Facial Emotion Detection Using CNN (Convolutional Neural Network)

Akshat Pareta

Chandigarh University

20BCS6567@cuchd.in

Abstract

Numerous intriguing studies on the automatic detection of facial emotions have already been carried out recently (FER). Human centre computing and the most recent developments in emotional artificial intelligence are two areas where FER has been applied to improve human-machine interactions (EAI). Researchers working in the subject of EAI want to improve how adept computers are at spotting patterns in human behaviour and facial expressions. The largest impact on this area has come from deep learning since alternative designs are being created to tackle ever-harder issues as a result of the considerable evolution of neural networks in recent years. This article will discuss the most recent advances on automatic expression recognition in relation to computational intelligence using the most recent deep learning models. We show how model which make use of architecture-related methods, including databases, and FER that is based on deep learning can work well together to produce results that are extremely accurate.

1. Introduction

It is easier to identify facial expressions on a person because of their major and distinctive features. FER is defined as a change in facial expression triggered by an individual's emotional condition on the inside. In addition to machine learning, image analysis, and artificial intelligence, it is used in a range of human-computer interface (HCI) applications, such as facial image analysis, face surveillance footage, and facial animations.. The challenging subject of automatic facial expression identification has drawn the attention of numerous academics recently. The feature extraction stage is crucial in FER. In the literature, According to Alek et al.[1] , oral and written communications represent 38 percent and 7 percent, respectively, to overall transmission, whereas facial emotions account for 55 percent of it.

A FER system can be designed using one of two main methods. Some methods start with a succession of images that range from a neutrality face to the strongest emotion. As a result of the limited data they can access, other systems, on the other hand, only employ a single image of a person's face to detect ’s exactly, and thus often perform worse than top approaches [2,3]. A single or both of those characteristic groups are utilised by a FER system., and in addition to the approach type it models, features used in the recognition process are also categorized in this manner. The position of the first set of traits are supplied by the face organs and skin texture. The second type of feature is a geometric feature, which contains information about various facial positions and points and is used to evaluate a static image or a collection of photos by observing how well the locations and points change over time. Face landmarks can be utilized as a starting place for geometrical feature extraction in one technique. During a facial analysis, landmarks are significant areas of the face that can be utilized to collect information. Numerous studies on the subject of recognising facial landmarks have been done; nevertheless, they are not pertinent to our work. This study employs the Python module dlib to locate these spots.

2. Related Work

There have been significant advances in the development of automatic expression classifiers. in recent years [7, 8, 9]. Some systems for recognizing facial expressions classify the expression into a range of fundamental emotionss., including joy, sad and rage.. In an effort to provide an unbiased description of the face, others have made an effort to identify the

specific muscle movements that the face is capable of doing[11]. The most widely used psychological framework is the Face Action Coding System (FACS)[12]. for summarizing practically all facial movements. Using Action Units, the FACS system classifies human facial movements according to how they appear on the face (AU). A facial expression usually originates from one of the 46 atomic units (AUs) of face movement or related deformation that can be detected. Several AUs are usually added together to form an expression [7, 8]. Additionally, Multilevel Hidden Markov Model, Neural Networks, and Bayesian Networks (HMM) have all seen development in the approaches employed for face emotion identification [13],[14]. Several of them have problems with timing or detection rates. combining two or more techniques to accomplish precise recognition allows for the extraction of features as necessary. Because of illumination and feature extraction, Each technique's effectiveness is reliant on image pre-processing.

3. Methods

To assess how well these models performed at identifying facial expressions, we created CNNs with varying depths. For our analysis, we took into account the following network design.

[Conv2D - (SBN) – RELU - (Dropout)–(Max-pool)]
M - [Affine-(BN)-RELU-(Dropout)]
N - Affine - SoftMax.

