Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 116 (2017) 523–529

Procedia Computer Science

www.elsevier.com/locate/procedia

2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017, Bali, Indonesia

# Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition

Diah Anggraeni Pitaloka[a,*], Ajeng Wulandari[a], T. Basaruddin[a], Dewi Yanti Liliana[a]

[a] *Faculty of Computer Science, Universitas Indonesia, Depok-16424, West Java, Indonesia*

**Abstract**

Emotion recognition from facial expression is the subfield of social signal processing which is applied in wide variety of areas, specifically for human and computer interaction. Many researches have been proposed for automatic emotion recognition, which is fundamentally using machine learning approach. However, recognizing basic emotions such as angry, happy, disgust, fear, sad, and surprise is still becoming a challenging problem in computer vision. Lately, deep learning has gained more attention to solve many real-world problems, including emotion recognition. In this research, we enhanced Convolutional Neural Network method to recognize 6 basic emotions and compared some preprocessing methods to show the influences of its in CNN performance. The compared data preprocessing methods are: resizing, face detection, cropping, adding noises, and data normalization consists of local normalization, global contrast normalization and histogram equalization. Face detection as single pre-processing phase achieved significant result with 86.08 % of accuracy, compared with another pre-processing phase and raw data. However, by combining those techniques can boost performance of CNN and achieved 97.06% of accuracy.


Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.

*Keywords:* Emotion recognition;Facial expression; Convolutional neural network;Machine learning;Computer vision;Normalization;

* Corresponding author. Tel.: +62-852-398-70552 .
*E-mail address:* diah.anggraeni61@ui.ac.id

1877-0509 
Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.
10.1016/j.procs.2017.10.038

## 1. Introduction

Recently, automatic facial expression analysis has caught up researcher attention in Computer Science field. Modelling the interaction between human and computer becomes challenging research topic. According to Maja et. al., in order to represent social facts, social signals can be captured from facial expression in the form of informatic signal [1]. Emotion is a kind of social facts that can be analysed through facial expressions. Basic emotions consist of anger, happy, disgust, fear, sadness, happy, and surprise [2]. The application of automatic facial expression can be applied in wide variety of areas such as emotion and paralinguistic communication, clinical psychology, psychiatry, neurology, pain assessment, lie detection, intelligent environments, and multimodal human computer interface (HCI) [3].

There are two approaches in facial expression analysis, i.e. feature extraction and action unit detection from Facial action coding system (FACS). FACS is a framework proposed by Ekman et al. Feature extraction has two main techniques: geometric feature-based and appearance-based feature extraction [3]. The former represents facial points to form feature vectors and to show the face from geometrical perspective. The latter is applied in the extraction of feature vectors, either in specific or holistic face area of facial image, such as in the Gabor Wavelets, or LBP application.

Deep learning is a part of machine learning approaches that can be utilized as emotion recognition and facial expression analysis. However, its performance depends on the data size. The greater the data the better the performance is [4]. The size of facial expressions datasets are still insufficient for deep learning to be implemented in. Therefore, some researches apply augmentation techniques in the pre-processing step such as cropping, scaling, translation, or mirroring to increase the variance, hence the size of the data. This pre-processing techniques are quite effective to improve the performance of deep learning [5 6 7 8].

This research aims to recognize emotions using deep learning and show the influences of data pre-processing in deep learning performance. Data pre-processing methods including resizing, face detection, cropping, adding noises, and normalizations. We compare the accuracy of each pre-processing methods and the combination between them, then analyse it in order to see the variability of accuracies.

## 2. Related Work

Many works have been proposed in the automatic facial expression analysis. For instance, Shan, et al., who used local binary pattern (LBP) as the appearance-based feature extraction and support vector machine (SVM) as the classifier. They claimed that the results are quite effective and efficient for emotion recognition [9]. They made comparisons between different types of image resolution and showed a stable result although the images resolution were low. Dewi et al. proposed Active Appearance Model (AAM) and Fuzzy C-Means to recognize emotions: happy, sad, anger, fear, disgust, contempt, surprise, and neutral [10]. AAM is a template-matching-based feature extraction and is used in training phase. There are 68 points were selected for shape analysis. While in the inference, Fuzzy C-Means is used in which the input belongs to a specific emotion. They performed experiment using CK+ dataset and gave 80.71% of the experimental result.

