IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 16, May 2023

Facial Expression Detection using Machine Learning

Akhilesh Dilip Surve, Vivekanand Sunil Singnapure, Joslyn Sheeril Manuel, Abhishek Sharad Waghchaure
Department of Computer Engineering
Sinhgad Institute of Technology, Lonavala, Maharashtra, India.

Abstract: *In this project, our aim was to create convolutional neural networks (CNN) specifically designed for recognizing facial expressions. The objective was to categorize each facial image into one of seven emotion types considered in our study. We trained CNN models with varying levels of complexity using grayscale images obtained from the Kaggle platform. Our implementation was based on the Torch framework, and we leveraged the power of Graphics Processing Units (GPUs) to accelerate the training process. Alongside the networks' ability to process raw pixel data, we adopted a novel approach that combined this information with Histogram of Oriented Gradients (HOG) features. This hybrid feature strategy was employed to enhance the models' performance.*

To tackle the problem of overfitting, we employed various techniques, including dropout, batch normalization, and L2 regularization. Additionally, we utilized cross-validation to identify the optimal hyperparameters for our models. To evaluate the effectiveness of our models, we closely examined their training histories and performance metrics.

Furthermore, we explored the visualization of different layers within the network to gain insights into the facial features learned by the CNN models.

In summary, our project focused on developing specialized CNN models for facial expression recognition. We experimented with different model complexities, feature strategies, regularization methods, and visualization techniques to achieve accurate classification of facial emotions.

Keywords: Facial Expression

I. INTRODUCTION

Communication between humans primarily relies on verbal speech, but it also involves nonverbal cues such as body language and facial expressions to convey emotions and emphasize certain aspects of communication.

Facial expressions, in particular, play a crucial role in interpersonal relations, transmitting nonverbal signals that contribute to the overall message. While humans effortlessly and quickly recognize facial expressions, achieving reliable automatic recognition by machines remains a challenging task. Despite recent advancements in face detection, feature extraction, and expression classification techniques, developing an automated system capable of accurately recognizing facial expressions remains difficult.

This paper presents an approach that utilizes Convolutional Neural Networks (CNN) for facial expression recognition. Our system takes an image as input and employs CNN to predict the corresponding facial expression label, which includes emotions such as anger, happiness, fear, sadness, disgust, and neutral.

II. RELATED WORK

In recent times, notable advancements have been made by researchers in the field of automatic expression classifiers [7, 8, 9]. Certain systems for expression recognition categorize facial expressions into prototypical emotions like happiness, sadness, and anger [10]. Others focus on identifying specific muscle movements in the face [11] to provide an objective description of facial expressions. The Facial Action Coding System (FACS) [12] is widely recognized as the primary psychological framework for comprehensively describing facial movements. FACS employs Action Units (AU) to classify distinct elements of visible facial movement or associated deformations. Expressions typically arise from the combination of multiple AUs [7, 8].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 16, May 2023

Furthermore, there have been notable developments in the techniques employed for facial expression recognition, including Bayesian Networks, Neural Networks, and the multilevel Hidden Markov Model (HMM) [13, 14]. However, some of these techniques suffer from limitations in terms of recognition rates or timing. Achieving accurate recognition often requires the combination of two or more techniques, with appropriate feature extraction as necessary. The success of each technique heavily relies on image pre-processing to address challenges such as illumination and feature extraction.

Methods:

To access the performance of our models in facial expression recognition, we developed convolutional neural networks (CNNs) with varying depths. The network architecture we employed in our investigation was as follows:

[Conv-(SBN)-ReLU-(Dropout)-(Max-pool)]M - [Affine-(BN)-ReLU-(Dropout)]N - Affine - Softmax.

The network structure consists of M convolutional layers, which can include spatial batch normalization (SBN), dropout, and max-pooling in addition to the convolution layer and ReLU nonlinearity that are always present. Following the M convolution layers, the network proceeds to N fully connected layers, which always involve an affine operation and ReLU nonlinearity. These layers can also incorporate batch normalization (BN) and dropout. Finally, the network concludes with an affine layer for computing scores and applying the softmax loss function.

Our developed model offers flexibility to the user in determining the number of convolutional and fully connected layers, as well as the inclusion of batch normalization, dropout, and max-pooling layers. In addition to dropout and batch normalization techniques, we implemented L2 regularization in our implementation. The user can specify the number of filters, strides, and zero-padding, or default values will be considered if not provided.