All of these layers have the convolution layer and ReLU non - linear., are referred to as the first component of the network. These layers also include dropout, max-pooling, and spatial batch normalization (SBN), which can be present. After M convolutional layers, which are always Affine-operating & ReLU non-linear and may additionally include batch normalisation (BN) and dropout, the network is led to N fully linked layers. In the show's processing order, the affine layer follows the network and computes the score and softmax loss function.. The user has control over the created model's convolution and fully - connected layers layer counts along with the presence of batch normalisation, dropout, and max-pooling layers. We used L2 regularisation in addition to dropouts and batch normalisation methods. Additionally, the user can specify the quantity of filters, strides, and zero-padding; otherwise, the default settings are used.

As we will describe in the next part, we proposed the notion of combining HOG features with those extracted by convolutional layers utilising raw pixel data. We utilised the same architecture as before to accomplish this, but we now added the HOG characteristics to the features exiting final convolutional layer. The hybrid feature set is then passed to the fully connected layers, who use it to compute scoring and losses.

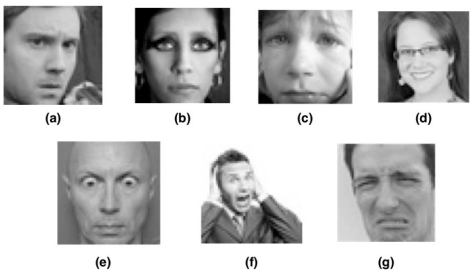


Figure 1: Examples of the 7 face expressions we are taking into account for this classification issue. Aside from being furious[a], you can also feel neutral[b], sadness [c], delighted[d], surprised [e], fearful[f], or disgusted[g].

Thus order to speed up the network train program, we developed that above models in Torch and used GPU acceleration deep learning characteristics.

4. Dataset and Feature

For this study, we was using a data from the Kaggle website, which consists of approximately 35000 well-structured 48x48 pixel grayscale photos of faces. Each face in each image about equals the amount of space it takes up in the processed photographs, which virtually perfectly centre the faces. The seven classes that represent various facial emotions must be applied to each image. These facial expressions have been divide in 6 classes: 0 represents rage, 1 disgusted, 2 fearfull, 3 happy, 4 sad, 5 surprised, and 6 neutral. Each category of facial expression is represented by one example in Figure 1. The provided photos are separated into three distinct sets, referred to as training, validation, and test sets, along with the images classes no. (between zero to six). In order to

normalise the raw pixel data after reading it, we subtracted the average of each image's training pictures, including those taken from the validation and testing set.. We created mirrored images by horizontally flipping photographs from the set is used to train for the enhancement of data.

The characteristics produced by convolution layers using the raw image data were mostly used to classify the expressions. As an additional experiment, we created learning models that fed the input features into Fully Connected (FC) layers by concatenating the HOG features with those produced by convolutional layers.

Parameters	Value
Learning Rate	0.001
Regularization	1e-6
Hidden Neurons	512

Table 1: The hyper-parameters obtained by cross validation for the shallow model

5. Analysis

5.1. Experiment

We first constructed a shallow CNN with the help of this project. One FC layer and two convolutional layers made up this network. 32 3x3 The first convolutional layer contained filtering with such a stride of length 1, batch normalisation, dropout, and no max-pooling. In the second convolution layers, 64 3x3 filter with such a strideod size of 1 were combined with batch normalisation, dropout, and max-pooling using 2x2 filters. In the FC layer, we employed 512 neurons inside a hidden units using Softmax as that of the loss function. Rectified-linear Unit (ReLU) was also used as an activation function for each layer.To ensure that the network's implementation was accurate before training our model, we performed a few sanity checks. The initial loss when regularization is not present was calculated as the first sanity check.

Since there are 7 different classes in our classifier, we anticipated receiving a result around 1.95. A tiny subset of the training data was used to attempt an over-fit of our model as a second sanity check. Both sanity tests on our shallow model were successful. Following that, we began completely training our model. We made use of Torch's Deep learning capabilities that are GPU-accelerated to hasten model training.

