic facial expressions. To build a training dataset, every first frame is taken as a target image, whereas the remainder of the sequence is considered as input images. Then the model which is pretrained further applied to other databases for fine-tuning. Adam optimizer is used with taking the batch size of 160, the momentum of 0.95 and drop out of 0.5 for all the connected layers; 200 epochs are used to train generative model and 50 used for the classification model.

The Extensive Cohn Kanade (CK+) is a very popular database. This database consists of 593 video sequences obtained from 123 subjects. Out of them, 327 video sequences taken from 117 subjects and they are labeled as seven expressions. Just because Extensive Cohn Kanade is unable to provide specified training, validation as well as testing, uniform algorithms can't be implemented over this database. Therefore, for the static-based methods, the data selection is bit different. The last frame is termed as a general procedure. The final three frames with peak information are taken from each sequence provided with the label and generated 983 images. For the subjects in any two subsets being mutually exclusive, we have performed 10-fold cross-validation on the basis of identity. The average accuracy is shown in Table 1 after 10 runs. It was more robust, accurate, and outperformed other

methods: LBP-Top [38], HOG 3D [39], 3D CNN [40], STM-Explet [41], DTAGN [42], and CNN [43]. The confusion matrix on Cohn Kanade Extensive database is shown in Figure 4.

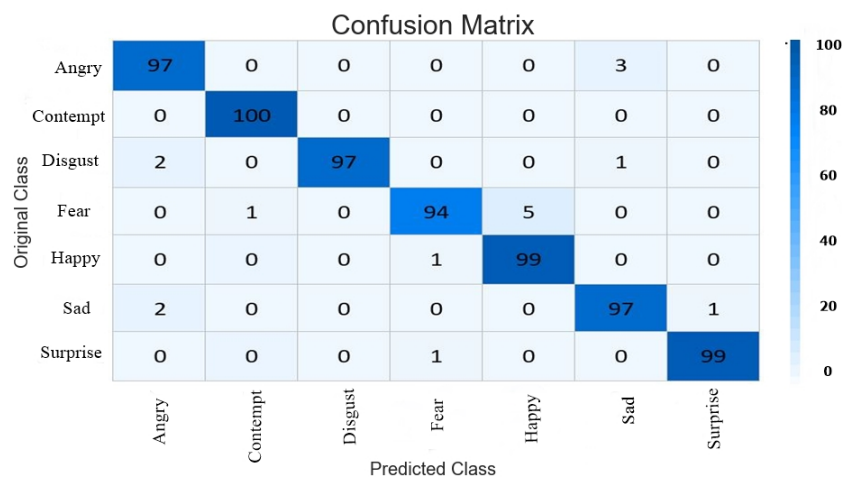


Figure 4. Confusion Matrix on CK+.

Oulu-CASIA is also a famous database, with the help of two cameras, it generates images under varying illumination conditions. VIS contains the images with Oulu-CASIA is also a famous database, with the help of two cameras it generates images under varying illumination conditions. VIS contains the images with strong illumination [44]. Only Oulu-CASIA VIS has used in our model for validation of the results. It contains 480 video sequences against 80 subjects and every subject contains six basic facial expressions. Similar to CK+, three last frames are selected and the 10-fold cross-subject independent split is done. Table 2 shows the average accuracy after 10 runs. Therefore, our proposed method outperformed all other handcrafted methods (HOG 3D, LBP-TOP, and STM-Explet) and CNN-based methods (DTAGN-Joint, PPDN, and FN2EN) [45–47]. The confusion matrix on Oulu-CASIA database is shown in Figure 5.

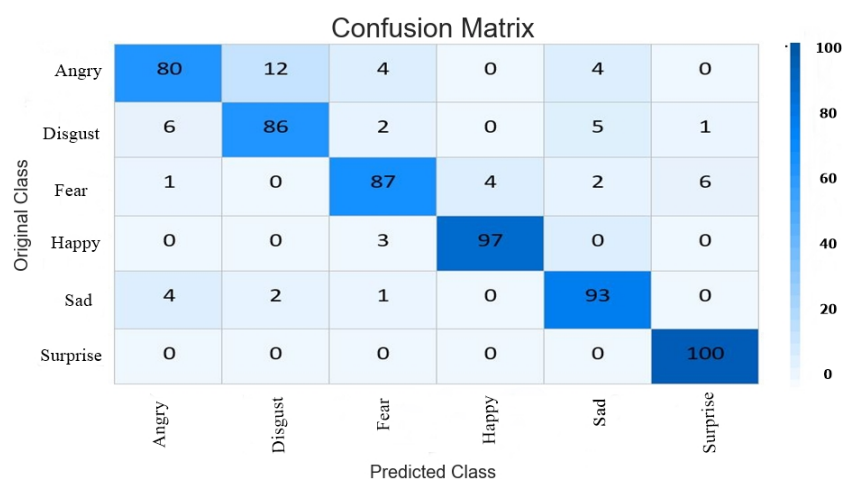


Figure 5. Confusion matrix on Oulu-CASIA.

