Acceleration of CNN-Based Facial Emotion Detection Using NVIDIA GPU

Bhakti Sonawane and Priyanka Sharma

Abstract Emotions often mediate and facilitate interactions among human beings and are conveyed by speech, gesture, face, and physiological signal. Facial expression is a form of nonverbal communication. Failure of correct interpretation of emotion may cause for interpersonal and social conflict. Automatic FER is an active research area and has extensive scope in medical field, crime investigation, marketing, etc. Performance of classical machine learning techniques used for emotion detection is not well when applied directly to images, as they do not consider the structure and composition of the image. In order to address the gaps in traditional machine learning techniques, convolutional neural networks (CNNs) which are a deep Learning algorithm are used. This paper comprises of results and analysis of facial expression for seven basic emotion detection using multiscale feature extractors which are CNNs. Maximum accuracy got using one CNN as 96.5% on JAFFE database. Implementation exploited Graphics Processing Unit (GPU) computation in order to expedite the training process of CNN using GeForce 920 M. In future scope, detection of nonbasic expression can be done using CNN and GPU processing.

Keywords CBIR · CNN · GPU processing · Emotion detection

1 Introduction

Facial expression which is a form of nonverbal communication plays an important role in interpersonal relations. Facial expressions represent the changes of facial appearance in reaction to persons inside emotional states, social communications, or intentions. Recently, active research is going on in the area of Automatic Facial

B. Sonawane (✉) · P. Sharma
Nirma University, Ahmedabad, India
e-mail: bhakti.sonawane@sakec.ac.in

P. Sharma
e-mail: priyanka.sharma@nirmauni.ac.in

S. Bhalla et al. (eds.), *Intelligent Computing and Information and Communication*,
Advances in Intelligent Systems and Computing 673,
https://doi.org/10.1007/978-981-10-7245-1_26

Expression Recognition (FER). It is useful in the field of medicine, human emotion analysis, etc., and will be one of the best steps for improving Human Machine Interaction (HMI) systems. Many factors make emotion detection as challenging problem. Among them is the problem of an unavailability of the standardized database for FER. Benchmark database that can fulfill the various requirements of the problem domain in order to become standard database for future research is a tough and challenging exercise [1]. Most of existing databases expressions is posed and not spontaneous. Biggest challenge is to capture spontaneous expressions on images and video. Like different subjects express the same emotions at different intensities and sometimes laboratory conditions become hurdle for the subject to display spontaneous expressions. Another major challenge is labeling of the data which is a time-consuming process and possibly error prone also. Challenges involved in capturing and recognizing spontaneous nonbasic expression are more than basic expressions. Most of the FER has lack of rotational movement freedom [1]. Here, our aim is to present an approach based on CNNs for FER and a systematic comparison of five different 12-layer CNNs. The input to CNN is an image to predict the facial expression label which should be one of these labels: anger, happiness, fear, sadness, disgust, and neutral. Among various database, JAFFE database is chosen for implementation as this is the most commonly used in other automatic FER systems.