To study the effects of adding convolution layer & FC-layers to a network, we trained a stronger C-NN with 4 convolution layer and two FC levels. There were 64 3x3 filters in the first convolutional layer, 128 5x5 filters in the second, 512 3x3 filters in the third, and 512 3x3 filters in the final layer. We have a stride of size 1, batch normalization, dropout, max-pooling, and ReLU as the activation function in all the convolutional layers. The second FC layer had 512 neurons, whereas the buried layer in the first FC layers had 256 neurons. As with the convolutional layers, batch normalization, dropout, and ReLU were utilized in both FC layers. Additionally, Softmax served as our loss function. Figure 2 shows the architecture of this convolutional model. We performed preliminary loss checking before training the network, as we did with the shallow model, and we looked at the possibility of overfitting the network using only a small fraction of the training set. The results of these sanity checks showed that the network implementation was correct. Then, with 35 epochs as well as batch size of 128, the network was trained utilizing each of the images from the training set.In order to obtain the model with the maximum accuracy, we also cross-validated the hyper-parameters.

Moreover, we developed networks with five and six convolution layer.to investigate the deeper CNNs, but these networks did not improve classification accuracy. the model has two FC layers and four convolutions layer was therefore deemed to be the best network for our dataset. We solely utilized the features produced by the convolution layers in both the shallow and deep models, the important components of our classification task are taken from the original image pixels. Because HOG characteristics are sensitive to edges, they are frequently employed for facial expression identification. In order to see how well the model performs when it combines two separate characteristics, we decided to see if

there was a method to Raw pixels and HOG characteristics should be added to our network. To do this, we developed a novel learning model consisting of two neural networks, the first of which contained convolution layer as well as the second of which only contained fully-connected-layers.The first network's features are combined with the HOG features to create hybrid features, which are then sent into the second network. In order to assess the hybrid network's performance, We created two networks by training them similar to the deep and shallow networks we developed for the previous experiment. This instance, the accuracy of the shallow model was pretty comparable to that of a shallow model that used only raw pixels. The accuracy of the deep model was similar to what we attained with our own deep model that used raw pixel as characteristics.

Parameter	Value
Learning Rate	0.01
Regularization	1e-7
Hidden Neurons	256, 512

Table 2: The hyper-parameters obtained by cross validation for the deep mode

5.2. Results

We charted the lost history and measured these models' precision. to assess how well the shallow model and deep model performed. These outcomes are shown in Figures 3 and 4. The validation accuracy was increased by 18.46% thanks to the deep net effort, as shown in Figure 4Along with L2 regularisation, additional non-linearity and hierarchy anti-overfitting techniques, such as batch normalisation and dropout,may be seen in the deep network, which has also been seen to lessen the overfitting behavior of the learning model. The training accuracy quickly reached its greatest value, as shown in Figure 3, and the shallow network converged more quickly.

The confusion matrices for the deep and shallow networks were also computed. A depiction of a confusion matrice is shown in Fig 5 and 6. These numbers show that for the majority of the labels, the deep network produces higher correct predictions. It's intriguing to observe how well both models predicted the happy label, suggesting that it may be simpler to acquire the characteristics a smile rather than other emotions.. These matrices also show the labels that the trained networks are most likely to misinterpret. As an illustration, we can observe the relationship between the furious label and the fear and sad labels. There are many situations where people are misclassified as fearful or depressed when their genuine emotion is anger. These errors match what we notice when viewing photos in

Emotions	Shallow Model	Deep Model
Angry	43.4%	54.6%
Disgust	31.2%	68.99%
Fear	53%	45%
Happy	71%	79.99%
Sad	32.9%	65.8%
Surprise	69.5%	64.5%
Neutral	37.9%	52.9%

Table 3: The accuracy of each expression in the shallow and deep models.

the data; Even for humans, determining whether a face is sad or angry can be difficult. This is due to the fact that not everyone displays their emotions the same way.

