International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 8 Issue 3, May-June 2024 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Convolutional Neural Networks for Facial Expression Recognition

Rashmi Yadav[1], Mr. Ghanshyam Sahu[2], Lalitkumar P Bhaiya[3]

[1]M Tech Scholar, BCET, Durg, Chhattisgarh, India
[2]Assistant Professor, CSE Department, BCET, Durg, Chhattisgarh, India
[3]Associate Professor, Bharti Vishwavidyalaya, Durg, Chhattisgarh, India

ABSTRACT

In the context of this study, convolutional neural networks (CNNs) have been developed with the objective of identifying facial expressions. The primary aim of this study project is to categorize each facial image into one of the seven distinct categories of facial expressions under investigation. The training of Convolutional Neural Network (CNN) models with different levels of depth included the use of grayscale photographs sourced from the Kaggle website [1]. By using Torch [2], we successfully developed our models and used the computational capabilities of Graphics Processing Units (GPUs) to enhance the efficiency of the training procedure. In addition, we used a hybrid feature method alongside the networks that were operating on raw pixel data. By integrating raw pixel data with Histogram of Oriented Gradients (HOG) characteristics, we were able to train a distinctive CNN model [3]. We used several techniques, including dropout and batch normalization, along with L2 regularization, to mitigate the occurrence of overfitting in the models. Cross-validation was used to determine the optimal hyper-parameters, and the performance of the generated models was assessed by examining their individual training histories. Furthermore, we provide a visual representation of the several layers inside a network to illustrate the attributes of a facial feature that may be acquired using Convolutional Neural Network (CNN) models.

How to cite this paper: Rashmi Yadav | Mr. Ghanshyam Sahu | Lalitkumar P Bhaiya "Convolutional Neural Networks for Facial Expression Recognition" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-8 | Issue-3, June 2024, pp.277-283, URL: www.ijtsrd.com/papers/ijtsrd64783.pdf

IJTSRD64783

KEYWORDS: *Face Recognition, Image Processing, Computer Vision, Emotion Detection, OpenCV*

1. INTRODUCTION:

Humans primarily engage in communication via words, as well as using bodily gestures, to highlight certain elements of their speech and express their emotions. Facial expressions play a crucial role in human communication, serving as a prominent means of conveying emotions. Despite the absence of verbal communication, there exists a substantial amount of information that may be inferred from the signals sent via nonverbal means. Facial expressions serve as a means of conveying non-verbal communication. Linguistic cues are crucial in facilitating interpersonal connections between individuals [4, 5]. The automatic recognition of facial expressions has significant promise as a fundamental element in facilitating natural interactions between humans and machines. Indeed, it has the capacity to be used in the realm of behavioral research, as well as in the context of therapeutic intervention. However, despite the inherent ability of humans to effortlessly and promptly detect facial expressions, robots continue to encounter challenges in reliably recognizing emotional states on the face. In recent years, significant progress has been made in the fields of face identification, feature extraction techniques, and expression classification methodologies. However, the development of an automated system capable of successfully doing this job remains challenging [6]. The objective of this research is to propose a Convolutional Neural Network (CNN)-based approach for the identification and classification of facial expressions. The input for our system is a picture, and we use Convolutional Neural Networks (CNN) to predict the face expression label. The label should fall into one of the following categories: Anger, joy, apprehension, sadness, hatred, or neutrality.

2. Related Work

In recent years, researchers have made notable progress in the development of autonomous expression classifiers [7, 8, 9]. Several facial expression recognition algorithms classify the face into a set of attributes that are often associated with emotions, including pleasure, sadness, and rage. The number is 10. To provide an impartial depiction of the face, several individuals endeavor to ascertain the precise muscular motions that the face can do [11]. The Facial Action Coding System (FACS) [12] is a well recognized psychological framework that may be used to effectively characterize almost all facial movements. The FACS system employs Action Units (AU) as a means of classifying human facial movements based on their visual manifestation on the face. The formation of a phrase often arises from the amalgamation of many AUs [7, 8]. An atomic unit (AU) is one of the 46 atomic elements that have a role in perceptible facial movement or the associated deformation.

