Annals of Emerging Technologies in Computing (AETiC)
Vol. 7, No. 2, 2023

Research Article

Enhancing Feature Extraction Technique Through Spatial Deep Learning Model for Facial Emotion Detection

Nizamuddin Khan[1,*], Ajay Vikram Singh[1] and Rajeev Agrawal[2]

[1]Amity Institute of Information Technology, Amity University, Noida UP, India
nukhan0308@gmail.com; avsingh1@amity.edu
[2]Lloyd Institute of Engineering and Technology, G. Noida, UP, India
rajkecd@gmail.com
*Correspondence: nukhan0308@gmail.com

Received: 14th September 2022; Accepted: 15th March 2023; Published: 1st April 2023

Abstract: Automatic facial expression analysis is a fascinating and difficult subject that has implications in a wide range of fields, including human–computer interaction and data-driven approaches. Based on face traits, a variety of techniques are employed to identify emotions. This article examines various recent explorations into automatic data-driven approaches and handcrafted approaches for recognising face emotions. These approaches offer computationally complex solutions that provide good accuracy when training and testing are conducted on the same datasets, but they perform less well on the most difficult realistic dataset, FER-2013. The article's goal is to present a robust model with lower computational complexity that can predict emotion classes more accurately than current methods and aid society in finding a realistic, all-encompassing solution for the facial expression system. A crucial step in good facial expression identification is extracting appropriate features from the face images. In this paper, we examine how well-known deep learning techniques perform when it comes to facial expression recognition and propose a convolutional neural network-based enhanced version of a spatial deep learning model for the most relevant feature extraction with less computational complexity. That gives a significant improvement on the most challenging dataset, FER-2013, which has the problems of occlusions, scale, and illumination variations, resulting in the best feature extraction and classification and maximizing the accuracy, i.e., 74.92%. It also maximizes the correct prediction of emotions at 99.47%, and 98.5% for a large number of samples on the CK+ and FERG datasets, respectively. It is capable of focusing on the major features of the face and achieving greater accuracy over previous fashions.

Keywords: ***Convolutional neural network; Facial expression recognition; Spatial deep learning; Spatial transform network***

1. Introduction

Facial expression recognition is a hot topic in computer vision, with a wide range of applications including human behaviour analysis, mental disorder identification, and human-computer interaction, to name a few. Most recent research [1], [2], and [3-7] has concentrated on developing deep ANNs to achieve cutting-edge outcomes. Even though handcrafted feature-based artificial neural network models [8] and [9] provide results that are less accurate than deep learning networks, they have attracted less attention. Various methods are employed to identify emotions based on face traits. This manuscript examines many recent investigations into the automatic data-driven technique [3-7] and the handcrafted approach [1-2] to facial emotion recognition. In the most difficult real-world dataset, FER-2013, these approaches have computationally complex solutions that give good accuracy while training and testing on the same datasets. The objective of the research is to provide a most efficient system with less computing complexity to predict

Nizamuddin Khan, Ajay Singh and Rajeev Agrawal, "Enhancing feature extraction technique through Spatial Deep Learning model for Facial Emotion Detection", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 9-22, Vol. 7, No. 2, 1st April 2023, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2023.02.002, Available: http://aetic.theiaer.org/archive/v7/v7n2/p2.html.

emotion classes more accurately than current methodologies and to aid society in a practical, all-encompassing facial expression system solution.

The earliest attempts at recognising face expressions mainly depended on hand-crafted characteristics [10]. Following the success of the AlexNet [11] deep neural network in the ImageNet Large-Scale Visual Recognition Challenge [12], deep learning has become widely used in the field of computer vision. Some of the early papers to propose deep learning algorithms for facial emotion identification were presented at the 2013 Facial Expression Recognition (FER) Challenge, according to [13]. Surprisingly, a deep convolutional neural network achieved the highest score in the 2013 FER Challenge, while the best handcrafted model came in fourth [9].

Deep learning has been used in the bulk of recent studies on facial emotion identification, with a few exceptions [8], [10]. For improved performance, several recent articles have suggested training a group of CNNs, while others have mixed handmade features like SIFT [14] or HOG [1] with deep features. In earlier studies, a classifier SVM or neural network was used to detect the emotions after extracting features from images, which is one of two crucial stages for emotion recognition. The histograms of directed gradients [1], local binary patterns, haar features, and Gabor wavelets are some well-known hand-crafted functions used for face feature detection. These techniques were effective when applied to smaller datasets, but as larger datasets (with higher magnitude variance) were accessible, their drawbacks became apparent. Use images in which only a piece of the face is visible or the face area is covered by eyeglasses or a hand to better grasp a range of challenging scenarios.

This article demonstrates that by combining automatic features learned by the base model of CNN with spatial variant pixels computed by the spatial deep learning model, we can outperform earlier state-of-the-art systems, especially when local learning is used during the training and validation phases. We test several methods for obtaining automatic characteristics, including hybrid CNN architectures and baseline architectures. Some of them have already been trained on other computer vision tasks, like face recognition and object recognition [12]. These CNN models are also fine-tuned by us for FER using an extended L2 regularisation training procedure that outperforms Dense-Sparse Dense.

We employ both automated and STN features separately and in combination and notice the major difference in the accuracy of FER, whereas STN stands for Spatial Transform Network, which removes spatial invariance from images by using a learnable affine transformation followed by interpolation. It is made up of three parts: a local network, a grid generator, and a sampler. We fine-tuned the independent CNN models by using softmax and cross entropy for classification. Using the one-vs.-all strategy, CNN for models that are combined, only the baseline model and the spatial transform model are investigated. On the FER-2013 Challenge dataset [13], the FERG dataset [15], and the CK+ dataset [16], we compare our proposed models with recent and relevant state-of-the-art approaches [5-7, 17]. Deep learning has recently been used to extract and train numerous features for a good FER system, notably convolutional neural networks (CNNs) [3-7]. However, many of the signals for facial emotions come from a few parts of the face, such as the eyes and lips, while other parts of the face, such as the hairs and ears, play minor roles in the output. This implies that the deep learning model should ideally focus solely on the essential facial regions and ignore the rest of the face.

This research, present a spatial deep learning model for FER that incorporates the aforementioned observation and focuses on the most significant element of the face using an attention mechanism [18]. We demonstrate that by employing STN, even a simple network with a few layers can achieve a very high accuracy rate. The following contributions are highlighted in this paper:

1) We present a STN-based approach that is more accurate than traditional CNN-based networks and focuses on areas of the face with rich feature information. 2) In addition, we employ the visualisation technique proposed in [19] to highlight the most relevant regions of the face image, i.e., the areas of the image that have the greatest influence on the classifier's output (see Figure 1).