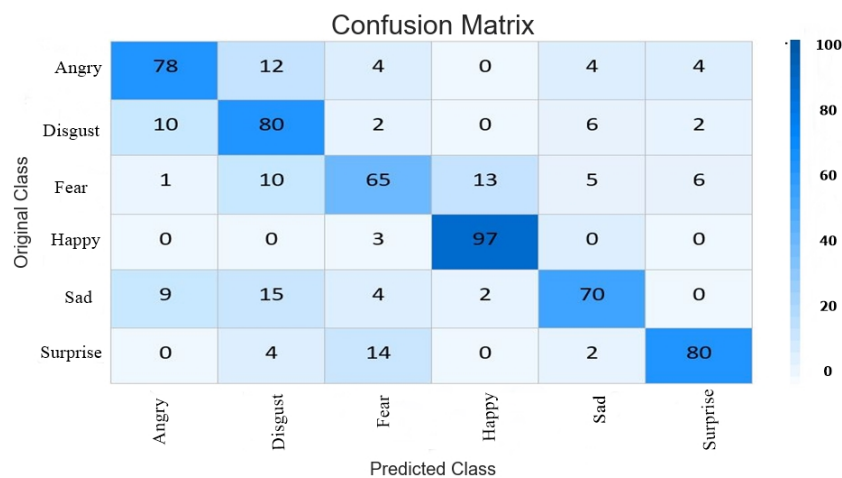
Table 1. Average accuracy rates of seven facial expressions classification on CK+ database.

Method	Setting	Accuracy
LBP-TOP [38]	sequence-based	88.99
HOG 3D [39]	sequence-based	91.44
3DCNN [40]	sequence-based	85.9
STM-Explet [41]	sequence-based	94.19
DTAGN [42]	sequence-based	97.27
CNN [43]	image-based	89.50
IACNN(Baseline) [6]	image-based	95.37
DLP-DeRL(Ours)	image-based	97.57

Table 2. Average accuracy rates of seven facial expressions classification on Oulu-CASIA database.

Method	Setting	Accuracy
LBP-TOP [38]	sequence-based	68.13
HOG 3D [39]	sequence-based	70.63
STM-Explet [41]	sequence-based	74.59
Atlases [45]	sequence-based	75.52
DTAGN-Joint [44]	sequence-based	81.46
FN2EN [46]	image-based	87.71
PPDN [47]	image-based	84.59
CNN [43]	image-based	72.92
IACNN(Baseline) [6]	image-based	82.00
DLP-DeRL(Ours)	image-based	90.5

The MMI database contains 236 image sequences. We took 208 frontal view sequences. Each sequence contains six basic facial expressions. MMI is different then CK+ as sequences are onset-apex-offset labeled, i.e., every sequence starts with neutral expression, reaches the peak in the middle, and then ends over neutral expression. MMI database is having more tough conditions, subjects perform same expressions with varying inter-personal variations, i.e., wearing glasses, hats etc for experimentation. Three frames are taken in middle of every sequence. Just like CK+, after 10 runs cross validation is done and the average accuracy is achieved. Table 3 displays the dominance of our proposed method over other methods. The confusion matrix on MMI database is shown in Figure 6.

**Figure 6.** Confusion matrix on MMI database.

The BU-3DFE database consists of a bulk of texture and static 3d face model images from subjects with different races and ages, having four intensity levels and six basic facial expressions. This database is mainly used for multiview 3D facial expression analysis just like Multi-PIE. A subject independent 10-fold split is done. The average accuracy is shown in Table 4 after 10 runs. It has outperformed all other existing methods [48–52]. The confusion matrix on BU-3DFE database is shown in Figure 7. At the end, per class precision and f1 scores are shown in Figures 8 and 9, respectively.