As described in the following section, we proposed a novel approach that combines Histogram of Oriented Gradients (HOG) features with those extracted by the convolutional layers using raw pixel data. For this purpose, we utilized the same architecture mentioned earlier, but with the addition of HOG features to the output of the last convolutional layer. The hybrid feature set is then fed into the fully connected layers for score computation and loss calculation.

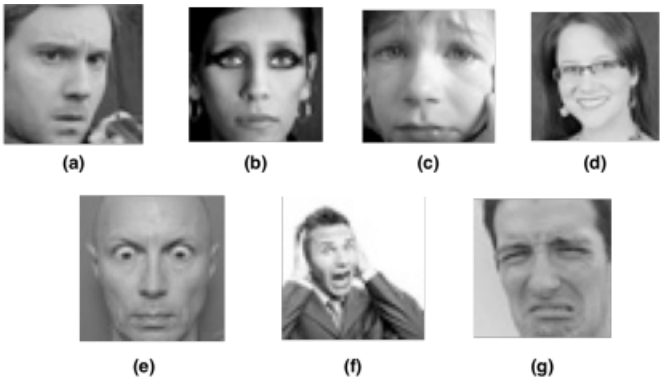

Figure 1 illustrates examples of the seven facial emotions considered in our classification problem, namely angry, neutral, sad, happy, surprise, fear, and disgust.

Our model implementation was carried out using Torch, taking advantage of the GPU-accelerated deep learning capabilities to expedite the training process.

Dataset and Features:

For this project, we utilized a dataset obtained from the Kaggle website. The dataset comprises approximately 37,000 grayscale images of faces, each with a size of 48 × 48 pixels. These images were meticulously processed to ensure that the faces are well-centered and occupy a consistent amount of space in each image. The primary task was to categorize each image into one of seven classes, representing different facial emotions. The assigned numerical labels for these emotions are as follows: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, and 6=Neutral. Figure 1 provides an illustrative example for each facial expression category.

In addition to the class number, which corresponds to a value between 0 and 6, the dataset is divided into three distinct sets: training, validation, and test sets. The training set consists of around 29,000 images, while the validation and test sets contain approximately 4,000 images each.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 16, May 2023

To pre-process the images, we normalized the raw pixel data by subtracting the mean of the training images from each individual image, including those in the validation and test sets. To enhance the dataset through data augmentation, we generated mirrored images by horizontally flipping the images in the training set.

In terms of feature extraction, our primary focus was on utilizing the features generated by the convolutional layers using the raw pixel data. Additionally, as an exploratory approach, we developed learning models that incorporated Histogram of Oriented Gradients (HOG) features by concatenating them with the features generated by the convolutional layers. These combined features were then used as inputs for the Fully Connected (FC) layers.

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

Table 1 presents the hyperparameters that were obtained through cross-validation for the shallow model.

III. ANALYSIS

Experiments

In this project, we initially constructed a shallow Convolutional Neural Network (CNN) consisting of two convolutional layers and one Fully Connected (FC) layer. The first convolutional layer incorporated 32 filters of size 3×3, accompanied by batch normalization, dropout, and ReLU activation. In the second convolutional layer, we employed 64 filters of size 3×3, along with batch normalization, dropout, max-pooling using a filter size of 2×2, and ReLU activation. The FC layer contained a hidden layer with 512 neurons and utilized Softmax as the loss function. The entire network employed ReLU activation in all layers. Prior to training, we conducted sanity checks to verify the correct implementation of the network. These checks involved computing the initial loss without regularization and attempting to overfit the model using a small subset of the training set. The shallow model passed both checks successfully.

Subsequently, we trained the shallow model from scratch using all the images in the training set, leveraging GPU acceleration for faster model training in Torch. The training process involved 30 epochs with a batch size of 128. To determine the optimal hyperparameters, we performed cross-validation experiments, varying the values of regularization, learning rate, and the number of hidden neurons. Validation was conducted using the validation set, while the test set was used to evaluate the model's performance. The best shallow model achieved an accuracy of 55% on the validation set and 54% on the test set. Table [1] provides a summary of the hyperparameters obtained through cross-validation for the shallow model.