Deep learning methods such as multilayer perceptron NN and SVM have been investigated as a family of solutions to improve the durability and efficiency of machine learning classification algorithms. Because humans work in a variety of contexts, human behavior analysis must be robust in order to provide the necessary versatility and durability when dealing with new types of data. CNN learns by using the backpropagation method to limit the change in weights based on the training target. Reaction learning in the human brain is analogous to using a backpropagation method to optimize a fitness function. CNN has

a deep, multi-layered, hierarchical structure. It can extract high-, medium-, and low-level data. Lower and mid-level characteristics are combined to form high-level characteristics. CNN's hierarchical feature extraction capabilities mimic the human brain's central nervous system, which is a sophisticated and layered learning system that dynamically absorbs features from input data.

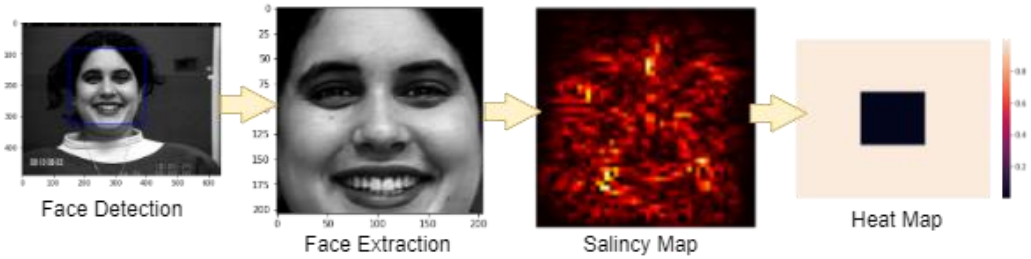

Figure 1. Using our model, we identified salient regions for various facial expressions. The image is part of the larger Cohn-Kanade dataset.

2. Related Works

Handcrafted characteristics are used in traditional approaches in the literature; however, CNN-based models have significantly improved performance in a variety of soft computing tasks. Object detection, segmentation, face recognition, and facial emotion recognition are examples [9], [11], [17], and [20-24]. In this section, we will look at some CNN-based FER models that have performed well in recent years. Comparison of various state-of-the-art techniques for FER is given in Table 1. Subsections 2.1 to 2.7 show some architectural details of related network models.

2.1. FER Using Attentional CNN

Shervin and Mehdi proposed the attentional convolutional model [17], which has less than ten layers, with four convolutional layers using max-pooling and activation functions (ReLU) for feature extraction and two convolutional layers each using max-pooling, ReLU, and two fully connected layers for localization. They claim to have outperformed the competition (70.02 percent for FER-2013). Attentional mechanisms are also used in articles [7] and [18].

2.2. FER Using Deep Networks

Since Mollahosseini and Ali proposed a deep model [20] based on the Inception layer for Deep Neural Network applications, it appears only reasonable to use cutting-edge object recognition algorithms to solve the FER problem. Lower convolutions are used locally, while higher convolutions are used to approach global features. The origin layer enables better local feature detection as well as theoretical advantages from the network's sparsity and hence relative depth. The majority of emotions can be distinguished by examining details such as the eyes and lips. Using the Inception layer structure and Lin et al.'s network-in-network theory, we can predict large increases in local feature performance, which logically equates to improved FER outputs.

Then, in the spirit of "Inception," it adds two modules composed of 1x1, 3x3, and 5x5 convolution layers with ReLU that operate in parallel using network design methodologies. This layer's output is then concatenated, resulting in two completely connected layer classifications with ReLU activation functions.

2.3. Improve the Bag of Visual Words model for FER Using Local Learning

Ionescu and Tudor [9] introduce a new computer vision system for recognising human face expressions in low-resolution photos. It extracts dense SIFT descriptors from the entire image or from a spatial pyramid that divides it into finer and finer sub regions. Multiple kernel learning includes creating linear concatenating kernels for a range of weighted sums of localised presence vectors. According to empirical findings, combining presence vectors, local learning, and spatial information improves recognition accuracy by more than 5%. Finally, in the FER Challenge, his model came in fourth place with a final test set accuracy of 67.48 percent.

2.4. Handcrafted Deep Features for FER

This model was proposed by Georgescu and Mariana [21]. Dense Sparse Dense is an automatic feature experiment that employs several CNN models, pre-trained architectures, and training strategies. The test image for which the SVM was trained is used to predict the class name. Despite the fact that it has been utilised in the past with handcrafted features, regional learning has never been employed in conjunction with our knowledge's deep features. Using the data set from the FER-2013 Contest, the author demonstrates that his approach (Figure-2) produces cutting-edge results. The maximum accuracy in this dataset was 66.31 percent.

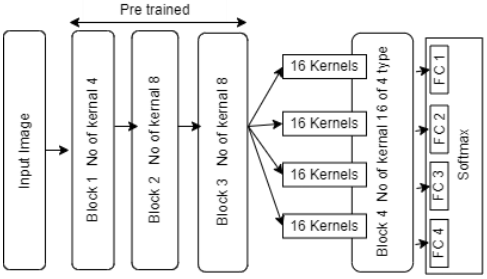

Figure 2. Neural Network Model composed of 3 pre trained CNN blocks and task oriented columns

2.5. GoogLeNet

It was the embodiment of the "Inception" network proposed in the ILSVRC 2014 challenge [22]. When only parameter-containing layers are counted, the network has 22 layers, whereas the total number of layers used is close to 100. Parts of the network run in parallel. These elements are referred to as "inception modules." GoogLeNet, for example, has 12 times fewer parameters than the AlexNet design, which will be deployed next. Sparse matrices, which are grouped into tightly packed submatrices, produce cutting-edge results. To train the network, the DistBelief system [25] was used, with an asynchronous SGD descent with 0.9 momentum and a constant rate of learning with a 4% rate of drop per eight epochs. The first top accuracy rate is used to evaluate the classifier's performance, and the fifth top error rate is calculated when comparing the ground truth to the first five predicted classes.

2.6. FER Domain Adaption using Generative Adversarial Networks is Unsupervised

It is an unsupervised method. We begin by teaching a CNN to recognise facial expressions using the raw images. We want to improve CNN's cross-dataset performance after it has been trained with the source database without using ground truth label information from the target database [23]. In order to do so, we'll have to work around the target dataset's small sample size. GAN provides a solution. We fine tune with the original dataset because the results are better than fine-tuning with only the samples generated. Figure 3 shows.