Furthermore, significant progress has been achieved in the techniques used for the identification of face expressions. The aforementioned models include Bayesian Networks, Neural Networks, and the multi-level Hidden Markov Model (HMM) [13, 14]. There is a potential for some individuals to have difficulties pertaining to the pace of recognition or the time. Typically, to achieve accurate recognition, it is feasible to integrate many algorithms and then extract pertinent features as needed. The efficacy of each strategy is dependent on the pre-processing of the photographs, since it is influenced by lighting conditions and feature extraction techniques.

3. Methods

In order to assess the efficacy of these models in the context of face emotion identification, we constructed Convolutional Neural Networks (CNNs) with different levels of depth. During the analytical process, we considered the following network architecture:

The given expression is denoted as [Conv-(SBN)-ReLU-(Dropout)-(Max-pool)].The letter M is represented by [Affine-(BN)-ReLU-(Dropout)].The N-Affine-Softmax is being referred to.

The network's first component consists of M convolutional layers. The aforementioned layers exhibit spatial batch normalization (SBN), dropout, and max-pooling, alongside the convolution layer and ReLU nonlinearity, which are consistently seen in these layers. After the M convolution layers have been completed, the network is then steered to N fully connected layers. Affine operation and ReLU non-linearity are consistently present in these layers, which may also include batch normalization (BN) and dropout methods. The last layer of the network is the affine layer, which is responsible for calculating the scores and the softmax loss function. The model allows the user to exercise control over the number of convolutional and fully connected layers, as well as the inclusion of batch normalization, dropout, and max-pooling layers. To include L2 regularization into our system, we used dropout and batch normalization approaches. Furthermore, the user has the capability to designate the quantity of filters, strides, and zero-padding. In the absence of these specifications, the default values remain in effect.

In the next section, we propose the idea of combining HOG features with those obtained from convolutional layers using raw pixel data. We have offered this strategy. In order to do this, we used the same architecture as previously stated, with the exception that we included the HOG characteristics with the features that were exiting the final convolution layer. The hybrid feature set is then included into the fully linked layers, where it is used for the calculation of both score and loss.

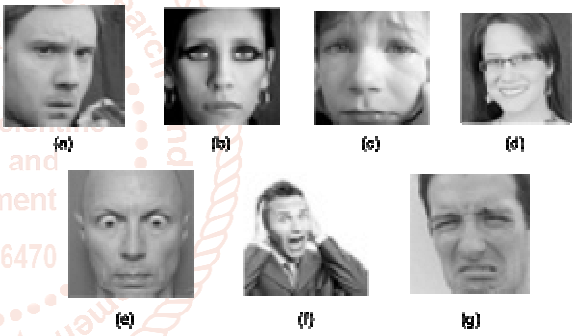

Figure 1: Examples of seven facial emotions that we consider in this classification problem. (a) angry, (b) neutral, (c) sad, (d) happy, (e) surprise, (f) fear, (g) disgust

We used Torch to create the model that was stated before, and we took advantage of GPU-accelerated deep learning features in order to speed up the process of training the model.

4. Dataset and Features

For this study, we made use of a dataset that was made available to us via the Google website. This dataset contains around 37,000 grayscale photos of faces that are well-structured and 48 by 48 pixels in size. During the processing of the photos, the faces are positioned in such a manner that they are almost perfectly centered, and each face takes up approximately the same amount of space in each image. Every picture needs to be placed into one of the seven categories that correspond to the various expressions that can be seen on the face. These facial expressions have been classified as follows: 0

represents anger, 1 represents disgust, 2 represents fear, 3 represents happiness, 4 represents sadness, 5 represents surprise, and 6 represents neutrality. All of the categories of facial expressions are illustrated with a single example in Figure 1. In addition to the picture class number, which is some number between 0 and 6, the images that are provided are separated into three distinct groups. These sets are referred to as the training set, the validation set, and the test set. In all, there are around 29,000 photos for training, 4,000 images for validation, and 4,000 images allocated for testing. In order to normalize the raw pixel data, we first read it and then subtracted the mean of the training photos from each image, including those in the validation and test sets. This allowed us to normalize the data. We produced mirrored pictures for the purpose of data augmentation by flipping photos in the training set horizontally. This allowed us to produce mirrored images.

During the process of classifying the expressions, we relied mostly on the features that were produced by the convolution layers by making use of the raw pixel data. As a further investigation, we developed learning models that concatenate the HOG features with those generated by convolutional layers and offer them as input features into Fully Connected (FC) layers. This was done as an additional research.