Besides of those methods, Lopes et al. developed facial expression recognition with Convolution Neural Networks (CNN) [8]. They apply some pre-processing steps to tackle the limitation of facial expression datasets for data augmentation. This research achieves competitive results: 96.75% in CK+. The experiments employed CK+, JAFFE and BU-3DFE which commonly used for facial expression competition and benchmark problem. Jeon, et al. also proposed CNN to perform real-time facial expression recognition and achieved 70.74% of accuracy using Kaggle database for happy and surprise [11]. Another research related real-time facial expression recognition was developed by Duncan, et al [12]. they proposed transfer learning from VGG_s and achieved accuracy 90.0% of training steps and 57% for testing. They utilized non-standardized homemade dataset for this case. Outlett, et al. used CNN as feature extraction to perform face recognition for video game and replaced the classifier with SVM [13]. The experimental result gave 94.4% accuracy. However, Mayya, et al. also implemented the same method and achieved higher accuracy of 98.08% for CK+ and 98.12% for JAFFE [14].

## 3. Proposed Method

In this work, we propose feature extraction and classification of 6 basic emotions: angry, happy, disgust, fear, sad, and surprise based on convolution neural network (CNN). We applied pre-processing steps such as face detection, cropping, resize, adding noise, and data normalization. About the details explanation were described below.

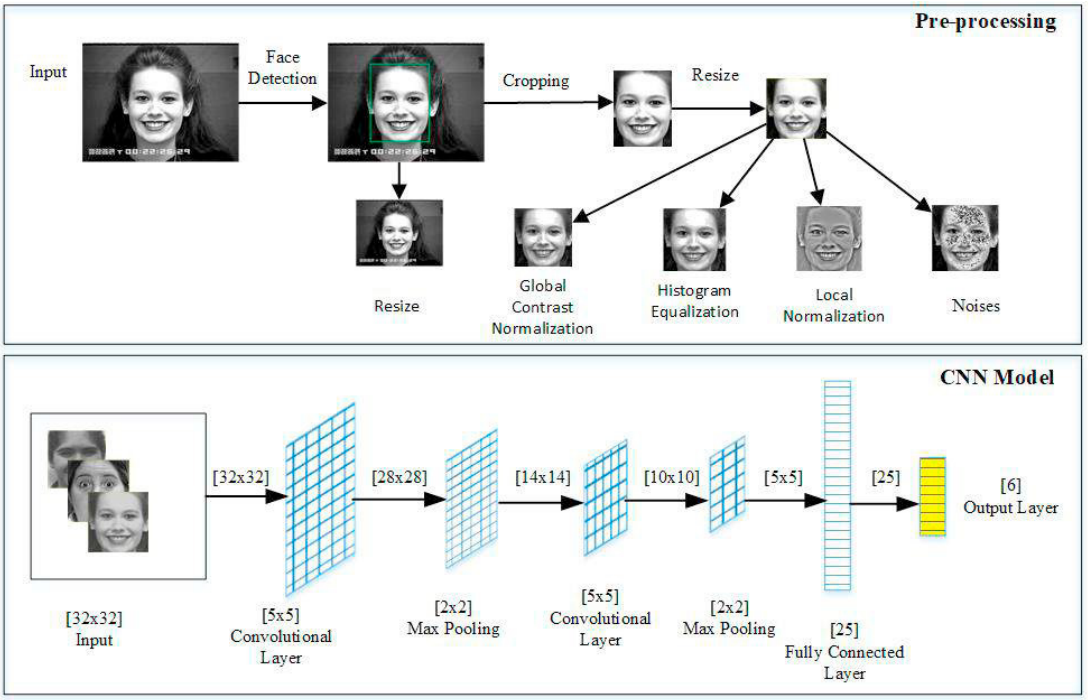


Fig. 1. Research method

### 3.1 Datasets

The experiments were performed using posed dataset JAFFE, CK+, and MUG which mean every subject was given instructions by expert to demonstrate 6 emotions. CK+ consist of 123 participants were 18-50 years old, 69% female, 81% Euro-American, 13% Afro-American, and 6% other groups [15]. It is widely used for emotion recognition and facial expression analysis. Furthermore, The Japanese Female Facial Expression (JAFFE) has 213 images that represent 6 basic emotions from 10 Japanese females [16]. The last is MUG dataset [17]. There are 86 subjects of Caucasian origin and ages between 20 and 35 years, 35 females and 51 males. The following data are available in 896 x 896 pixels and were categorized by 6 types of emotions.