Additionally to confusion matrices, we determined the accuracy of each model for each expression.

This data is shown in [3]. The In both deep and shallow models, the accuracy in detecting a happy expression has the highest value of all the emotions., as can be seen in the table. Deep network technology has also improved categorization accuracy for the majority of phrases. In other words, adding more features does not always result in superior features for particular expressions.

As stated in the previous section, we created learning models that combined the HOG features with those produced by the convolutional layers

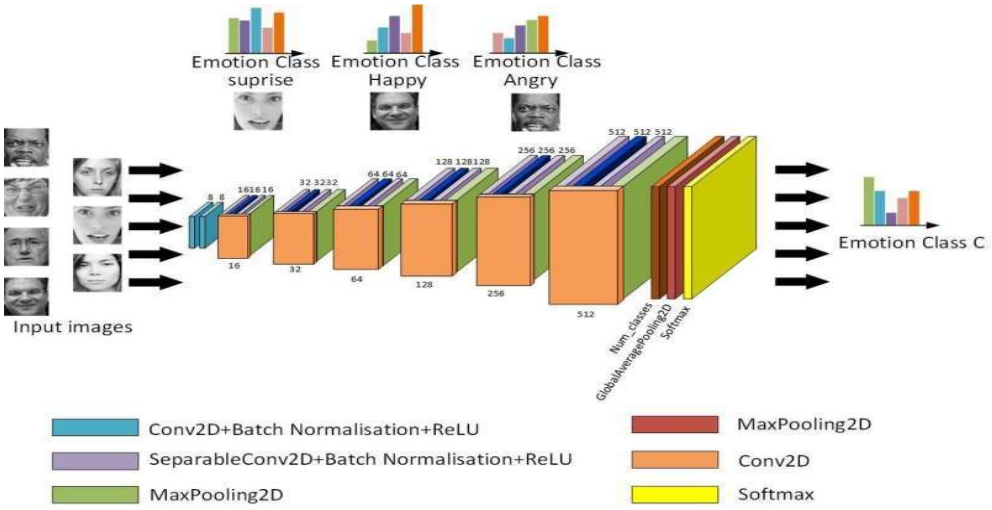


Fig 2: The deep network's design consists of two fully linked layers and four convolutional layers.