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

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

5. Analysis

5.1. Experiments

To begin this endeavor, we first constructed a hal-low Convolutional Neural Network (CNN). The system consisted of two convolutional layers and one fully connected (FC) layer. The first convolutional layer consisted of 32 3 3 filters, each with a stride size of 1. Additionally, batch normalization and dropout were used, although max-pooling was not performed. In the subsequent convolutional layer, a total of 64 3 3 filters were used, each with a stride size of 1. Additionally, batch normalization and dropout were applied, and max-pooling was performed using a filter size of 2 2. The FC layer consisted of a hidden layer with 512 neurons, using the loss function Softmax. Furthermore, the Rectified Linear Unit (ReLU) was used as the activation function in all of the layers. Prior to training our model, we conducted sanity tests to verify the accuracy of the network's implementation. In the first sanity check, the loss was calculated in the absence of regularization. Given that our classifier consists of 7 distinct classes, we anticipated obtaining a result of about 1.95. For the second sanity check, we attempted to subject our model to overfitting by using a limited portion of the training data. Both of these sanity tests were successfully passed by our shallow model. Next, we began the training of our model from the beginning. In order to enhance the efficiency of the model training process, we used the GPU-accelerated deep learning capabilities available on Torch. During the training phase, we used the whole of the pictures inside the training set, using 30 epochs and a batch size of 128. Additionally, we conducted cross-validation of the model's hyper-parameters by varying the values for regularization, learning rate, and the number of hidden neurons. In each iteration, we used the validation set to verify our model and the test set to assess its performance. The most optimal shallow model achieved an accuracy of 55% on the validation set and 54% on the test set. The hyper-parameters acquired via cross validation for the shallow model are summarized in Table [1].

In order to examine the impact of including convolutional layers and FC layers into the network, we conducted training on a more advanced CNN model consisting of four convolutional layers and two FC layers. The first convolutional layer consisted of 64 3 3 filters, followed by a subsequent layer of 128 5 5 filters. The third layer was composed of 512 3 3 filters, while the final layer further consisted of 512 3 3 filters. The activation function used in all convolutional layers includes a stride size of 1, batch normalization, dropout, max-pooling, and ReLU. The first FC layers consisted of a hidden layer including 256 neurons, whereas the subsequent FC layer consisted of 512 neurons. We used batch normalization, dropout, and ReLU in both FC layers, as well as in the convolutional layers. In addition, the Softmax loss function was used. The architecture of the deep network is shown in Figure 2. Prior to training the network, we conducted an initial loss assessment and evaluated the network's susceptibility to overfitting using a limited portion of the training dataset, similar to the shallow model. The outcomes of these sanity tests demonstrated the accuracy of the network's implementation. Subsequently, the network was trained using 35 epochs and a batch size of 128, including all the pictures included in the training set. Furthermore, we conducted cross-validation of the hyperparameters in order to get the model that exhibits the maximum level of accuracy. In this instance, we achieved a validation set accuracy of 65% and a test set accuracy of 64%. The values for each hyper-parameter in the model with the best accuracy are shown in Table [2].

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

Table 2: The hyper-parameters obtained by crossvalidation for the deep model

In addition, we conducted training on CNNs with 5 and 6 convolutional layers to investigate their deeper capabilities. However, these networks did not result in an improvement in classification accuracy. For our dataset, we have identified the model with 4 convolutional layers and 2 FC layers as the optimal network. For our classification assignment, we only used the features produced by the convolution layers in both the shallow and deep models. These features were derived from the raw pixel data. HOG features are often used in the field of face expression identification because to their heightened sensitivity towards edges. We want to investigate the feasibility of including HOG features in conjunction with raw pixels into our network and evaluate the model's performance when it incorporates both features. In this study, a novel learning model was constructed, including two distinct neural networks. The first model had convolutional layers, whereas the subsequent model only consisted of fully connected layers. The first network's features are combined with the HOG features, resulting in hybrid features that are then inputted into the second network. In order to assess the efficacy of the network including hybrid features, we conducted training on two distinct networks: a shallow network and a deep network. These networks, which were trained in a previous experiment, exhibited identical characteristics to the shallow and deep networks constructed in the preceding experiment. In this specific case, the precision of the shallow model closely resembled the precision achieved by the shallow model that just relied on raw particles. Furthermore, the precision of the deep model exhibited a level of accuracy that was similar to the precision achieved by our deep model while using raw pixels as distinctive features.