2 Background

Generally, face detection, facial feature extraction, and facial expression classification are the parts of an automatic FER system using traditional machine learning techniques. In the face detection step, given an input image system performs some image processing techniques on it in order to locate the face region. In feature extraction step, from located face, geometric features and appearance features are the two types of features that are generally extracted to represent facial expression. Geometric features describe shape of face and its components like lips, nose or mouth corners, etc. Whereas appearance features depict the changes in texture of face when expression is performed. Classification is the last part of the FER system which based on machine learning theory. Output of previous stage which is a set of features retrieved from face region is given as an input to the classifier like Support Vector Machines, K-Nearest Neighbors, Hidden Markov Models, Artificial Neural Networks, Bayesian Networks, or Boosting Techniques [2]. Some expression recognition systems classify the face into a set of standard emotions. Other system aims to find out movements of the individual muscle that the face can produce. In [1], author provided an extensive list of researches between 2001 and 2008 on FER and analysis. Since past few years, there were several advances to perform FER using traditional machine learning methods which involve different techniques of face tracking and detection, feature extraction, training classifier, and classification. In [3], to provide a solution for low resolution images, framework for expression recognition based on appearance features of selected few prominent facial patches which are active when emotion are expressed is proposed. In order to get discriminative features for classification, salient patches are obtained after processing selected patches further. One-against-one classification task is performed using these features and recognition of the expression is done based on majority vote. Experimentation of the proposed method is carried on CK and JAFFE facial expression databases. In [4], feature extraction is done using PCA along with LBP and SVM classifier used to obtain results. Database used are JAFFE database and MUFE database and obtained results show that both PCA and LBP gave high performance together. In [5], live video stream frames are extracted containing face using Gabor feature extraction method and neural network and modified k means with PCA is used for classification of emotion. JAFFE database is used for simulation of framework. In [6] Active Appearance Models (AAM) were used to identify the face and extract its graphic features. For expression prediction, HMM is used and to identify the person in the image, K-NN is used. In [7], improved Directional Ternary Patterns (DTP) feature extraction and SVM classifier are used for real-time purpose emotion detection by facial expressions on JAFFE database.

In FER using traditional machine learning techniques, programmer has to be very specific about what he is interested which involves laborious process of feature extraction. Domain knowledge is expected for feature extraction. Thus, success rate of system depends on programmer's ability to accurately define a feature set. In addition, whenever the problem domain changes the whole system needs to change requiring a redesign of the algorithm from the start [8]. In [9], authors proposed a novel FER system based on features resulting from principal component analysis (PCA) which are fine-tuned by applying particle swarm. The best classification result achieved was 97% for CK database.

Most of researcher also used neural networks as its ability to extract undefined features from the training database. Most of the time it is observed that if neural networks that are trained on large amounts of data are able to extract generalized features well to scenarios that the network has not been trained on. In [10], constructive training algorithm for MLP neural networks has been proposed as classification step for the FER system. Experiments carried on three well-known databases show that the best recognition rate has been obtained using the constructive training algorithm as compared to the fixed MLP architecture. In [11], authors proposed Neural Network and K-NN based model for facial expression classification. For extraction of facial features on JAFEE database, ICA is used. Recent approaches include increased use of deep neural networks (neural networks with many numbers of hidden layers) for automatic FER problem. With growing computing power, for finding complex patterns in images, sound, and text, deep neural network architectures provide learning architecture similar to the development of brain-like structures which can learn multiple levels of representation and abstraction. Extreme variability patterns with robustness to distortions and simple geometric transformations are recognized by CNNs which are deep neural networks. It has been proven by a wide range of applications that are using CNN such as face detection, face recognition, gender recognition, and so forth that minimal domain knowledge of the problem at hand is sufficient to perform efficient pattern recognition tasks [8, 12–16]. CNNs have become the traditional approach for researchers examining vision and deep learning. Starting with LeNet-5 [17], variations of this basic design are prevalent in the image

classification literature with the best results. The recent trend is to increase the number of layers and layer size for larger datasets such as ImageNet and use of dropout in order to deal with the problem of overfitting [18, 19].

In [20], authors proposed network consists of two convolutional layers each followed by max pooling with next four inception layers and conducted experiments on seven publicly available facial expression databases. In [21], two different deep network models are proposed, for extraction temporal appearance features from image sequences and for extraction temporal geometry features from temporal facial landmark points. A new integration method for combining these two models is required in order to boost the performance of the FER. For Emotions in the Wild [22] contest for static images in [23], multiple deep convolutional neural networks are trained as committee members and combine their decisions, generating up to 62% test accuracy.