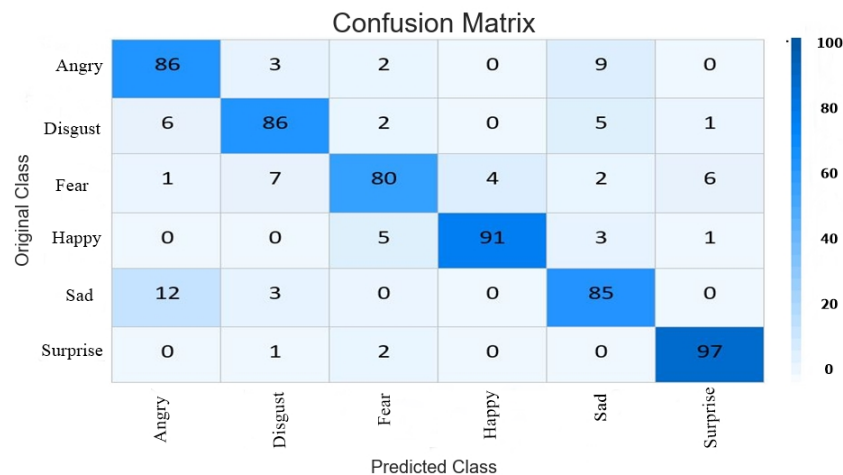


Figure 7. Confusion Matrix on BU-3DFE database.

Table 3. Average accuracy rates of seven facial expressions classification on MMI database.

Method	Setting	Accuracy
LBP-TOP [38]	sequence-based	59.51
HOG 3D [39]	sequence-based	60.89
STM-Explet [41]	sequence-based	75.12
DTAGN-Joint [44]	sequence-based	70.24
CNN [43]	image-based	57.00
IACNN(Baseline) [6]	image-based	71.55
DLP-DeRL(Ours)	image-based	78.33

Table 4. Average accuracy rates of seven facial expressions classification on BU-3DFE database.

Method	Setting	Accuracy
Wang et al. [48]	3D	61.79
Berreti et al. [49]	3D	77.54
Yang et al. [50]	3D	84.80
Lo et al. [51]	2D Image + 3D	86.32
Lopes [52]	image-based	72.89
CNN [43]	image-based	73.2
IACNN(Baseline) [6]	image-based	83.15
DLP-DeRL(Ours)	image-based	87.5

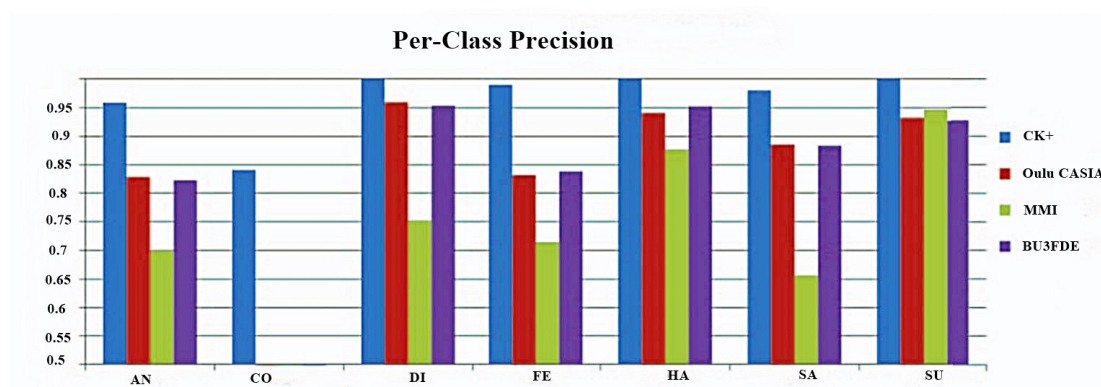


Figure 8. Precision of every class for the aforementioned databases.