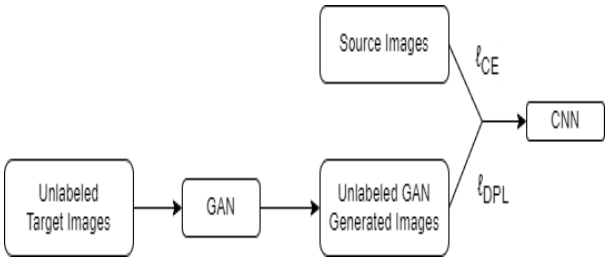

Figure 3. unlabelled GAN images in this training framework

ℓ_{CE} stands for cross-entropy loss, while ℓ_{DPL} stands for distributed pseudo label loss, which is employed for the unlabelled GAN images in this training network.

2.7. Alex Net

It was designed by Krizhevsky and Sutskever and first entered the 2012 ILSVRC competition [11]. Eight completely connected nodes make up the network. To build a 1000 class label distribution, the output of the final fully connected layer is passed into the 1000-way softmax function. This means that the first layer has

55x55x96 = 290400 neurons, each with 11x11x3 = 363 weights and 1 bias, for a total of 105705600 (=290400x364) parameters. The second CNN layer contains 256 kernels with dimensions of 5x5x48, the third layer contains 384 kernels with dimensions of 3x3x256, and so on. SGD was used to train the network, with a batch size of 128, 0.9 momentum, and 0.0005 weight decay. A Gaussian distribution with a standard deviation of 0.01 and a mean of zero was used to calculate the weights in each layer. Finally, after the validation error rate stopped improving, all layers were given the same learning rate, which was then divided by ten. Because of the computational cost of the training process, the architecture is split into two channels, one for each of the two GPUs utilised in the training.

Summary of literature review with performance analysis is as follows (Table 1):

Table 1 Performance Comparison of various recent state-of-the-art techniques for FER

Techniques	Performance Analysis				Reference
	Database	Performance on FER-2013 dataset	Gap	Strength	
Attentional CNN	CK+, FERG, FER-2013, JAFFE	70.02%	They don't test FER-2013 using the regular size of data.	2 fully connected layers and fewer than 10 convolutional layers	[17]
Deep Neural Network	FERG, JAFEE, FER-2013, CK+	66.40%	With higher convolutions, global - characteristics are approached.	Locally, lower convolutions are employed. Better local feature detection is made possible by the deep NN inception layer.	[20]
Bag of visual Words model using local learning	FER-2013	67.48%	High computational complexity is involved in multiple kernel learning.	Empirical results show that the accuracy of recognition is increased by 5% when presence vector, local learning, and spatial information are combined. capable of identifying face emotion in pictures with low resolution	[9]
Handcrafted Deep Features	FER-2013	66.31%	Uneven heavy network and hyperparameters	Sparse Dense An artificial feature experiment called Dense makes use of a number of CNN models, pre-trained architectures, and training techniques. SVM is applied to categorization. FER-2013 dataset cutting-edge results production	[21]
GoogleLeNet	CK+, FER-2013, FERG, JAFFE	65.20%	utilising huge networks and high training parameters	Sparse matrices, which are arranged into closely packed submatrices and have 12 times less parameters than AlexNet design, produce state-of-the-art outcomes.	[22]
Unsupervised GAN (Generative Adversarial Network)	JAFFE, FERG, CK+, FER-2013	65.30%	More complexity in computation	Boost CNN's performance across datasets	[23]

3. Proposed Model based on Spatial Deep Learning Method

We present a very effective spatial deep learning model for feature extraction to categorise potential emotions in face images for the FERG, CK+, and FER-2013 databases, which outperforms the existing state-of-the-art method [5-7], [17]. Increasing the size of a CNN, assisting in the direction of the sharpest drop in error, or improving spectral normalisation are all common ways to improve a deep neural network, particularly for challenges with a large number of classes. Our straightforward model, which is constructed from the ground up with fewer hyper parameters, provided good results with little computing effort. This model is only sensitive to the most important part of the face and ignores the rest of the face. As shown in [19], we can visualise it.

It is obvious from a facial image that not all portions of the face are required for interpreting various emotions. In many circumstances, all that is required is a concentrated effort in a few critical areas to understand the underlying feelings. Based on this discovery, we added two Spatial Transformer Networks

[26] to our network to concentrate on the most important features of the face. Figure-4 depicts the proposed model for spatial deep learning architecture.

Our model more efficiently pulls features from the feature set, with this set focusing on the most significant elements of the facial image we can imagine, as suggested by Zeiler and Fergus [19], which play a significant role in communicating a person's mood and producing promising results.

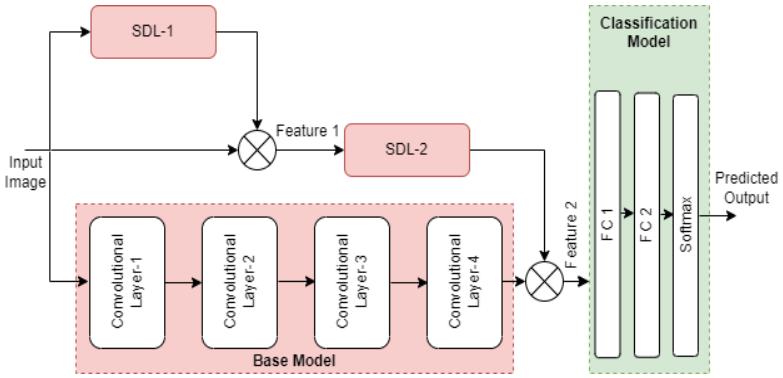

Figure 4. Proposed Spatial deep learning model based on STN

There are two feature extraction parts, each of which is made up of one of the rectified linear unit's activation functions. This is followed by a convolutional layer and an average pooling layer, all of which are part of a single Spatial Deep Learning network (SDL). The first SDL feature extraction model is very effective; it takes the original image and transforms important pixels with the help of a learnable affine transformation from the different parts of the face. This is based on spatial transform networks (STN)

Using the ReLU activation function [26-27], we can avoid the vanishing gradient problem. These two feature extraction parts are linked and joined in parallel with the baseline model (Figure 4), which is made up of four convolutional layers (max-pooling and ReLU). Following that, these two feature-extracting parts were linked together to form two fully connected layers for classification.

3.1. Spatial Deep Learning Module (SDL)

This section shows the working and architectural details of the SDL module. The spatial transformer (location network) is the foundation for the SDL model, which removes spatial invariance from images by using a learnable affine transformation followed by interpolation (Figure 5). It is made up of two fully connected layers and three convolutional parts (each convolutional layer followed by maxpooling and ReLU).