3 JAFFE Database and Proposed CNN Architecture

Proposed CNN architectures are tested on JAFFE database set of facial expression images for posed emotions (six different emotions and neutral face displays) of 10 Japanese female subjects. These six expressed emotions are the basic emotions given by Ekman and Friesen [24]. Figure 2 shows sample images from JAFFE database [25]. Expressed emotion seems to be universal across human ethnicities and cultures which are happiness, sadness, fear, disgust, surprise, and anger. The grayscale images are 256 × 256 pixels size. The images were labeled into 6 + 1 = 7 emotion classes. Some head pose variations can be featured by these images [26].

Five different 12-Layer CNNs architectures are proposed for facial expression classification up to seven different basic emotions. In all five CNNs, input layer is followed by convolutional layer with different filter size and number of filters. This layer is followed by relu layer and max pooling layer with pool filter size as 2 which outputs maximum among the four values. Max pooling layers are trailed by convolutional layer with different filter sizes and number of filters for different CNNs again followed with relu and max pooling layer. Next layer is a fully connected layer with a number of output neurons which varies in different CNNs and followed by relu layer. And last fully connected layer is with seven output neurons and output of this layer is given to final softmax and classification layer. Table 1 shows detail regarding proposed 5 different CNNs. For CNN_1, CNN_2, CNN_3, and CNN_4, number of training images are 164 and testing images are 26. In CNN_1 and CNN_2, training images are repeated to increase the total number of training image set. For CNN_5, number of training images are 178 and testing images are 35.

Table 1 Details for proposed CNNs for emotion detection

		Convolution	Convolution	Fully connected	Fully connected
Sr. No.	CNNS	Layer 1	Layer 2	Layer 1	Layer 2
01	CNN 1	3(3 × 3)	9(3 × 3 × 3)	512	7
02	CNN 2	3(12 × 12)	6(9 × 9 × 3)	16	7
03	CNN 3	3(12 × 12)	6(9 × 9 × 3)	16	7
04	CNN 4	3(15 × 15)	6(9 × 9 × 3)	16	7
05	CNN 5	3(15 × 15)	6(9 × 9 × 3)	16	7

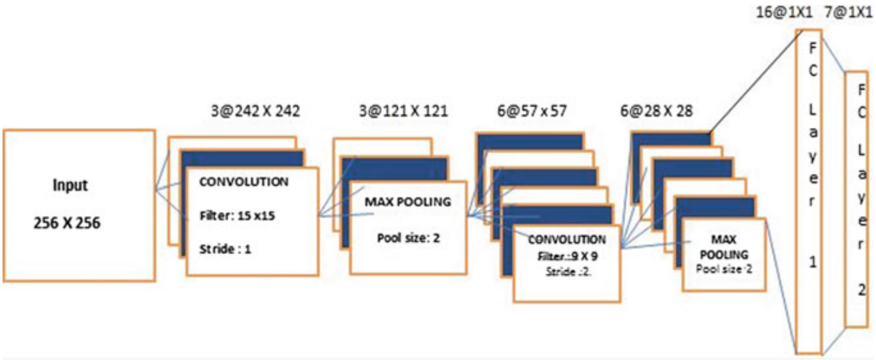

Fig. 1 Proposed CNN_5 architecture

In Table 1, numbers in convolutional layer column give number of feature maps generated with filter size and numbers in fully connected layer column gives number of output neurons. Thus in CNN_1, number of training images is 164 and first convolutional layer generates 3 feature map using 3 × 3 filter size. Second convolutional layer generates 9 feature map using 3 × 3 filter size. Fully connected layer 1 has output neurons 512 and last fully connected layer has number of output neurons as 7 representing basic emotion.

Proposed 12 layers architecture are shown in Fig. 1. Similar architecture is (referring Table 1) for CNN_1 CNN_2, CNN_3, and CNN_4 only with different filter size, feature map, and number of output neurons in fully connected layer.