Table 1. The differences between CK+ and JAFFE datasets

| | CK+ | JAFFE | MUG |
|---|---|---|---|
| Label | 6 emotions | 6 emotions | 6 emotions |
| The Number | 593 (327 labeled, 266 unlabeled) | 213 | 281 |
| Participants | 123 (male and female) | 10 (Japanese female) | 86 (male and female) |
| Resolution | 640x490 or 640x480 | 256x256 | 896x896 |
| Format | .png | .tiff | .jpg |

### 3.2 Pre-processing

There are 4 methods for pre-processing step: face detection & cropping, resize, adding noise, and normalizations. Face detection methods in this work using Haar algorithm. It aims to remove background and non-face areas, then crop the face area. Next pre-processing is down-sampling (resize) the resolution to the image to be 32x32, 64x64,128x128. We want to examine the performance of CNN architecture againsts the input size. The fourth is adding noise like Salt & Pepper, and Speckle as data augmentation. We want to show that CNN performance still good and stable despite of given noises. And the last is normalizations which consist of global contrast normalization (GCN), local normalization, and histogram equalization.

GCN performs subtracting each pixel value of image by the mean and divide it by standard deviation. It aims to prevent images from having varying amounts of contrast. Images with very low but non-zero contrast often contains less information and it becomes a problem for facial expression recognition. Dividing by true standard deviation usually accomplishes nothing than amplifying sensor noise. Goodfellow et al. introduced positive regularization parameter $\lambda$ to bias the estimate of standard deviation. Alternately, we can constraint the denominator to be at least $\varepsilon$ [18]. GCN can be defined as:

$$X'_{i,j,k} = s\frac{X_{i,j,k} - \bar{X}}{\max\left\{\varepsilon, \sqrt{\lambda + \frac{1}{3rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{3}(X_{i,j,k} - \bar{X})^2}\right\}} \quad (1)$$

Where $X_{i,j,k}$ has row $i$, column $j$, and colour depth $k$, and $\bar{X}$ is the mean intensity of the entire image.

The second is local normalization. It is implemented by using separable convolution to compute local means and local standard deviations, then using element-wise subtraction and element-wise division on different feature maps [18]. And the last is histogram equalization which is doing manipulation of intensity for image enhancement [19]. It uses cumulative density function of the image then changes brightness of an image by flattening the histogram and stretching the image contrast to be distributed over all grey levels [20].

### 3.3 Convolution Neural Networks

Facial expression recognition using CNN is performed by fed the image from pre-processing step to first convolution layer with a kernel size 5 x 5 and stride 1. It aims to extract features like edges, oriented-edges, corners, and shapes [18]. Consider an input image size 32 x 32 and applied 64 number of filters, thus the output of first convolution layer is 64 of feature-maps size 28 x 28. The 28 x 28 output image is passed to max pooling layer of 2 x 2 kernel size with stride 2 for each dimension and reduce the image to size 14 x 14. Max pooling tried to find the underlying features of an image in each dimension [18]. The results of this previous step were convoluted by second convolution layer with a kernel size 5 x 5 and stride 1 and gained 10 x 10 pixels of output images. We also apply max pooling in fourth layer, reduce the image to size 10 x 10, and it is followed by fully connected layer that has 25 neurons. Finally, the output of this layer is connected to the output layer that has 6 output nodes which represent 6 basic emotions. The activation function for all layers using Rectified linier unit (ReLu) which aims to preserve the properties of each output [18]. ReLu is defined as maximum value that greater than or equal to zero: $g(z) = \max\{0, z\}$. Training of CNN model is done by using RMSProp. RMSProp was modified from AdaGrad to improve the performance especially for non-convex setting, and change the gradient into exponential. Cross-entropy is used for error function.