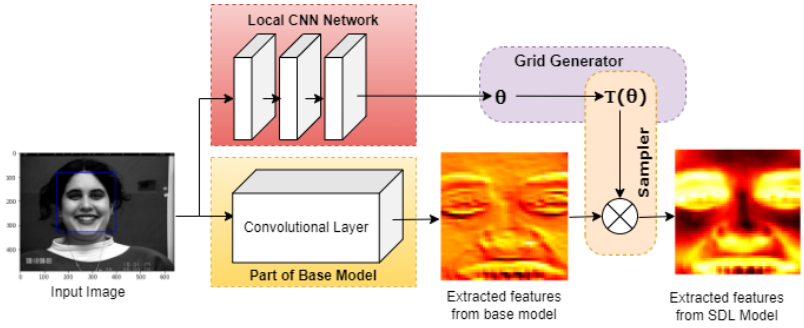

Figure-5 Spatial Deep Learning module

After backtracking the transformation parameters, the input is changed to the sample grid T(θ), resulting in the output of the same data using the identity matrix as in eq. (1).

Our proposed model feature extraction phases (Figure-6) start with an image and extract the relevant features for emotion classification; these features are then sent to the second SDL module.

The second component, SDL-2 (Figure-7), takes input from the first SDL-1 module and extracts the most prominent features from the input feature map using a learnable affine transformation, which is very useful for emotion classification.

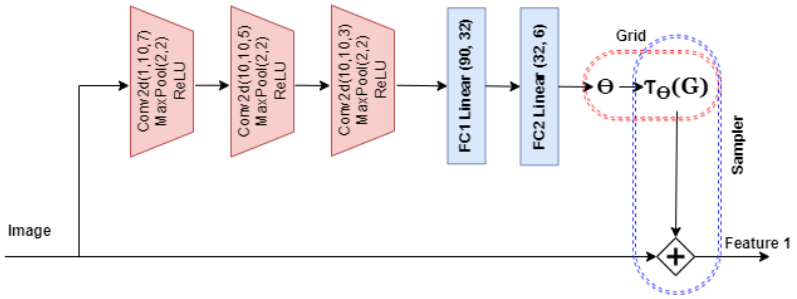

Figure 6. Relevant feature extraction part from the proposed model (SDL-1)

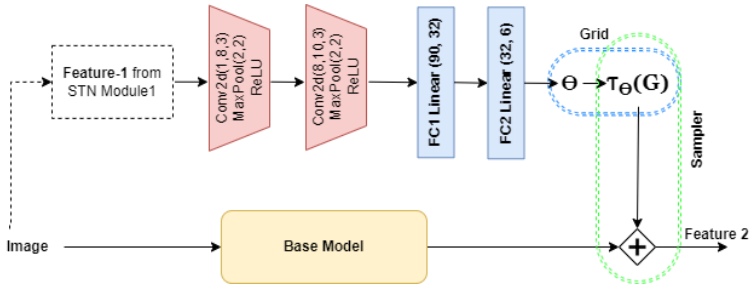

Figure 7. Prominent feature extraction part from the proposed model (SDL-2)

SDL module removes spatial invariance from images by using a learnable affine transformation followed by interpolation. The SDL block can be used in a CNN and can function almost entirely on its own, with only a few transformation stages.

Make an affine transformation matrix θ to represent the linear transformation first.

$$\theta = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \end{bmatrix} \quad (1)$$

After that, rather than applying the transformation directly to the initial image (U), we construct a sampling mesh grid of the same size (U), A mesh grid is a set of (x_t, y_t) values that span the entire image space. There is no pixel value associated with it.

Apply the linear transformation matrix in (1) to the meshgrid we created earlier to generate a new set of sampling points as in eq. (2).

$$\begin{bmatrix} X_i^s \\ Y_i^s \end{bmatrix} = \theta \begin{bmatrix} X_i^t \\ Y_i^t \\ 1 \end{bmatrix} \quad (2)$$

Finally, create the sampled output (V) using the initial feature map in (2), the modified mesh grid, and the differentiable interpolation function (e.g. bilinear).

$$V_i^C = \sum_n^H \sum_m^W U_{nm}^C \max(0, 1 - |x_i^s - m|) \, max(0, 1 - |y_i^s - n|) \quad (3)$$

Finally, we'll create a localization network whose sole purpose is to learn and then provide the right for a specific image based on the loss we back-propagated through the sampler.

The spatial deep learning transformation module aims to concentrate on the image's most important features by estimating samples in the region of interest. Several transformations can be used with an affine transformation to change the direction of the input into the output [26].

The proposed enhanced feature extraction model is then trained by minimizing the error function using the AdamW optimization algorithm. This algorithm is a stochastic gradient descent extension that produces good results for our proposed deep learning model. The error function (given in (4) and (5)) in this paper is simply the sum of cross entropy and L2-regularization (penalty) in the final two completely linked layers.

$$E_{Total} = E_{Cross\_entropy} + L_2\text{-regularization} \quad (4)$$

$$L_2\text{-regularization} = 0.5(\lambda_1 \sum W_1^2 + \lambda_2 \sum W_2^2) \quad (5)$$

To select the matching value validation set with the best validation performance, the regularization parameters $\lambda 1$ and $\lambda 2$ are modified based on the model's validation set performance [28]. We trained separate models for the FERG, CK+ and FER2013 databases used in this study. There was other research into a network with a similar layout but more than fifty levels, However, the precision did not significantly

improve. As a result, we chose a network with fewer layers, which results in faster, more predictable times and is better suited to real-world applications.

We used the AdamW optimizer in the proposed model, and the weight decay (in eq. (6)) changed according to the equation 5.

$$\text{Weight decay} = \lambda_1 \sum W_1 + \lambda_2 \sum W_2 \quad (6)$$

Data augmentation was used to increase the number of images used to train the model. It used subtle changes such as flipping, minor rotation, and small distortion. As a result, the model will be well-trained for use with a large database. In datasets with large class imbalances, we used oversampling to ensure that all classes are in the same sequence, especially in classes with fewer samples.

4. Data Sets

The proposed model is validated using the enlarged Cohn–Kanade [16], FER-2013 [29] and FERG [15] datasets, FER-2013 includes emotions with only half a face or a face hidden by a hand or spectacles. This data set is difficult due to the presence of several prominent facial expression identification datasets, such as the JAFFE [30] and the Facial Expression Research Group Dataset [15]. Before we get into the results, let's take a look at these databases in general (subsection 4.1 to 4.4).

4.1. FER 2013

This database was first presented at the 2013 International Conference on Machine Learning (ICML) representation learning challenges [29]. The majority of the 35,887 images in this collection were taken in natural settings and have a resolution of 48 by 48 pixels. The validation set contained 3589 images and the same number of images as the testing set, there were 28,709 images in the training set. Faces were captured automatically in this dataset, which was created using Google's image search API. On the faces, any of the six primary expressions, as well as neutral, can be seen.

Figure 8. FER-2013

Face blockage (with object), half faces, low-resolution images, and specs are more prevalent in FER than in the other datasets.