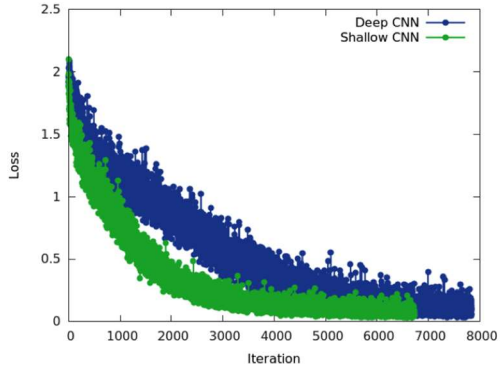


Fig 3: The deep and shallow models' loss histories

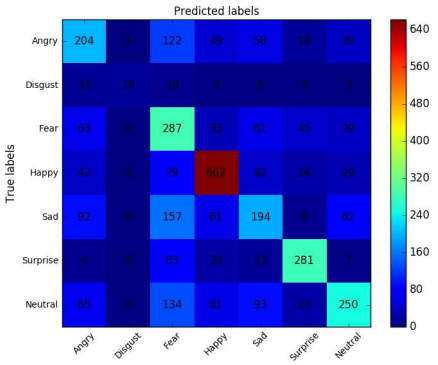


Fig 5: Confusion-matrix for shallow-model

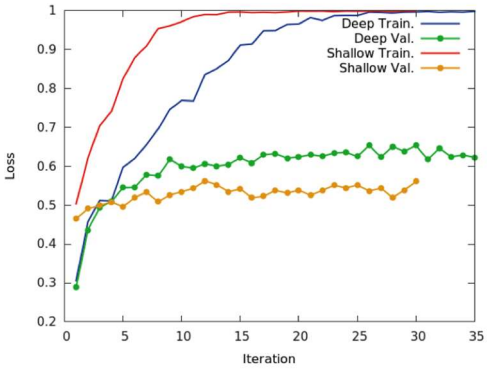


Figure 4: The effectiveness of a deep and shallow models for various iterations

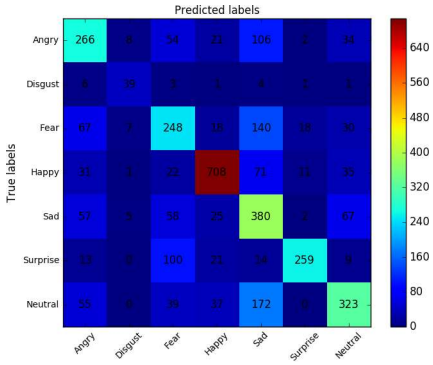


Fig 6: Confusion-matrix for Deep-model

utilized them as input features to the FC layers in order to test the impact of using various features in our CNN model. We trained one deep network and one Shallow network using this concept. The acquired accuracy for the s deep and shallow

The acquired accuracy for the deep and shallow models, respectively, is shown in Fig 8 and 7. These figures show that the model's accuracy is extremely close to the accuracy we obtained from the model without HOG elements. This suggests that with only raw pixel data, CNN is powerful

enough to extract enough information, including that derived from HOG features.

During the forward pass, we displayed the activation maps of several layers to track the features that each layer of our training set retrieves. This visualization is displayed in Figure 9. We may observe that the activation maps becoming sparser and more localized as the training goes on.

We also showed the weights of the first layer to show the qualification of the trained network. We offer quiet filters that don't produce noisy patterns., as shown in Figure 10. It shows that our network has received sufficient training, and the regularization strength is likewise adequate.

In order to uncover more complex patterns in our photos, we also applied the DeepDream [16, 17] technique to our top predictive model. Each expression's example and DeepDream output are shown in Figure 11 together.