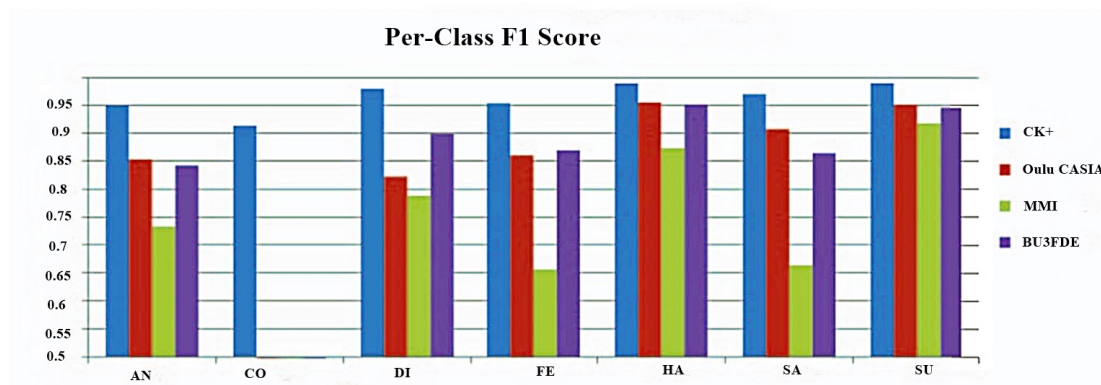


Figure 9. F1 score of every class for the aforementioned databases.

Threats to Validation

Considering the aforementioned expressions, a present threat in the environment or signal offensive aggression, e.g., a perceived arousal from face turned away, would be less compared to that with the face forward view. Furthermore, we had considered a simple paradigm by evaluating a face in the crowd by considering three scenarios, i.e, first, having all angry, sad, or neutral faces; second, by having discrepant face as an angry in the crowd of happy and neutral faces; and, third, having discrepant face as a happy in the crowd of angry and neutral faces. Responses were faster and less error prone for angry expressions in the crowd compared to happy expressions. Attentional processing by angry faces has been disrupted to a bigger extent than compared with happy faces. An angry face should be more easily detected than a happy face. Fear expression with the gasping face and wide eyed is interpreted as an expression of threat. Likewise, for disgust the noise wrinkle stereotype is assumed to be evolved for limiting the exposure to efficient threat detection. While by removing eye brows in an image the sadness expression was indicated instead of angry, it was because of paying the price of having just one different feature (mouth) between the expressions. However, a question may arise that why happy expression is not detected instead of sad. It is because sad and angry share almost similar facial expressions. Adaptive value in detecting angry stimuli is greater than that of sad stimuli. Therefore, the ambiguous expressions must be detected as angry at first.

5. Conclusions

In the past decade, much study has been done over handling occlusion, still, most of the facial expression recognition systems able to tackling it are at a very initial stage. Moreover, some very good results have been achieved with the use of hybrid architectures, although solving microexpressions is still a problem, as they are subtle facial movements that are more spontaneous and occur involuntarily. In this paper, our work mainly focused on real-world affected facial images. Initially, the data is preprocessed for fine tuning the facial image and then it is forwarded to remove intraclass variations

for further processing. Then we introduced a novel approach and an optimized algorithm for facial expression recognition. This approach is used for reliable crowdsourcing and deep locality preserving based on de expression residue learning. Our research method is implemented on both spontaneous and posed facial expression datasets, and it outperformed the state-of-the-art image and also outperformed sequence-based methods. The results have also been validated over cross databases. In realistic applications human expressive behaviors involve different encoding perspectives, where facial expression is just a single modality. Although promising results can be obtained by incorporating other models with facial expression recognition with visible face images into a high-level framework. Participants in Audio Video Emotion Challenges (AVEC) and EmotiW challenges now know the audio model is the second most vital element, and for multimodal affect recognition, they implemented different fusion techniques. As a multi-discipline field, facial expression recognition will provide more promising results in the closed related fields like psychology, neuroscience, and cognitive science. Our future work will aim to do a fusion of multiple modalities, such as physiological data, depth information of 3D face models, and infrared images.

