Automatic Caption Generation from Images Based on Facial Emotions

G. Priyanka, T. Revathi, K. Muneeswaran

Abstract*: In the current era, the vast number of internet users accesses the social network sites like Facebook, Whatsapp, Twitter, and Instagram for uploading their images. User may upload all kind of images with different emotions. In this research article, we introduce novel facial emotion recognition technique using Convolutional Neural Network (CNN) approach and associated a caption automatically based on the identified emotion. We consider only still images for emotion classification. Using face detection algorithms faces were detected from the facial image which is followed by emotion prediction. We automated this detection process for images having frontal faces. In images having non-frontal faces, we manually plot the eye points for rotating the face in such a way that algorithm can detect faces easily. Inferences obtained as a result of experiments shows that the proposed work is capable of identifying minute differences between different emotions and predict accordingly. Based on the emotions identified corresponding captions are generated with better accuracy.

***Keywords*: caption, Convolutional Neural Networks, emotions, facial images**

I. INTRODUCTION

In the world of social media around 3.03 billion users are active, every user is sharing their photos daily and the difficult part is putting captions for those images. Caption is a short description about an image or a picture. People feel that they are wasting their time in thinking about the catchy captions for those shared images. As a statistics, in every minute of a day, Snapchat users share 5,27,760 photos, Instagram users post 49,380 photos, Tumblr users post 79,740 and Facebook users uploads 1,36,000 photos.

The proposed work helps the social media users to share the images without thinking about captions for it. Our system detects the emotions from the image based on facial emotions and generates the captions based on predicted emotion. With the help of this project we can save the time and energy for the users from generating captions for shared photos.

Main motivation for creating this system is to save the time and energy for the social media users from generating captions for every photo they are sharing in social media. Users will be feeling delighted to use this system. Based on the emotion predicted from the system caption will be randomly picked from storage. Captions that are generated from this system are manually created in a large amount with care.

II. RELATED WORKS

Several systems are available for emotion representation. The International Affective Picture System (IAPS) is generally utilized in investigations of emotions and has been described principally along with the components of valence, excitement and predominance. Some of the approaches are dimensional and categorical. The dimensional methodology proposed in [1], permits to speak to mapping emotions in a 2-measurement vector space, by utilizing few emotional measurements, for example, valence and excitement to uncover the primary attributes of feelings. The other methodology is categorical methodology, comprising of a couple of fundamental emotional names, for example, happiness, sadness, fear and anger. Emotions in the pictures are classified based on Discrete emotion model and Dimensional Emotion model in [1]. But only few emotions labels were used. Similarly mapping emotion to 2D space leads to loss of information.

As per the reports specified in [2], for the IAPS images, to classify the emotions gender is not going to contribute and hence can be neglected while processing facial images for emotions. This information demonstrates that there is a possibility of getting mixture of emotions in human faces.

In [3] two major emotions sadness and anger is discriminated based on kind of occasion of the particular kind. Sadness and anger effectively affect whether a future occasion is seen cynically or hopefully [4], and sadness prompts hazard avoidance, while anger prompts chance looking for [5].

The framework for emotion order dependent on a variant of auto-encoder called Generative Adversarial Network (GAN) with generator and discriminator was introduced in [6] and it is used to expand testing dataset and mix it up to all the identified seven emotion classes. In this way, the oddity of investigation comprises in expanding the quantity of classes from N to 2N (in the learning stage) by associating additional generated images to the actual data store. Thus, making the system more robust and agile by comprising the use of images with additional samples for age, race, and gender. In that work, both generator and discriminator are independent of each other and until handshaking the same process continues.

Revised Manuscript Received on December 5, 2019.
***** Correspondence Author
G. Priyanka*, Assistant Professor (Senior Grade), Department of CSE, Mepco Schlenk Engineering College, Sivakasi – 626005, India. Email: priyanka@mepcoeng.ac.in
T. Revathi, Professor and Head, Department of IT, Mepco Schlenk Engineering College, Sivakasi – 626005, India. Email: trevathi@mepcoeng.ac.in
K. Muneeswaran, Professor and Head, Department of CSE, Mepco Schlenk Engineering College, Sivakasi – 626005, India. Email: kmuni@mepcoeng.ac.in

As indicated by a report distributed by MarketsandMarkets in 2017 [7], the worldwide facial acknowledgement showcase was assessed at 3.37 billion USD in 2016 and it is relied upon to grow up to 7.76 billion USD by 2022, with an annual growth rate of 13.9 %.