4 Results and Discussions

The proposed design is tested on a 2.40 GHz Intel i7-5500U quad core processor; 8 GB RAM, with windows 10, 64 bit system. As CNN is computationally intensive which requires GPU processing for faster computation, thus system used was CUDA-enabled NVIDIA GPU with compute capability higher than 3.0, DirectX Runtime Version 12 (graphics card GeForce 920 M). MATLAB used for

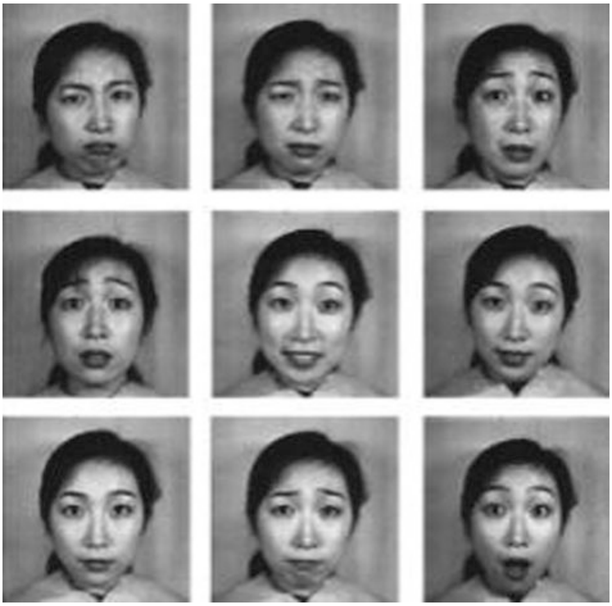

Fig. 2 JAFFE sample images [25]

Table 2 Comparison for CNNs

Sr. No.	CNNS	Accuracy (%)
1	CNN 1	80.76
2	CNN 2	88.46
3	CNN 3	73.07
4	CNN 4	88.46
5	CNN 5	96.15

implementation and system tested on JAFFE database From Tables 1 and 2 it is observed that more image detail gets captured using 12 × 12 or 15 × 15 filters. In CNN_1 and CNN_2, number of training set is increased by repeating that set but its effect is similar to increasing the epoch during training. If first layer captures good detail from the input image of 256 × 256 size and number of epoch is more this leads final accuracy. Maximum accuracy achieved in CNN_5 in which total number of images to be trained is more as compared to other CNNs.

5 Conclusion and Future Scope

Minimal preprocessing involved as CNNs are designed to recognize visual patterns directly from pixels of images. This is completely in contrast with the conventional pattern recognition tasks in which prior knowledge of the problem at hand is needed in order to apply a suitable algorithm to extract the right features. In this paper, five new CNN architectures have been proposed for automatic facial expression recognition. Among them, CNN_5 resulted in the highest accuracy with a classification accuracy of 96.15% achieved on test samples of JAFFE database using 2.5 GHz i7-5500U quad core processor, 8 GB RAM with GeForce 920 M. Proposed architecture can be extended for detection of nonbasic expression in future.