4.2. CK+ Database

The extended Cohn–Kanade facial emotion database [16] is a publicly available dataset for recognizing action units and emotions. There are both posed and natural expressions featured, and 593 sequences were collected from the CK+'s 123 individuals. In previous research [17, 28, 31-32], the last frame of these sequences was typically extracted and used for face emotion recognition.

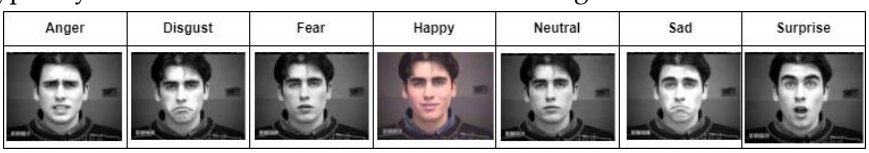

Figure 9 CK+ Dataset

4.3. JAFFE Dataset

This dataset includes 213 images of ten Japanese female models posing with seven different facial expressions. Six emotional descriptors were used by 60 Japanese people to score each image [30].

Figure 10. JAFFE Dataset

4.4. FERG Dataset

It features a cast of characters with annotated facial expressions in a stylised style. In the collection, there are 55,767 annotated face pictures of six stylised creatures. MAYA was used to create the characters. The facial expressions of each character are divided into seven categories [15]. We primarily intended to use this database to evaluate the performance of our algorithm when dealing with cartoon characters.

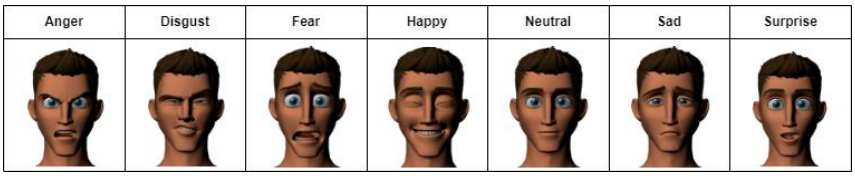

Figure 11. FERG Dataset

5. Result Analysis and Comparison

We trained the proposed SDL model on Google Colab for 300 epochs using a Tesla T4 GPU; the model's starting weights were random Gaussian numbers with a mean value of zero and a standard deviation of 0.05. The learning rate starts at 0.03 and decreases at a rate of 0.9 per 10 epochs up to 300 epochs. L2 regularization values λ_1 and λ_2 start at 0.001 and 0.01 respectively and are tuned to improve model performance and weight decay. In the proposed model, we used the AdamW optimizer for the first 300 epochs with constant weight decay according to equation 5, and data augmentation was used to train and validate the proposed SDL architecture based on a greater number of samples. During training on epoch number 184, the training loss was 0.00484363, the validation loss was 0.00454711, and on the FER-2013 dataset, we achieved a training accuracy of 74.92 percent and a validation accuracy of 76.26 percent.

All of our tests were run on three popular datasets: CK+ [16], FER-2013 [29] and the FERG [15] dataset. On the FERG and CK+ datasets, we achieved a training accuracy of 99.47% and 98.5%, respectively.

As previously stated, the FER-2013 dataset makes it more difficult to recognise facial emotions than the others we used. Aside from the intra-class FER variation, the uneven nature of different emotional classes is another key concern in this dataset. There are many more cases in some classes, such as neutral and joyful.

Table 2. FER-2013

FER-2013	*Angry*	*Disgust*	*Fear*	*Happy*	*Sad*	*Surprise*	*Neutral*	*Total*
Training	3995	436	4096	7214	4830	3171	4965	28707
Validation & Testing	958	111	1024	1774	1247	831	1232	7177

We demonstrate how well the proposed SDL architecture performs on the aforementioned datasets (Table-2). In each case, the proposed SDL architecture was trained on 80% of the images in the dataset, leaving 20% for testing and validation, and the accuracy over the test set (10% of images) was reported. The model was trained using 80 percent (28,707) of the images from the training set. which was then validated with 10% (3588) of the images from the validation set and tested with 10% (3589) of the images to determine the accuracy of the test set. In the test images, we achieved an accuracy rate of 74.92 percent.

Table 3 compares the results of our proposed hybrid model to some of the earlier FER-2013 studies.

Table 3. Result comparison on FER-2013 dataset with proposed model

Method	*Accuracy*
SoC based FER [24]	66%
Yang & Lv [7]	71.44%
Google Net [22]	65.2%
Sho and Cheng [6]	66.17%
Xiaoqing and Wang [23]	65.3%
Georgescu and Mariana [21]	66.31%
Shervin and Mehdi [17]	70.02%
Deep learning STN model [5]	71.36%
The proposed Model	**74.923%**

The CK+ dataset contains video sequences that can be used to train and test facial expression recognizers. For each sequence, the face regularly transforms from a neutral to a peak facial expression. Six fundamental facial emotions (pleased, angry, sad, disgust, surprise, and fear) are included in CK+, as well as one non-basic expression (contempt). Only 327 of the 593 sequences from 123 patients have been tagged with expression labels.

Table-4. CK+ Dataset

CK+	*Angry*	*Disgust*	*Fear*	*Happy*	*Sad*	*Surprise*	*Contempt*	*Total*
Training & Testing	45	59	25	69	28	83	18	327

Table 5 shows a comparison of our model to previous research on the bigger CK+ dataset

Table 5. Result comparison on CK+ dataset with proposed model

Method	*Accuracy*
Yang and Lv [7]	96.97%
Zhang and Zheng [31]	97.2%
Zhao and Liang [28]	97.3%
Eleyan and Akdemir [32]	97.8%
Shervin and Mehdi [17]	98.0%
Deep learning STN model [5]	98.3%
The proposed Model	**98.5%**

In the FERG dataset, we used around 34,000 pictures for training, 14,000 for validation, and 7000 for testing. We chose 1000 images at random for each face expression to test. We were able to obtain a rate of 99.3 percent accuracy. Table 6 shows a comparison of the proposed approach with some of the earlier research on the FERG database.