To assess the impact of adding convolutional and FC layers to the network, we proceeded to train a deeper CNN with four convolutional layers and two FC layers. The first convolutional layer consisted of 64 filters of size 3×3, followed by a second layer with 128 filters of size 5×5, a third layer with 512 filters of size 3×3, and a final layer with 512 filters of size 3×3. All convolutional layers employed a stride of size 1, batch normalization, dropout, max-pooling, and ReLU activation. The first FC layer had 256 neurons, while the second FC layer had 512 neurons. Similar to the convolutional layers, both FC layers incorporated batch normalization, dropout, and ReLU activation. Softmax was employed as the loss function. The architecture of this deep network is depicted in Figure 2. Prior to training, we performed sanity checks to ensure correct network implementation, including initial loss computation and overfitting tests using a small training subset. The results of these checks confirmed the accurate implementation of the network.

We then trained the deep network using 35 epochs and a batch size of 128, utilizing all the images in the training set. Cross-validation was conducted to identify the hyperparameters that yielded the highest accuracy. The deep model achieved an accuracy of 65% on the validation set and 64% on the test set.

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

Table [2] presents the values of each hyperparameter for the model with the highest accuracy.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 16, May 2023

Further experiments involved training networks with 5 and 6 convolutional layers, but they did not result in increased classification accuracy. As a result, we concluded that the model with 4 convolutional layers and 2 FC layers was the most effective for our dataset.

In both the shallow and deep models, we solely utilized the features generated by the convolutional layers using raw pixel data as the primary features for the classification task. However, HOG features are commonly used for facial expression recognition due to their sensitivity to edges. We aimed to explore the potential of combining HOG features with raw pixels in our network and assess the model's performance when incorporating.

IV. RESULTS

To evaluate and compare the performance of the shallow and deep models, we analyzed the loss history and accuracy achieved by each model. The results are depicted in Figures 3 and 4. Figure 4 illustrates that the deep network significantly improved the validation accuracy by 18.46%. Additionally, the deep network exhibited reduced overfitting tendencies due to the incorporation of non-linearity, hierarchical anti-overfitting techniques such as dropout, batch normalization, and L2 regularization. On the other hand, Figure 3 indicates that the shallow network converged faster, with the training accuracy quickly reaching its peak.

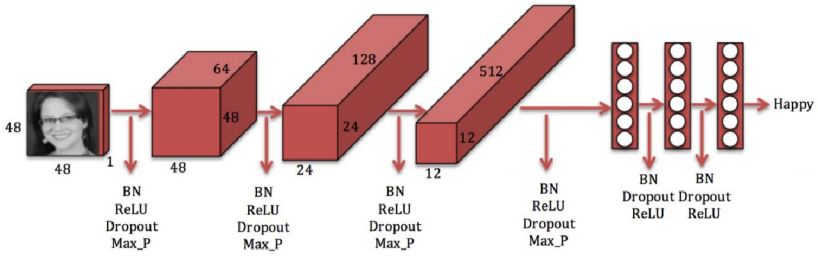

Figure 2: The architecture of the deep network: 4 convolutional layers and 2 fully connected layers

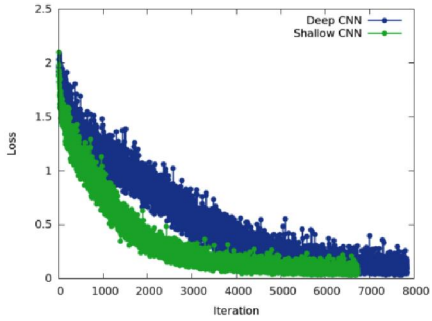

Figure 3: The loss history of the shallow and deep models

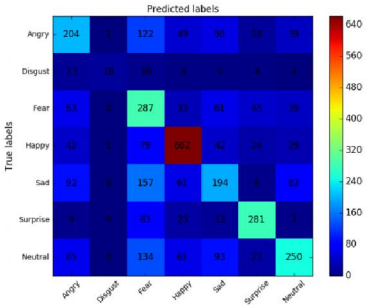

Figure 5: The confusion matrix for the shallow model

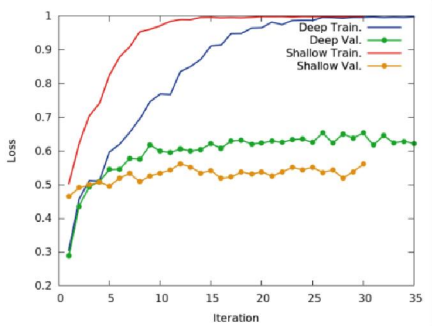

Figure 4: The accuracy of the shallow and deep models for different numbers of iterations