5.2. Results

In order to assess the performance of the shallow model compared to the deep model, it was essential for us to graph the loss history and the attained accuracy for both models. The results are shown in both Figure 3 and Figure 4. Based on the analysis of Figure 4, it is evident that the use of the deep neural network facilitated a notable improvement in the validation accuracy, amounting to 18.46%. Moreover, it is evident that the deep neural network has effectively mitigated the issue of overfitting in the learning model. The aforementioned objective was achieved with the use of enhanced non-linearity and the hierarchical utilization of anti-overfitting techniques, including dropout and batch normalization, alongside L2 regularization. As seen in Figure 3, the shallow network exhibited a higher rate of convergence, resulting in a speedy attainment of the highest training accuracy within a short timeframe.

Furthermore, the confusion matrices were computed for both the shallow and deep networks [15]. The confusion matrices are shown in Figures 5 and 6, correspondingly. The figures above demonstrate that the deep neural network generates more precise predictions for most of the labels. Considering the high performance of both models in accurately predicting the happy label, it is important to highlight that this implies that acquiring knowledge about the attributes of a happy face is easier than acquiring knowledge about the attributes of other facial expressions. Furthermore, these matrices provide insights into the labels that are prone to being misidentified by the trained networks. When using the anger label as an illustrative example, it becomes evident that there exists a correlation between this label and both the fear label and the sad label. There are other instances when the individual being categorized has erroneously categorized them as afraid or sad, despite their actual categorization being angry. These inaccuracies align with the observations we make when examining images in documents.

Expression	Shallow Model	Deep Model
Angry	41%	53%
Disgust	32%	70%
Fear	54%	46%
Happy	75%	80.5%
Sad	32%	63%
Surprise	67.5%	62.5%
Neutral	39.9%	51.5%

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

the dataset; even for a human being, it might be It is difficult to ascertain if an angry expression really indicates grief or rage. This phenomenon arises due to the inherent variability in people' modes of emotional expression.

Furthermore, the accuracy of each model was computed for every statement. In addition to the confusion matrices, this action was undertaken. The results of this study are shown in Table 3. The table demonstrates that the accuracy of forecasting a happy expression is the highest among all emotions. This assertion is applicable to both superficial and profound facial expression models. Deep neural

networks have led to improved classification accuracy for most expressions. Deeper networks not only did not enhance accuracy, but they actually led to a decrease in the accuracy of predictions for some emotions, such as surprise and horror. In other words, it suggests that delving deeper does not always lead to enhanced attributes for some expressions.

A learning model was developed whereby the HOG features were combined with the features provided by the convolutional layers. These combined features were then used as input features for the FC layers. The purpose of this study was to investigate the impact of including different characteristics into our Convolutional Neural Network (CNN) model, as mentioned in the preceding section. This technique was used to train both a shallow network and a deep network. Figure 7 and Figure 8 show the accuracy attained throughout several iterations for the shallow and deep models, respectively. Based on these data, it is evident that the model's accuracy closely resembles the accuracy achieved by the model without any HOG features. This suggests that CNN has the capability to effectively extract a substantial amount of information, including that derived from HOG features, by employing just raw pixel data. The activation maps of the several layers were viewed during the forward pass to see the characteristics extracted by our trained network at each layer. This facilitated the observation of the characteristics extracted by our network. The visualization is shown in Figure 9. It is evident that as the training advances, the activation maps exhibit an increasing degree of sparsity and localization.

In addition, we displayed the weights of the first layer to assess the proficiency of the trained network. Figure 10 demonstrates the presence of smooth filters devoid of any noisy pattern. This suggests that our network has had a sufficient training period and the regularization strength is enough.

Furthermore, we used the Deep Dream technique [16, 17] on our most advanced prediction model to identify improved patterns in our photos. The example for each statement, together with its corresponding Deep Dream output, is shown in Figure 11.