Table 6. Result comparison on FERG dataset with proposed model

Method	*Accuracy*
Deepali and Colburn [15]	89.02%
Hang & Liu [33]	97.0%
Deep learning STN model [5]	98.3%
Shervin and Mehdi [17]	99.3%
The proposed Model	**99.47%**

5.1. Statistically Analysis

Additionally, we tested our model using conventional values found in equations (7) through (10). Since CNN-based models are not consistent, each time we train, we get a little different accuracy. In order to acquire the 95% confidence interval, we performed the experiment n=15 times for each dataset. Confidence interval can be calculated using equation (7)

$$\text{Confidence Interval} = x \pm Z * (\sigma / \sqrt{n}) \quad (7)$$

For 95% Z value is 1.96, x is average value and σ is standard deviation. The results in Table 7 are as follows

Table 7. 95% confidence level intervals

Datasets	**Sample mean (x)**	**Std. div. (σ)**	**Sample size (n)**	**Confidence level**	**Sample conf. interval**
FERG	99.468	0.17	15	95%	99.47 ± 0.0863
CK+	99.512	0.26	15	95%	99.51 ± 0.1316
FER-2013	74.923	0.32	15	95%	74.92 ± 0.1619

Recall It is the proportion of accurately categorised true positive events.

$$\text{Recall} = \text{ConfMatrix}[x,x] / \sum x^{th} \text{ Row of ConfMatrix} \quad (8)$$

Precision It is the proportion of appropriately classified positive cases.

$$\text{Precision} = \text{ConfMatrix}[x,x] / \sum x^{th} \text{ Column of ConfMatrix} \quad (9)$$

Where ConfMatrix[x,x] is diagonal element of x^{th} column

Table 8. Evaluated metrics for emotion classification on FER-2013 dataset

Emotion class	**Precision**	**Recall**	**F1-Score**
Angry	0.512765957	0.490835031	0.501560874
Disgust	1	1	1
Fear	0.575862069	0.316287879	0.408312958
Happy	0.974416018	0.996587031	0.985376828
Sad	0.516693164	0.547138047	0.531479967
Surprise	1	0.995192308	0.997590361
Neutral	0.734375	0.976038339	0.838134431

F1-Score is the harmonic mean of precision and recall scores for a classification problem. It is advantageous when there is an unequal distribution.

$$\text{F1-Score} = 2 \text{ x Recall} * \text{Precision} / (\text{Recall} + \text{Precision}) \quad (10)$$

Accuracy it's the percentage of positive outcomes out of the total number of cases.

Table 9. Evaluated metrics for emotion classification on FERG dataset

Emotion class	Precision	Recall	F1-Score
Angry	0.995995996	0.995	0.995497749
Disgust	0.995987964	0.993	0.994491738
Fear	0.993006993	0.994	0.993503248
Happy	0.997	0.997	0.997
Sad	0.994005994	0.995	0.994502749
Surprise	0.992031873	0.996	0.994011976
Neutral	0.99498998	0.993	0.993993994

5.2. Confusion Matrix

To further investigate our proposed model, we plot the confusion matrix from the prediction class of the test data. Each row corresponds to a unique instance of the class and each column shows an instance of the anticipated class, There are seven target classes, and there are seven of them. Happy: 876/879, angry: 241/491, fear: 167/528, sad: 325/594, surprised: 414/416, and neutral: 611/626. The proposed model recognizes 2689 correct facial expressions out of 3589 emotions, for a 74.923 % accuracy.

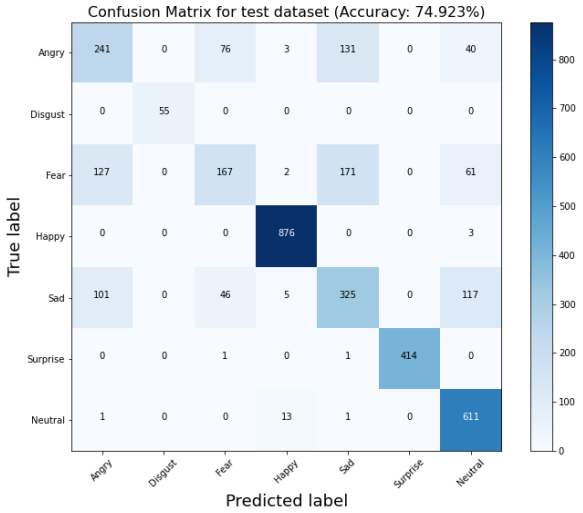

Figure 12. Confusion Matrix for proposed SDL model for FER-2013

Figures 12 and 13 depict the confusion matrix and their normalized view, which were obtained by testing our proposed spatial deep learning model on the FER-2013 database.

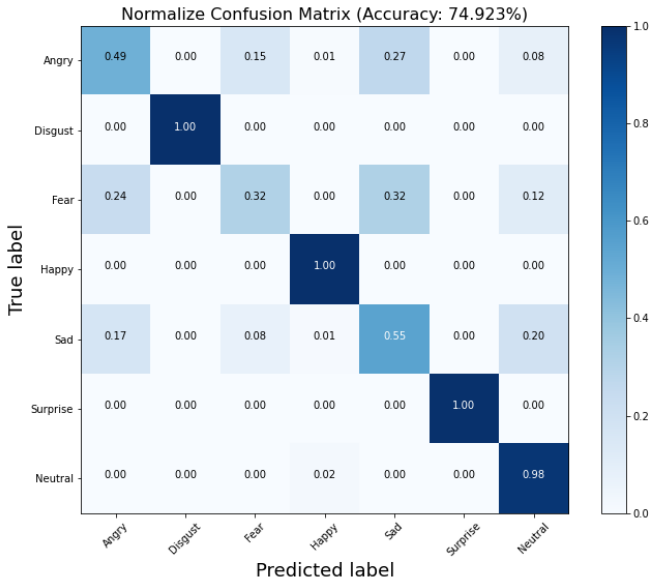

Figure 13. Confusion Matrix in Normalized form on FER-2013 dataset

Figure 14 depicts the suggested model's confusion matrix on the FERG dataset's test set.

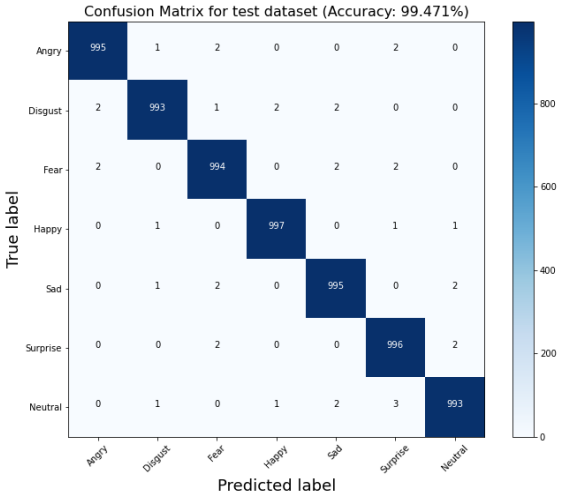

Figure 14. Confusion Matrix for proposed SDL model for FERG dataset

5.3. Visualization of Proposed Model

Saliency mapping is a term used to describe a CNN visualisation technique [19]. Saliency maps make it clear what a CNN is looking at when classifying data. In computer vision and deep learning, it is a crucial idea. How does proposed model know to concentrate on emotion-related pixels while ignoring the rest of the image's background while training over dataset is depicted in Figure 15. In a heat map, "hotness" denotes the areas of the image that have a significant impact on determining the emotion's class. In order to direct the choice of attended sites based on the spatial distribution of saliency, it seeks to identify the regions that are conspicuous or noticeable at every point in the visual field.