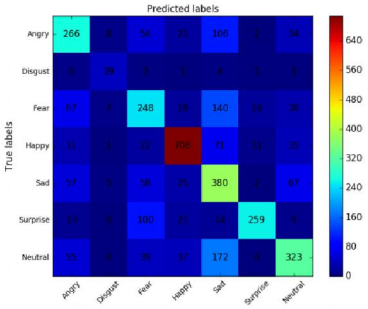

Figure 6: The confusion matrix for the deep model

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 16, May 2023

Confusion matrices were computed for both the shallow and deep networks to examine their predictive capabilities. Figures 5 and 6 visualize these matrices, revealing that the deep network achieved higher correct predictions for most labels. Notably, both models performed well in predicting the "happy" label, indicating that recognizing happy expressions is relatively easier compared to other emotions. The confusion matrices also highlight the labels that are prone to confusion by the trained networks. For instance, there is a correlation between the "angry" label and the "fear" and "sad" labels, with instances misclassified as fear or sadness instead of anger. These misclassifications align with the challenges faced when visually identifying angry versus sad expressions, as individuals express emotions differently.

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 shallow and deep models.

Additionally, the accuracy for each expression was computed for both models, as presented in Table [3]. The results demonstrate that the prediction accuracy for the "happy" expression was the highest among all emotions in both the shallow and deep models. Moreover, the use of deeper networks increased the classification accuracy for most expressions. However, for certain emotions like "surprise" and "fear," deeper networks did not enhance accuracy and, in some cases, even decreased it. This suggests that deeper networks do not necessarily provide improved features for all expressions.

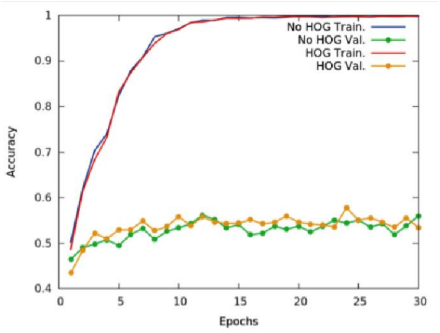

Figure 7: The accuracy of the shallow model with hybrid features for different numbers of iterations

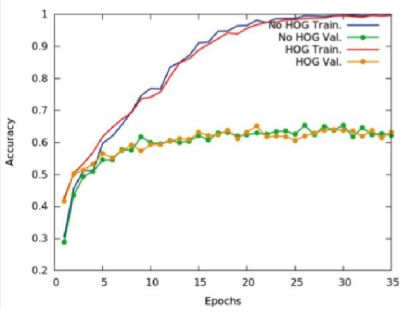

Figure 8: The accuracy of the deep model with hybrid features for different numbers of iterations

To investigate the impact of incorporating different features in our CNN model, we developedlearning models that concatenated HOG features with the features generated by the convolutional layers. Both a shallow and a deep network were trained using this approach. Figures 7 and 8 showcase the accuracy obtained in various iterations for the shallow and deep models, respectively. Remarkably, the accuracy achieved by the models with HOG features was very close to the accuracy of the models without HOG features. This indicates that the CNN is capable of extracting sufficient information, including HOG features, solely from raw pixel data.

To visualize the features extracted by our trained network at each layer, we examined the activation maps during the forward pass. Figure 9 presents this visualization, revealing that as training progresses, the activation maps become more sparse and localized.

Furthermore, we visualized the weights of the first layer to assess the quality of the trained network. Figure 10 illustrates smooth filters without any noisy patterns, indicating that the network was adequately trained with sufficient regularization strength.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 16, May 2023

As an additional analysis, we applied the DeepDream technique to our best predictive model to identify enhanced patterns in the images. Figure 11 displays one example for each expression along with its DeepDream output.