## 4. Experiments and Results

Pre-processing step and CNN were performed using python libraries such OpenCV and TensorFlow on GPU NVIDIA version 375.74 from nvidia-375, processor type 16xi7-5960X, memory 65 GB, and Ubuntu 16.04 as operating system. The scenario of this research is performed using CK+, JAFFE, and MUG dataset. We split the dataset into 90:10 for learning and testing. We collected single image of the subject for each emotion in all dataset

(CK+, JAFFE, and MUG) which represents peak emotion and got 852 number of frames (766 for training and 86 for testing). There are 6 pre-processing steps to enhance CNN performances: (a) raw data or without pre-processing step, (b) face detection and cropping (c) global contrast normalization, (d) local normalization, (e) histogram equalization, (f) adding noises, and (g) b + f. All the pre-processing steps except (g) step using unique subject so the partition is quite fair because there is no overlapping subject with same emotion belongs to training set and validation set. While the (g) step contains data from (b) step and (f) step as augmentation data which increase the probability of overlapping data. We shuffle the data to solve ordered index of datasets. It intends to make sure CNN will learn all emotion for each dataset and recognize new data based on learning process.

Table 2 shows the mean of accuracy for each class. We highlighted the best accuracies for each class of emotion and the average of pre-processing step. Region of interest extraction (b) can improve CNN performance up to +24.27%. It is the baseline for further pre-processing techniques. The best normalization techniques for this case is led by histogram equalization since it can reduce the data variance by adjust the image contrast to be balance and to be better distributed over all intensity values. Pre-processing steps b+f achieved the best accuracy about 93.14%. We used data from (f) as data augmentation. But for only step (f) or adding noise, it caused the higher error rate for emotion recognition. The most discriminable emotion that can be recognized by CNN are anger, happy, and surprise. CNN gained 100% of accuracy for those emotions. Sadness is the most difficult emotion to recognize since many missclassified of sadness into anger and fear. it means, CNN is hard to distinguish sadness due sadness looks similar to anger and fear.

Table 2. The influences of accuracy in pre-processing stage for each class of all dataset using 32x32 pixels.

| Pre-processing Step | Anger | Disgust | Fear | Happy | Sadness | Surprise | Average |
|---|---|---|---|---|---|---|---|
| **(a)** | 90.91% | 50.00% | 45.45% | 68.42% | 28.57% | 87.50% | 61.81% |
| **(b)** | **100.00%** | 85.71% | 72.73% | **100.00%** | 64.29% | 93.75% | 86.08% |
| **(c)** | 90.91% | 85.71% | **81.82%** | **100.00%** | 71.43% | **100.00%** | 88.31% |
| **(d)** | **100.00%** | 92.86% | **81.82%** | **100.00%** | 57.14% | 93.75% | 87.60% |
| **(e)** | 90.91% | 85.71% | 72.73% | **100.00%** | 85.71% | **100.00%** | 89.18% |
| **(f)** | 81.82% | 64.29% | **81.82%** | **100.00%** | 64.29% | **100.00%** | 82.04% |
| **(g)** | 88.24% | **97.06%** | 76.67% | **100.00%** | **96.88%** | **100.00%** | **93.14%** |

The accuracy of testing phase is shown in Table 3. Performance of CNN increased and gave significant results in face detection and cropping phase. CNN can directly learn from region of interest because all background has been remove during pre-processing phase. Adding noise as augmentation techniques showed quite good impact although not as good as cropping phase. However, by combining result from (b) and (f) as data augmentation technique, it can boost performance of CNN and accuracy. Global contrast normalization reaches the best accuracy over all normalization techniques: (d) local normalization and (e) histogram equalization. GCN is used by Gudi et al. for action unit estimation and Goodfellow on the Cifar 10 dataset [21] to improve the accuracy by reduce the difference magnitude between high contrast and low contrast. This is only 0.78% higher than the step (e) or histogram equalization. The visually effects of these technique are depicted in Figure 1 which are not too different than the original image. While the transformations of step (d) or local normalization is much more differ than the original one since it modifies each pixel using small window and involve local neighbourhood rather than modifies entire pixels over the image. Perhaps it caused information loss that represents underlying structure of emotions.