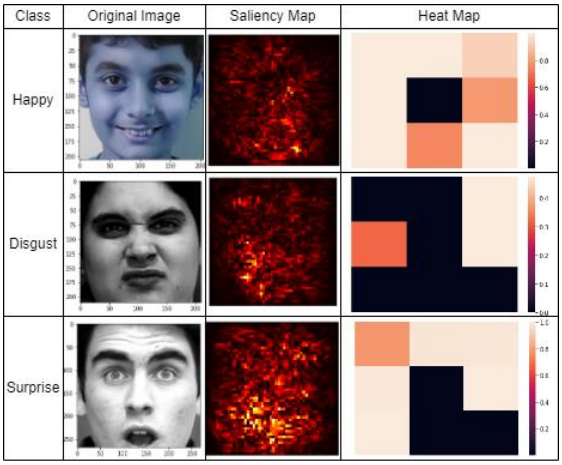

Figure 15. Some examples of saliency maps and heat maps

6. Conclusion and Future Scope

The proposed spatial deep learning model represents the state of the art in deep learning, and it is built on CNN and two Spatial Deep Learning modules that provide outstanding accuracy of 74.923% on the FER 2013 dataset. It also improved emotion prediction accuracy on the FERG and CK+ datasets, i.e., 99.47% and 98.5%, respectively. This model makes a unique contribution in that it improves feature extraction and classification techniques. It takes the features from the domain of the spatial feature set, which makes it most robust. This feature set focuses on the most significant aspects of the facial image, which plays an important role in expressing a person's mood and yields promising results across a wide number of test and validation samples. Our model is smaller in size than most other CNN models for common facial expression datasets; in addition, the suggested model is computationally simpler because it includes fewer hyper parameters and only 4 convolutional layers instead of others' equal to or more than 10-layer models. In the future, we can enhance the model to include micro-facial expressions, extract the features from wrinkles on the face, and work on a cross-data and cross-age data set.

Acknowledgement

We would like to thank the AIIT Amity University, Noida, India, and the IMS Engineering College, Ghaziabad, India, for helping us use the machine learning lab for doing our work.

References

[1] Siddheshwar S. Gangonda, Prashant P. Patavardhan and Kailash J. Karande, "VGHN: variations aware geometric moments and histogram features normalization for robust uncontrolled face recognition", *International Journal of Information Technology*, ISSN: 2511-2104, pp. 1823–1834, Vol. 14, 2022, Springer Nature, DOI: 10.1007/s41870-021-00703-0.

[2] K. Jayanthi, S. Mohan and B. Lakshmipriya, "An integrated framework for emotion recognition using speech andstatic images with deep classifier fusion approach", *International Journal of Information Technology*, ISSN: 2511-2104, pp. 3401–3411, Vol. 14, 2022, Springer Nature, DOI: 10.1007/s41870-022-00900-5.

[3] Xingcam Liang, Jinfu Liu, Zhipeng Liu, Wenxiang Zhang, Yan Zhang *et al.*, "A convolution-transformer dual branch network for head-pose and occlusion facial expression recognition", *The Visual Computer*, ISSN: 0178-2789, 13th February, 2022, Springer Nature, DOI: 10.1007/s00371-022-02413-5.

[4] Sumeet Saurav, Prashant Gidde, Ravi Saini and Sanjay Singh, "Dual integrated convolutional neural network for real-time facial expression recognition in the wild", *The Visual Computer*, ISSN: 0178-2789, pp. 1083–1096, Vol. 38, No. 3, 2022, DOI: 10.1007/s00371-021-02069-7.

[5] Nizamuddin Khan, Ajay Vikram Singh and Rajeev Agrawal, "Enhance Deep Learning Hybrid model of CNN based on Spatial Transformer Network for Facial Expression Recognition", *International Journal of Pattern Recognition and Artificial Intelligence*, ISSN: 0218-0014, Vol. 36, No. 14, 2252028, November 2022, Published by World Scientific, DOI: 10.1142/S0218001422520280.

[6] Jie Shao and Qiyu Cheng, ''E-FCNN for tiny facial expression recognition'', *Applied Intelligence*, ISSN: 0924669X, pp.549–559, Vol. 51, No. 1, January 2021, DOI:10.1007/s10489-020-01855-5.

[7] Jiahong Yang, Zhisheng LV, Kai Kuang, Sen Yang, Liuming Xiao *et al.*, "RASN: Using Attention and Sharing Affinity Features to Address Sample Imbalance in Facial Expression Recognition", in *IEEE Access*, ISSN: 2169-3536, pp. 103264-103274, Vol. 10, 2022, DOI: 10.1109/ACCESS.2022.3210109.

[8] Jamal Hussain Shah, Muhammad Sharif, Mussarat Yasmin and Steven Lawrence Fernandes, "Facial expressions classification and false label reduction using LDA and threefold SVM", *Pattern Recognition Letters*, ISSN: 0167-8655, pp. 166-173, Vol. 139, November, 2020, DOI: 10.1016/j.patrec.2017.06.021.

[9] Radu Tudor Ionescu, Marius Popescu and Cristian Grozea, "Local learning to improve bag of visual words model for facial expression recognition", in *Proceedings on challenges in representation learning, ICML 2013 Workshop on Representation Learning*, 2013, Atlanta, Georgia, USA, Published by Citeseer, Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=97088cbbac03bf8e9a209403f097bc9af46a4ebb.

[10] Yingli Tian, Takeo Kanade and Jeffrey F. Cohn, "Facial expression recognition", In *Studies in Handbook of Face Recognition*, Springer, 2011, pages 487–519, ISBN: 978-0-85729-931-4, DOI: 10.1007/978-0-85729-932-1_19.

[11] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks", Part of *Advances in Neural Information Processing Systems (NIPS)*, 25th International Conference, 2012, ISBN: 9781627480031, Vol. 25, pp. 1097–1105, 2012, Published by Curran Associates, Inc., Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonatheesh Krause, Sanjeev Satheesh *et al.*, "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, pp. 211–252, Vol. 115, No. 3, 2015. DOI: 10.1007/s11263-015-0816-y.

[13] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza *et al.*, "Challenges in Representation Learning: A report on three machine learning contests", In *Proceedings* of *International Conference on Neural Information Processing (ICONIP)*, 3-7 November 2013, Daegu, Korea, Online ISBN: 978-3-642-42050-4, Vol. 8228, pp. 117–124, Published by Springer, Berlin, Heidelberg, DOI: 10.1007/978-3-642-42051-1_16.