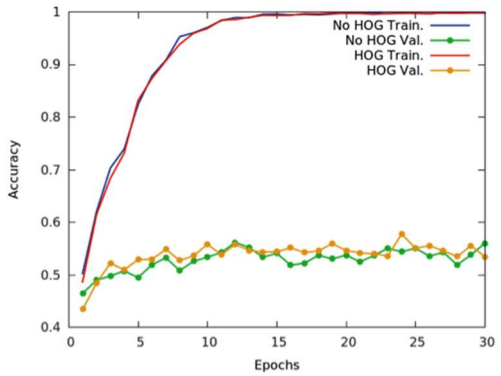


Fig 7: Accuracy of shallow model

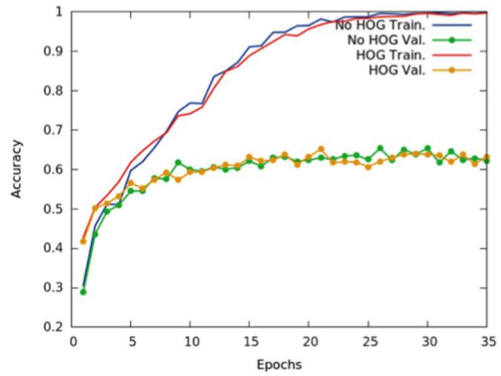


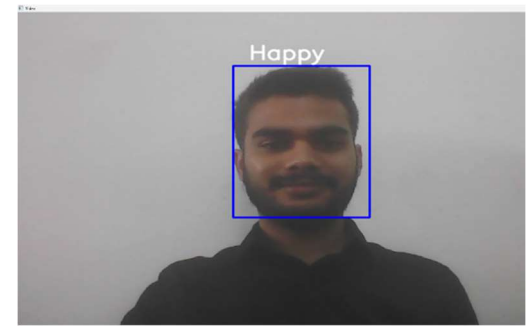
Fig 8: Accuracy of deep model

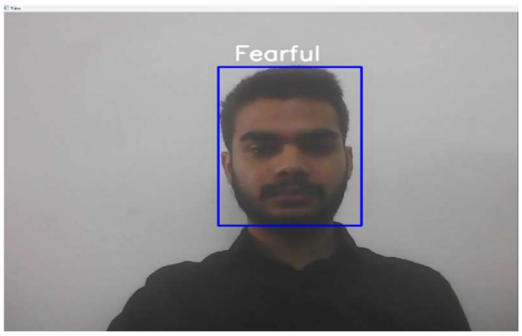
1. Summary

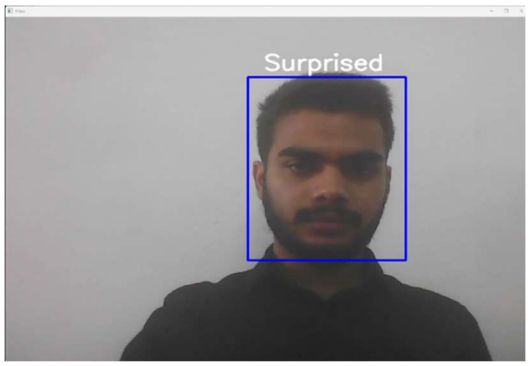
1.1. Conclusion

We developed a variety of CNNs for a recognition of facial expressions job and evaluated its efficiency using different post-processing and visualisation methods. The results demonstrated that deep CNNs can learn facial characteristics and enhance face expression recognition. Additionally, the hybrid feature sets had little effect on the accuracy of the model, indicating that convolutional networks may naturally learn the main facial traits even when given simply raw pixel data.

Following are some images of the real time detection :







1.1. Future Work

We used Torch's CNN packages to create all of the models for this project from scratch. In subsequent work, we hope to expand our model to include colour images.

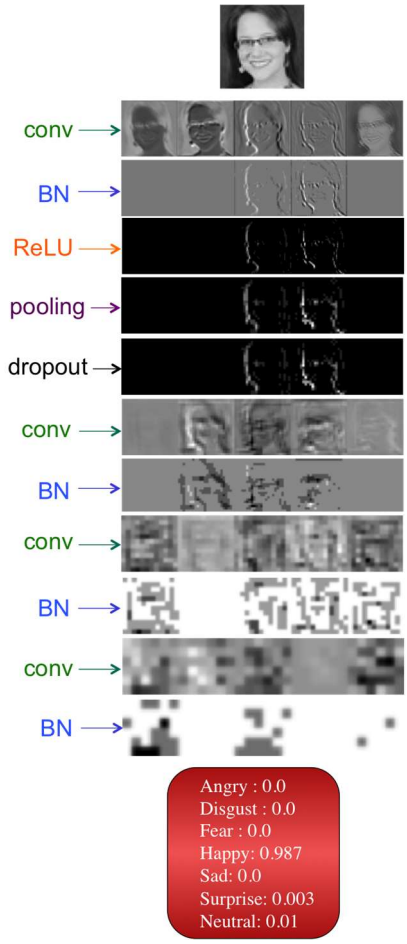


Fig 9: An illustration of our CNN's activation maps for several layers

This allow us to examine effectiveness of pretrained model for face emotion recognition, such as AlexNet or VGGNet. Implementing a face identification approach and then an emotion prediction mechanism would be another extension.