Author Contributions: Data curation, M.S.A.; Formal analysis, U.S.; Investigation, U.A.; Project administration, A.U.; Supervision, J.W.; Lead role, Z.F.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baltrusaitis, T.; Zadeh, A.; Yao, C.L.; Morency, L.P. OpenFace 2.0: Facial Behavior Analysis Toolkit. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Lille, France, 14–18 May 2019; pp. 59–66.
2. Ullah, A.; Wang, J.; Anwar, M.S.; Ahmad, U.; Wang, J.; Saeed, U. 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), Beijing, China, 13–15 October 2018; pp. 1–6.
3. Zhang, L.; Verma, B.; Tjondronegoro, D.; Chandran, V. Facial Expression Analysis under Partial Occlusion: A Survey. *ACM Comput. Surv.* **2018**, *51*, 25. [[CrossRef](#)]
4. Polli, E.; Bersani, F.S.; De, R.C.; Liberati, D.; Valeriani, G.; Weisz, F.; Colletti, C.; Anastasia, A.; Bersani, G. Facial Action Coding System (FACS): An instrument for the objective evaluation of facial expression and its potential applications to the study of schizophrenia. *Riv. Psichiatr.* **2012**, *47*, 126. [[PubMed](#)]
5. Vail, A.K.; Baltrusaitis, T.; Pennant, L.; Liebson, E.; Baker, J.; Morency, L.P. Visual attention in schizophrenia: Eye contact and gaze aversion during clinical interactions. In Proceedings of the International Conference on Affective Computing and Intelligent Interaction, San Antonio, TX, USA, 23–26 October 2017.
6. Meng, Z.; Liu, P.; Cai, J.; Han, S.; Tong, Y. 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), Washington, DC, USA, 1–3 June 2011; pp. 558–565.
7. Qi, W.; Shen, X.; Fu, X. The Machine Knows What You Are Hiding: An Automatic Micro-expression Recognition System. In Proceedings of the International Conference on Affective Computing and Intelligent Interaction, Memphis, TN, USA, 9–12 October 2011.
8. Felzenszwalb, P.F.; Girshick, R.B.; Mcallester, D. Cascade object detection with deformable part models. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010.
9. Yu, Z.; Zhang, C. Image based static facial expression recognition with multiple deep network learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 435–442.
10. Kotsia, I.; Pitas, I.; Buciu, I. An analysis of facial expression recognition under partial facial image occlusion. *Image Vis. Comput.* **2008**, *26*, 1052–1067. [[CrossRef](#)]

11. Wright, J.; Yang, A.Y.; Ganesh, A.; Sastry, S.S.; Ma, Y. Robust Face Recognition via Sparse Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 210–227. [[CrossRef](#)] [[PubMed](#)]
12. Osherov, E.; Lindenbaum, M. Increasing CNN Robustness to Occlusions by Reducing Filter Support. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
13. Rui, M.; Hadid, A.; Dugelay, J. Improving the recognition of faces occluded by facial accessories. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, Santa Barbara, CA, USA, 21–25 March 2011.
14. Huang, X.; Zhao, W.; Pietikäinen, G.; Zheng, M. Towards a dynamic expression recognition system under facial occlusion. *Pattern Recognit. Lett.* **2012**, *33*, 2181–2191. [[CrossRef](#)]
15. Lin, J.-C.; Wu, C.-H.; Wei, W.-L. Facial action unit prediction under partial occlusion based on Error Weighted Cross-Correlation Model. In Proceedings of the IEEE International Conference on Acoustics, Vancouver, BC, Canada, 26–31 May 2013.
16. Zhang, L.; Tjondronegoro, D.; Chandran, V. Random Gabor based templates for facial expression recognition in images with facial occlusion. *Neurocomputing* **2014**, *145*, 451–464. [[CrossRef](#)]
17. Dapogny, A.; Bailly, K.; Dubuisson, S. Confidence-Weighted Local Expression Predictions for Occlusion Handling in Expression Recognition and Action Unit Detection. *Int. J. Comput. Vis.* **2017**, *3*, 1–17. [[CrossRef](#)]
18. Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; Zheng, N. Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
19. Afifi, M.; Abdelhamed, A. Deep Gender Classification based on AdaBoost-based Fusion of Isolated Facial Features and Foggy Faces. *J. Vis. Commun. Image Represent.* **2017**, *62*, 77–86. [[CrossRef](#)]
20. Sugimoto, A. Facial expression recognition by re-ranking with global and local generic features. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 4118–4123.
21. Itti, L.; Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2001**, *2*, 194–203. [[CrossRef](#)] [[PubMed](#)]
22. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
23. Zheng, H.; Fu, J.; Tao, M.; Luo, J. Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–27 October 2017.
24. Zhao, L.; Xi, L.; Wang, J.; Zhuang, Y. Deeply-Learned Part-Aligned Representations for Person Re-Identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–27 October 2017.
25. Chen, Z.; Zhao, Y.; Huang, S.; Tu, K.; Yi, M. Structured Attentions for Visual Question Answering. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–27 October 2017.
26. Juefei-Xu, F.; Verma, E.; Goel, P.; Cherodan, A.; Savvides, M. DeepGender: Occlusion and Low Resolution Robust Facial Gender Classification via Progressively Trained Convolutional Neural Networks with Attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, Nevada, 26 June–1 July 2016.
27. Norouzi, M.N.; Araabi, E.; Ahmadabadi, B.N. Attention control with reinforcement learning for face recognition under partial occlusion. *Mach. Vis. Appl.* **2011**, *22*, 337–348. [[CrossRef](#)]
28. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Trans. Image Process.* **2018**, *28*, 2439–2450. [[CrossRef](#)]
29. Gogić, I.; Manhart, M.; Pandžić, I.S.; Ahlberg, J. Fast facial expression recognition using local binary features and shallow neural networks. *Vis. Comput.* **2018**, *1*–16. [[CrossRef](#)]
30. Mahmood, A.; Hussain, S.; Iqbal, K.; Elkilani, W.S. Recognition of Facial Expressions under Varying Conditions Using Dual-Feature Fusion. *Math. Probl. Eng.* **2019**, *2019*, 9185481. [[CrossRef](#)]
31. Pan, B.; Wang, S.; Xia, B. Occluded Facial Expression Recognition Enhanced through Privileged Information. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 566–573.

32. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2003.
33. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J. 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 Francisco, CA, USA, 13–18 June 2010.
34. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; Pietikinen, M. Facial expression recognition from near-infrared videos? *Image Vis. Comput.* **2011**, *29*, 607–619. [[CrossRef](#)]
35. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005.
36. Yin, L.; Wei, X.; Sun, Y.; Wang, J.; Rosato, M.J. A 3D Facial Expression Database For Facial Behavior Research. In Proceedings of the International Conference on Automatic Face and Gesture Recognition, Southampton, UK, 10–12 April 2006.
37. Ullah, A.; Wang, J.; Anwar, M.S.; Ahmad, U.; Saeed, U.; Wang, J. 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), Beijing, China, 12–16 August 2018; pp. 418–423.
38. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 915–928. [[CrossRef](#)] [[PubMed](#)]
39. Klaser, A.; Marszałek, M.; Schmid, C. A Spatio-Temporal Descriptor Based on 3D-Gradients. In Proceedings of the BMVC'08, Leeds, UK, 13–15 September 2008.
40. Liu, M.; Li, S.; Shan, S.; Wang, R.; Chen, X. Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014.
41. Liu, M.; Shan, S.; Wang, R.; Chen, X. Learning Expressionlets on Spatio-temporal Manifold for Dynamic Facial Expression Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, Ohio, 24–27 June 2014.
42. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
43. Gauthier, J. Conditional generative adversarial nets for convolutional face generation. In *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter Semester*; Stanford University: Stanford, CA, USA, 2014; Volume 2014, p. 2.
44. Yang, H.; Ciftci, U.; Yin, L. Facial Expression Recognition by De-Expression Residue Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, 19–21 June 2018.
45. Guo, Y.; Zhao, G.; Pietikäinen, M. Dynamic facial expression recognition using longitudinal facial expression atlases. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 631–644.
46. Ding, H.; Zhou, S.K.; Chellappa, R. *FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition*. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May –3 June 2017.
47. Zhao, X.; Liang, X.; Liu, L.; Teng, L.; Han, Y.; Vasconcelos, N.; Yan, S. Peak-Piloted Deep Network for Facial Expression Recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.
48. Wang, J.; Yin, L.; Wei, X.; Yi, S. 3D Facial Expression Recognition Based on Primitive Surface Feature Distribution. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006.
49. Berretti, S.; Bimbo, A.D.; Pala, P.; Amor, B.B.; Daoudi, M. A Set of Selected SIFT Features for 3D Facial Expression Recognition. In Proceedings of the International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010.

50. Yang, X.; Di, H.; Wang, Y.; Chen, L. Automatic 3D facial expression recognition using geometric scattering representation. In Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Ljubljana, Slovenia, 4–8 May 2015.
51. Li, H.; Ding, H.; Di, H.; Wang, Y.; Zhao, X.; Morvan, J.M.; Chen, L. An efficient multimodal 2D + 3D feature-based approach to automatic facial expression recognition. *Comput. Vis. Image Underst.* **2015**, *140*, 83–92. [[CrossRef](#)]
52. Lopes, A.T.; Aguiar, E.D.; Souza, A.F.D.; Oliveira-Santos, T. Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order. *Pattern Recognit.* **2016**, *61*, 610–628. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).