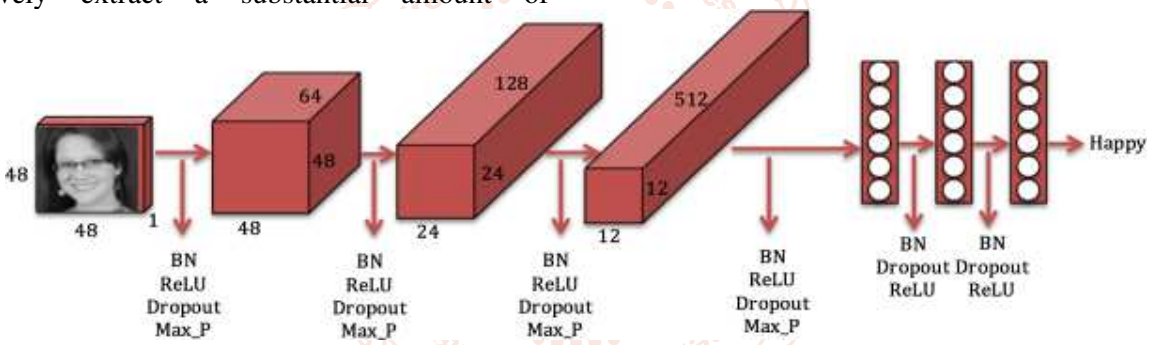

Figure 2: The architecture of the deep network: 4 convolutional layers and 2 fully connected layers

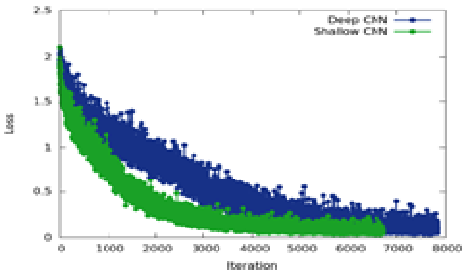

Figure 3: The loss history of the shallow and deep models

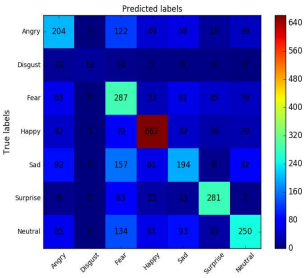

Figure 4: The confusion matrix for the shallow model

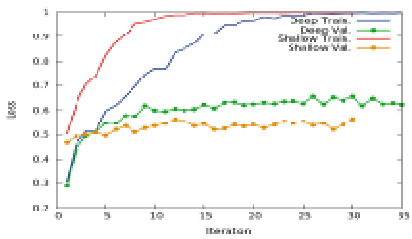

Figure 5: The accuracy of the shallow and deep models fordifferent numbers of iterations

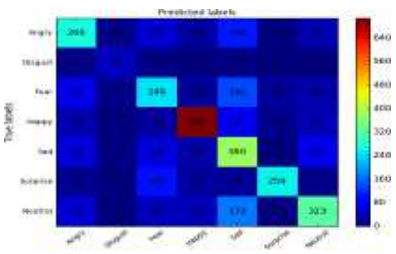

Figure 6: The confusion matrix for the deep model

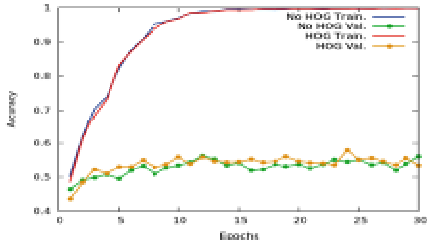

Figure 7: The accuracy of the shallow model with hybridfeatures for different numbers of iterations

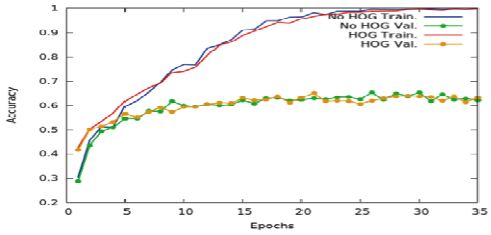

Figure 8: The accuracy of the deep model with hybridfeatures for different numbers of iterations

6. Summary

6.1. Conclusion

To address the problem of facial emotion identification, we developed many Convolutional Neural Networks (CNNs) and evaluated their effectiveness using various post-processing and visualization methods. The outcomes of the study demonstrate that deep convolutional neural networks (CNNs) has the ability to acquire knowledge about facial characteristics and enhance the accuracy of facial emotion identification. Furthermore, it was observed that the incorporation of hybrid feature sets did not provide any significant improvement in the accuracy of the model. This finding suggests that convolutional networks had the capability to acquire the fundamental facial characteristics alone via the use of raw pixel input.