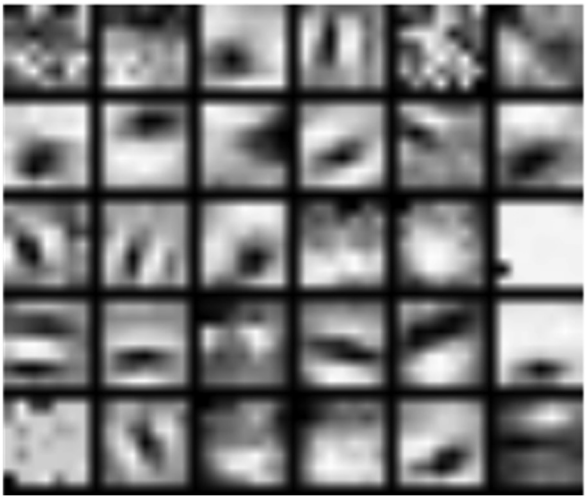


Fig 10: Visualization of the weights for the first layer in our CNN

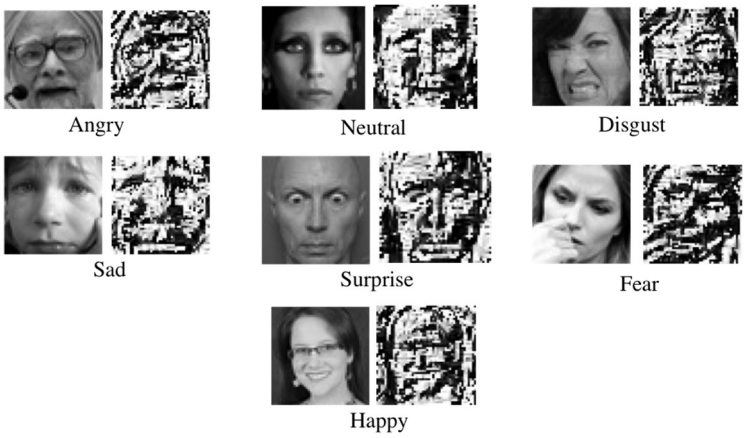


Figure 11: Examples of applying DeepDream on ourdataset

References

[1] https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data

[2] https://github.com/torch

[3] Dalal, Navneet, and Bill Triggs (2005). Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on. Vol. 1

[4] Bettadapura, Vinay (2012). Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722

[5] Lonare, Ashish, and Shweta V. Jain (2013). A Survey on Facial Expression Analysis for Emotion Recognition. International Journal of Advanced Research in Computer and Communication Engineering 2.12

[6] Nicu Sebe, Michael S. Lew, Ira Cohen, Yafei Sun, Theo Gevers, Thomas S. Huang (2007) Authentic Facial Expression Analysis. Image and Vision Computing 25.12: 1856-1863

[7] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2), 2001.

[8] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In Proceedings of the IEEE Conference on Automatic Facial and Gesture Recognition, 2006.

[9] M. Pantic and J.M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. IEEE Transactions on Systems, Man and Cybernetics, 34(3), 2004

[10] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. Image and Vision Computing, 24(6), 2006.

[11] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, and J.R. Movellan. Automatic recognition of facial actions in spontaneous expressions. Journal of Multimedia, 2006.

[12] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, 1978

[13] Cohen, Ira, et al. "Evaluation of expression recognition techniques." Image and Video Retrieval. Springer Berlin Heidelberg, 2003. 184-195

[14] Padgett, C., Cottrell, G.: Representing face images for emotion classification. In: Conf. Advances in Neural Information Processing Systems. (1996) 894900.

[15] http://scikit-learn.org/stable/

[16] http://deepdreamgenerator.com/

[17] https://github.com/google/deepdream

[18] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[19] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[20] P. S Aleksic and A. K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multistream hmms," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 1, pp. 3–11, 2006. View at: Publisher Site | Google Scholar

[21] H. Chouhayebi, J. Riffi, M. A. Mahraz, Y. Ali, and T. Hamid, "Facial expression recognition using machine learning," in *Proceedings of the 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)*, pp. 1–6, IEEE, Seoul, Korea, November 2021. View at: Publisher Site | Google Scholar