References

1. Face Expression Recognition and Analysis: The State of the Art Vinay Kumar Bettadapura Columbia University.
2. Nazia Perveen, Nazir Ahmad, M. Abdul Qadoos Bilal Khan, Rizwan Khalid, Salman Qadri, Facial expression recognition Through Machine Learning, International Journal of Scientific & Technology Research, Volume 5, Issue 03, March 2016, ISSN 2277-8616.
3. S L Happy, Aurobinda Routray Automatic Facial expression recognition Using Features of Salient Facial Patches, IEEE transactions on Affective Computing, VOL. 6, NO. 1, January–March 2015.
4. Muzammil Abdulrahman, Alaa Eleyan, Facial expression recognition Using Support Vector Machines, in Proceeding of 23nd Signal Processing and Communications Applications Conference, PP. 276–279, May 2015.
5. Debishree Dagar, Abir Hudait, H. K. Tripathy, M. N. Das Automatic Emotion Detection Model from Facial Expression, International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), ISBN No. 978-1-4673-9545-8,2016.
6. G. Ramkumar E. Logashanmugam, An Effectual Facial expression recognition using HMM, International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), ISBN No. 978-1-4673-9545-8,2016.
7. S. Tivatansakul, S. Puangpontip, T. Achalakul, and M. Ohkura, "Emotional healthcare system: Emotion detection by facial expressions using Japanese database," in Proceeding of 6th Computer Science and Electronic Engineering Conference (CEEC), PP. 41–46, Colchester, UK, Sept. 2014.
8. A.R. Syafeeza, M. Khalil-Hani, S.S. Liew, and R. Bakhteri, Convolutional neural network for face recognition with pose and illumination variation, Int. J. of Eng. and Technology, vol. 6 (1), pp. 44–57, 2014.
9. Vedantham Ramachandran, E Srinivasa Reddy, Facial Expression Recognition with enhanced feature extraction using PSO & EBPNN. International Journal of Applied Engineering Research, 11(10):69116915, 2016.
10. Boughrara, Hayet, et al. "Facial expression recognition based on a mlp neural network using constructive training algorithm." Multimedia Tools and Applications 75.2 (2016): 709–731.
11. Hai, Tran Son, and Nguyen Thanh Thuy. "Facial expression classification using artificial neural network and k-nearest neighbor." International Journal of Information Technology and Computer Science (IJITCS) 7.3 (2015): 27.
12. Shih, Frank Y., Chao-Fa Chuang, and Patrick SP Wang. "Performance comparisons of facial expression recognition in JAFFE database". International Journal of Pattern Recognition and Artificial Intelligence 22.03 (2008): 445–459.
13. C. Garcia and M. Delakis, "Convolutional face finder: a neural architecture for fast and robust face detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, pp. 1408–1423, 2004.

14. S. Chopra, R. Hadsell, and Y. LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification," in In Proceedings of CVPR (1) 2005, 2005, pp. 539–546.
15. T. Fok Hing Chi and A. Bouzerdoum, "A Gender Recognition System using Shunting Inhibitory Convolutional Neural Networks," in International Joint Conference on Neural Networks, 2006, pp. 5336–5341.
16. Caifeng Shan, Shaogang Gong, Peter W. McOwanb, Facial expression recognition based on Local Binary Patterns: A comprehensive study, Image and Vision Computing, V. 27 n. 6, pp. 803–816, May 2009. https://doi.org/10.1016/j.imavis.2008.08.005].
17. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural Comput., 1 (4):541551, Dec. 1989.
18. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. arXiv preprint arXiv:1409.0575, 2014.
19. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. arXiv preprint arXiv:1409.4842, 2014.
20. Mollahosseini, Ali, David Chan, and Mohammad H. Mahoor. "Going deeper in Facial expression recognition using deep neural networks." Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016.
21. Jung, Heechul, et al. "Joint fine-tuning in deep neural networks for facial expression recognition." Proceedings of the IEEE International Conference on Computer Vision. 2015.
22. Dhall A, Murthy OVR, Goecke R, Joshi J, Gedeon T (2015) Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, pp 423–426.
23. B. Kim, J. Roh, S. Dong, and S. Lee, Hierarchical committee of deep convolutional neural networks for robust facial expression recognition, Journal on Multimodal User Interfaces, pp. 117, 2016.
24. P. Ekman and W. Friesen. Constants Across Cultures in the Face and Emotion. Journal of Personality and Social Psychology, 17(2):124129, 1971.
25. Michael J. Lyons, Shigeru Akemastu, Miyuki Kamachi, Jiro Gyoba. Coding Facial Expressions with Gabor Wavelets, 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205 (1998).
26. Lyons M., Akamatsu S., Kamachi M., and Gyoba J. Coding Facial Expressions with Gabor Wavelets. In Third IEEE International Conference on Automatic Face and Gesture Recognition, pages 200205, April 1998.