V. SUMMARY

Conclusion

We conducted extensive experiments with different CNN models to address the task of facial expression recognition. Through the utilization of various post-processing and visualization techniques, we evaluated the performance of these models. The findings highlighted the effectiveness of deep CNNs in learning facial characteristics and enhancing the detection of facial emotions. Surprisingly, the inclusion of hybrid feature sets did not contribute to an improvement in model accuracy. This suggests that convolutional networks possess the inherent capability to learn crucial facial features solely from raw pixel data.

Future Work:

In our current project, we focused on training models from scratch using CNN packages in Torch. However, there are several avenues for future work that we would like to explore. Firstly, we plan to expand our model to include color images, enabling us to assess the performance of pre-trained models like AlexNet or VGGNet in facial emotion recognition.

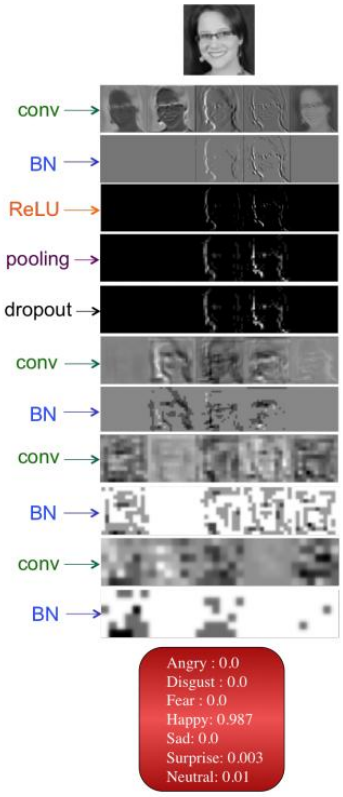

Figure 9: Visualization of the activation maps for different layers in our CNN

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

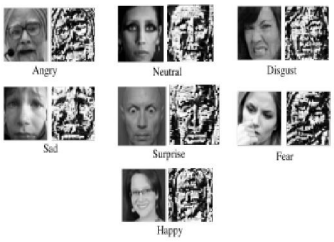

Figure 11: Examples of applying DeepDream on our

Additionally, we aim to incorporate a face detection step prior to emotion prediction, enhancing the overall system's effectiveness. These advancements will contribute to the continued improvement and refinement of our research in the field of facial expression recognition.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 16, May 2023

REFERENCES

[1] Kaggle: Challenges in Representation Learning - Facial Expression Recognition Challenge. Available at: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data
[2] Torch GitHub Repository. Available at: https://github.com/torch
[3] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on (Vol. 1).
[4] Bettadapura, V. (2012). Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722.
[5] Lonare, A., & Jain, S. V. (2013). A Survey on Facial Expression Analysis for Emotion Recognition. International Journal of Advanced Research in Computer and Communication Engineering, 2(12).
[6] Sebe, N., Lew, M. S., Cohen, I., Sun, Y., Gevers, T., & Huang, T. S. (2007). Authentic Facial Expression Analysis. Image and Vision Computing, 25(12), 1856-1863.
[7] Tian, Y., Kanade, T., & Cohn, J. (2001). Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2).
[8] Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2006). Fully automatic facial action recognition in spontaneous behavior. In Proceedings of the IEEE Conference on Automatic Facial and Gesture Recognition.
[9] Bartlett, M. S., Littlewort, G., Fasel, I., Susskind, J., & Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. Image and Vision Computing, 24(6).
[10] Bartlett, M. S., Littlewort, G., Frank, M. G., Lainscsek, C., Fasel, I., & Movellan, J. R. (2006). Automatic recognition of facial actions in spontaneous expressions. Journal of Multimedia.
[11] Ekman, P., & Friesen, W. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press.
[12] Cohen, I., Sebe, N., Garg, A., Chen, L., & Huang, T. S. (2003). Evaluation of expression recognition techniques. In Image and Video Retrieval (pp. 184-195).
[13] Padgett, C., & Cottrell, G. (1996). Representing face images for emotion classification. In Conf. Advances in Neural Information Processing Systems.
[14] scikit-learn: Machine Learning in Python. Available at: http://scikit-learn.org/stable/
[15] Deep Dream Generator. Available at: http://deepdreamgenerator.com/
[16] Google DeepDream GitHub Repository. Available at: https://github.com/google/deepdream
[17] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems.
[18] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556