6.2. Future Work

In this project, the models were trained from the ground up using Convolutional Neural Network (CNN) packages in Torch. In future endeavors, we want to expand our methodology to include color photos. By doing this study, we will be able to examine the effectiveness of pre-trained models like AlexNet[18] or VGGNet [19] in recognizing facial emotions. An further expansion might include the incorporation of a facial detection procedure, then followed by the prediction of emotions.

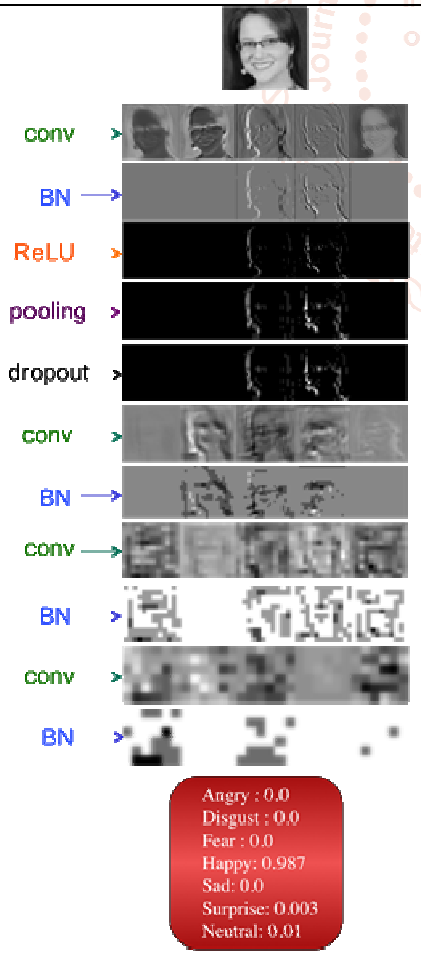

Figure 9: Visualization of the activation maps for differentlayers in our CNN

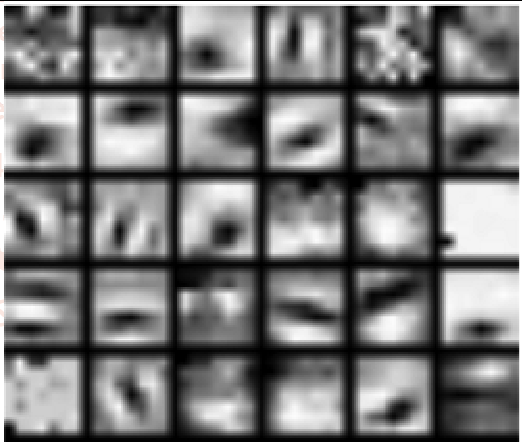

Figure 10: Visualization of the weights for the first layer inour CNN

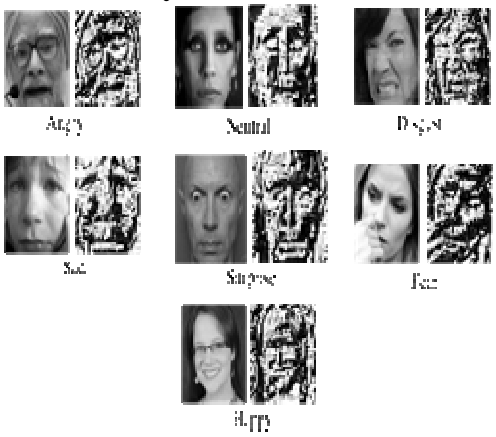

Figure 11: Examples of applying DeepDream on our dataset

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. arXivpreprint arXiv:1203.6722

[5] Lonare, Ashish, and Shweta V. Jain (2013). A Sur- vey 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 VisionComputing 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. Lain-scsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In Proceedings of the IEEE Conference on Auto- matic 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 ex- pressions. Journal of Multimedia, 2006.

[12] P. Ekman, W. Friesen, Facial Action Coding Sys-tem: 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 im-ages 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 im- age recognition.” arXiv preprint arXiv:1409.1556(2014).