Pre-processing step (g) which contains data from b and f achieved the best accuracy over all steps. However, the chance of the data to be overlap between training set and testing set increased since there are 2 version of the data: original and noisy data. This stage lead to some subject of the data for specific emotion either original or noisy data will belong to training and testing set. Finally, the impact of resolution on CNN performance. Model capacity of proposed CNN model works better with 32x32 or 64x64 resolution but not good enough for size 128x128 only 79.83%. It needs higher capacity to solve more complex task especially for large amounts of data or resolution.

Table 3. Training and Testing Results for all dataset

| Pre-processing Step | Resolution of Testing Phase | | | Average |
|---|---|---|---|---|
| | 32x32 | 64x64 | 128x128 | |
| **(a)** | **62.35%** | 56.47% | 51.76% | 56.86% |
| **(b)** | 87.06% | **89.41%** | 82.35% | 86.27% |
| **(c)** | **89.41%** | **89.41%** | 84.71% | **87.84%** |
| **(d)** | **88.24%** | 84.71% | 82.35% | 85.10% |
| **(e)** | **90.59%** | 89.41% | 81.18% | 87.06% |
| **(f)** | **83.53%** | 82.35% | 81.18% | 82.35% |
| **(g)** | 95.29% | **97.06%** | 95.29% | **95.88%** |
| **Average** | **85.21%** | 84.12% | 79.83% | |

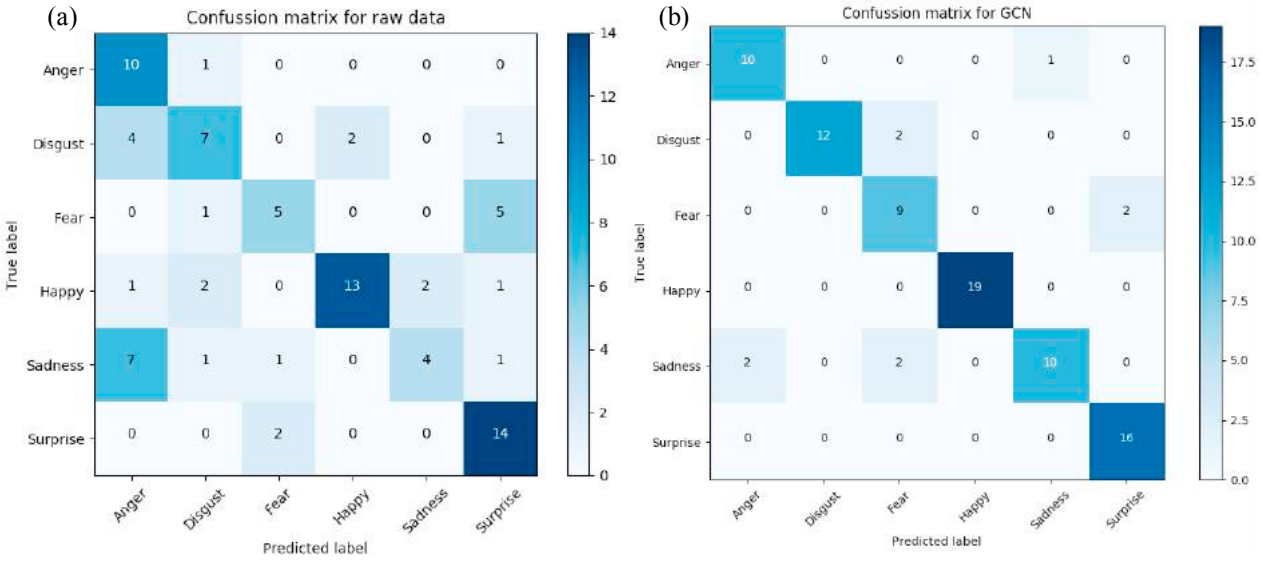


Fig. 2. Confusion matrix for six emotions of all dataset: (a) Without pre-processing step; (b) With pre-processing steps (face detection, cropping, and global contrast normalization)

## 5. Conclusion

In this paper, we have successfully implemented the proposed method to make comparison between pre-processing methods for facial expression recognition. Based on the experimental results obtained, explain that face detection and cropping to get region of interest (ROI) achieved the best improvement for the CNN performance. The global contrast normalization (GCN) step contributes better improvement than another normalization techniques to accuracy but not as good as getting the ROI. GCN tried to reduce variance of the data and prevent from having varying amounts of contrast. The proposed CNN model works better on 32x32 and 64x64 resolution. It seems the capacity of model satisfies the complexity task for facial expression recognition on those resolution. We can boost the performance of CNN using data augmentation like combining data from step (b) cropping and (f) adding noises. The feature work

involves exploring image synthesis techniques that may be considered as solution of augmentation data in deep learning. It aims to prevent data starvation and overfitting for small amount of data.