[14] Tee Connie, Mundher Al-Shabi, Wooi Ping Cheah and Michael Goh "Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator", In *Proceedings of Multi-disciplinary Trends in Artificial Intelligence, 11th International Workshop, MIWAI 2017*, 20-22 November, 2017, Gadong, Brunei, ISBN: 978-3-319-69455-9, pp. 139– 149, Vol. 10607, Published by Springer, DOI: 10.1007/978-3-319-69456-6_12.

[15] Deepali Aneja, Alex Colburn, Gary Faigin, L. Shapiro and Barbara Mones, "Modeling stylized character expressions via deep learning", In *Asian Conference on Computer Vision*; 2016, Springer: Cham, Switzerland, pp. 136–153, DOI: 10.1007/978-3-319-54184-6_9.

[16] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambad *et al.*, "The extended Cohn-Kanade dataset (CK+), A complete dataset for action unit and emotion-specified expression", *Computer Vision and Pattern*

Recognition Workshops (CVPRW), IEEE, 2010, San Francisco, CA, USA, pp. 94-101, DOI: 10.1109/CVPRW.2010.5543262.

[17] Shervin Minaee, Mehdi Minaai and Amirali Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network", *Sensors*, ISSN: 1424-8220 3046, Vol. 21, No. 9, 2021, DOI: 10.3390/s21093046.

[18] Khai Dinh Lai, Thuy Thanh Nguyen and Thai Hoang Le, "Detection of lung nodules on CT images based on the Convolutional Neural Network with Attention Mechanism", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 77-89, Vol. 5, No. 2, 1st April 2021, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2021.02.007, Available: http://aetic.theiaer.org/archive/v5/v5n2/p7.html.

[19] Matthew D. Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks", In *Proceedings of the Computer Vision - ECCV 2014 13th European Conference*, 6-12 September, 2014, Zurich, Switzerland, ISBN: 978-3-319-10589-5, Vol. 8689, pp 818-833, Published by Springer Verlag, DOI: 10.1007/978-3-319-10590-1_53.

[20] Ali Mollahosseini, David Chan and Mohammad H. Mahoor, "Going deeper in facial expression recognition using deep neural networks", In *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 7-10 March 2016, Lake Placid, NY, USA, , pp. 1-10, DOI: 10.1109/WACV.2016.7477450.

[21] Mariana-Iuliana Georgescu, Radu Tudor Ionescu and Marius Popescu, ''Local learning with deep and handcrafted features for facial expression recognition", *IEEE Access*, ISSN: 2169-3536, pp. 64827–64836, Vol. 7, 2020, DOI: 10.1109/ACCESS.2019.2917266.

[22] Panagiotis Giannopoulos, Isidoros Perikos and Ioannis Hatzilygeroudis, "Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013", *Advances in Hybridization of Intelligent Methods*, Springer Nature, ISBN 978-3-319-66789-8, pp. 1-16, Vol. 85, 2018, DOI: 10.1007/978-3-319-66790-4_1.

[23] Xiaoqing Wang, Xiangjun Wang and Yubo Ni, "Unsupervised domain adaptation for facial expression recognition using generative adversarial networks", *Computational Intelligence and Neuroscience*, ISSN: 1687-5273, pp. 1-10, Vol. 2018, Article ID 7208794, DOI: 10.1155/2018/7208794.

[24] Pham The Vinh and Truong Quang Vinh, "Facial Expression Recognition System on SoC FPGA", In *Proceedings of the IEEE 2019 International Symposium on Electrical and Electronics Engineering (ISEE)*, October 2019, Ho Chi Minh City, Vietnam, pp. 1-4, DOI: 10.1109/ISEE2.2019.8921140.

[25] Jeffrey Dean, Greg S. Corrado, Rajat Monga, Kai Chen, Matthieu Devin *et al*., "Large scale distributed deep networks", In *Advances in Neural Information Processing Systems*, pp. 1223 −1231, Vol. 25, 3rd December 2012, Curran Associates Inc., Available: https://typeset.io/papers/large-scale-distributed-deep-networks-22q3vnn2cn.

[26] Jaderberg, Max, K. Simonyan and A. Zisserman, "Spatial Transformer Networks ", in *Advances in Neural Information Processing Systems (NIPS)*, ISBN: 9781510825024, pp. 1-9, Vol. 28, 7-12 December 2015, Montreal, Canada, Curran Associates, Inc, Available: https://papers.nips.cc/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf.

[27] Feng Liu, Yurong Qian, Hua Li, Yongqiang Wang and Hao Zhang, "CAFFNet: Channel Attention and Feature Fusion Network for Multi-target Traffic Sign Detection", *International Journal of Pattern Recognition and Artificial Intelligence*, ISSN: 0218-0014, Vol. 35, No. 07, 2152008, 2021, DOI: 10.1142/S021800142152008X.

[28] Xiangyum Zhao, Xiaodam Liang, Luoqi Liu, Teng Li, Yugang Han *et al.* "Peak-piloted deep network for facial expression recognition", In *European Conference on Computer Vision*, 17th September 2016, Switzerland, ISBN: 978-3-319-46474-9, pp. 425–442, DOI: 10.1007/978-3-319-46475-6_27.

[29] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza *et al*., "Challenges in Representation Learning: A report on three machine learning contests", *Neural Networks*, ISSN: 0893-6080, pp. 59-63, Vol. 64, 2015, DOI: 10.1016/j.neunet.2014.09.005.

[30] Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi and Jiro Gyoba, "The Japanese female facial expression (JAFFE) database", in *Proceedings of the third international conference on Automatic Face and Gesture Recognition*, 1998, pp. 200-205, DOI: 10.5281/zenodo.3451524, Available: https://zenodo.org/record/3430156.

[31] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong and Yang Li, "Spatial-temporal recurrent neural network for emotion recognition", *IEEE Transactions on Cybernetics*, ISSN: 2168-2267, pp: 839-847, Vol. 49, No. 3, March 2019, DOI: 10.1109/TCYB.2017.2788081.

[32] Abubakar M. Ashir, Alaa Eleyan and Bayram Akdemir, "Facial expression recognition with dynamic cascaded classifier", *Neural Computing Applications*, ISSN: 0941-0643, pp: 6295–6309, Vol. 32, No. 10, May 2020, DOI: 10.1007/s00521-019-04138-4.

[33] Hang Zhao, Qing Liu and Yun Yang, "Transfer learning with ensemble of multiple feature representations", In *Proceedings of the IEEE 2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA)*, 13-15 June 2018, Kunming, China, pp. 54-61, Published by IEEE, DOI: 10.1109/SERA.2018.8477189.