## Acknowledgements

This research was supported by Publikasi International Terindeks (PITTA) of Universitas Indonesia. The first author would like to thank of this grants for graduate school students.

## References

1. Pantic M, Cowie R, DÉrrico F, Heylen D, Mehu M, Pelachaud C, et al. Social Signal Processing: The Research Agenda. In Moeslund TB, Hilton A, Krüger V, Sigal L, editors. Visual Analysis of Humans: Looking at People. London: Springer London; 2011. p. 511-538.
2. Ekman P. An argument for basic emotions. Cognition and Emotion. 1992;: p. 169-200.
3. Tian Y, Kanade T, Cohn JF. Facial Expression Recognition. In Li SZ, Jain AK, editors. Handbook of Face Recognition. London: Springer London; 2011. p. 487-519.
4. Chen XW, Lin X. Big Data Deep Learning: Challenges and Perspectives. IEEE Access. 2014; 2: p. 514-525.
5. Gudi A, Tasli HE, den Uyl TM, Maroulis A. Deep learning based FACS Action Unit occurrence and intensity estimation. In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG); 2015 May. p. 1-5.
6. Khorrami P, Paine TL, Huang TS. Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition? CoRR. 2015; abs/1510.02969.
7. Mollahosseini A, Chan D, Mahoor MH. Going deeper in facial expression recognition using deep neural networks. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV); 2016 March. p. 1-10.
8. Lopes AT, de Aguiar E, Souza AFD, Oliveira-Santos T. Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. Pattern Recognition. 2017; 61: p. 610-628.
9. Shan C, Gong S, McOwan PW. Facial expression recognition based on Local Binary Patterns: A comprehensive study. Image and Vision Computing. 2009; 27: p. 803-816.
10. Liliana DY, Widyanto MR, Basaruddin T. Human emotion recognition based on active appearance model and semi-supervised fuzzy C-means. In 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS); 2016 Oct. p. 439-445.
11. Jeon J, Park JC, Jo Y, Nam C, Bae KH, Hwang Y, et al. A Real-time Facial Expression Recognizer Using Deep Neural Network. In Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication; 2016; New York, NY, USA: ACM. p. 94:1--94:4.
12. Duncan D, Shine G, English C. Facial Emotion Recognition in Real Time. 2016.
13. Ouellet S. Real-time emotion recognition for gaming using deep convolutional network features. CoRR. 2014; abs/1408.3750.
14. Mayya V, Pai RM, Pai MMM. Automatic Facial Expression Recognition Using DCNN. Procedia Computer Science. 2016; 93: p. 453-461.
15. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops; 2010 June. p. 94-101.
16. Lyons M, Akamatsu S, Kamachi M, Gyoba J. Coding facial expressions with Gabor wavelets. In Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition; 1998 Apr. p. 200-205.
17. Aifanti N, Papachristou C, Delopoulos A. The MUG facial expression database. In 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10; 2010 April. p. 1-4.
18. Goodfellow I, Bengio Y, Courville A. Deep Learning: MIT Press; 2016.
19. Gonzales RC, Woods RE. Digital Image Processing Third Edition New Jersey: Pearson; 2010.
20. Vidyasaraswathi HN, Hanumantharaju MC. Review of Various Histogram Based Medical Image Enhancement Techniques. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering \&\#38; Technology (ICARCSET 2015); 2015; New York, NY, USA: ACM. p. 48:1--48:6.
21. Goodfellow IJ, Warde-farley D, Mirza M, Courville A, Bengio Y. Maxout networks. In In ICML; 2013.
22. Sage D. Biomedical Imaging Group. [Online].; 2002 [cited 2017 08 10. Available from: http://bigwww.epfl.ch/sage/soft/localnormalization/.