In [8], emotions are classified based on the fact that different discrete emotions have different valence and arousal ratings. But it could not differentiate anger and fear since they two exhibits more or less same emotions. The work proposed in [9] discusses about the emotion classification in video clips based on Hilbert–Huang Transform (HHT) based visual highlights, HHT-based sound highlights, and cross-relationship highlights. But this technique has a major limitation on predicting particular emotion since emotion clusters are used.

Some steps were taken to use CNN for both feature extraction and emotion classification. This is mainly due to the voluminous nature of readily available emotion datasets. Because CNN model can be trained with many such emotion samples for each category as a result of that CNN is proved to be better than manually handcrafted features. [10] proposed a deep CNN architecture based on visual sentiment concept classification method. In that work, label of the image is represented as Adjective Noun Pairs (ANPs). By using the mathematical approach of ANPs, emotion recognition in images were done. [11] utilized CNNs for determining sentiments from the images based on the visually discriminating content in it. In all the above works they had used CNNs for visual content and emotion determination from the images.

The proposed emotion recognition model in [12] used two variants of CNNs as binary CNN (B-CNN) to predict positive or negative emotion class and an eight-class CNN (E-CNN) to predict the actual emotion class label. B-CNN architecture used three convolutional layers and three fully-connected layers. The first convolutional layer takes the color image of size 224 ×224 as input and process it with 96 filters each of size 11×11 with a striding of 4 pixels. Next convolutional layer for filtering and extracting the features uses 256 filters each of size 5 ×5 with a stride of 2 pixels. The third convolutional layer used 384 such filters each of size 3 ×3 with default stride. Each of the above specified convolutional layer is also followed by a max-pooling layer and a normalization layer. Next to convolutional layers two dense layers each with 4096 neurons were designed. Final dense layer used two neurons and used softmax as activation function to produce the positive or negative prediction of image emotion.

E- CNN is designed for specific emotion class prediction based on the binary classification output. The first convolutional layer used 64 filters each of size 3×3 with 2 pixels striding. Next convolutional layer used 128 filters each of size 3×3 with 2 pixels striding. The third convolutional layer had 256 filters each of size 3×3 with 3 pixels striding. The fourth convolutional layer had 512 filters each of size 3×3 with same striding as previous convolutional layer. The fifth convolutional layer has 512 filters each of size 3×3 with three pixels striding like previous layer. Each of the above mentioned convolutional layer is followed by a max-pooling layer of size 3x3 with two pixels striding. Followed by the convolutional layer, two dense layers each having 4096 neurons were designed. Last dense layer uses 8 neurons to determine the eight emotion class label as output. This network leads to an overall accuracy of 64.6% for benchmark dataset like CK+.

III. PROPOSED SYSTEM

The proposed CNN model has 3 Convolution layers and 3 fully connected layer as shown in Table 1. Each Convolution layer is followed by a Rectifier linear unit layer and a Max-Pooling layer. In the proposed work, seven different kind of emotions is determined.

The proposed CNN does hyper parameters tuning for the E-CNN model proposed in [1]. Seven types of basic emotions that can be classified from facial images that can be identified by the proposed work are as follows: Anger, Disgust, Fear, Happy, Sad, Surprise and Neutral.

Table -I: Proposed CNN Architecture

Layer Name	Size	Number of Kernels
Input layer	224 X 224	-
Convolution Layer	11 X 11	112
Rectification Layer	-	-
Max-Pooling Layer	3 X 3	-
Convolution Layer	5 X 5	284
Rectification Layer	-	-
Max-Pooling Layer	3 X 3	-
Convolution Layer	3 X 3	324
Rectification Layer	-	-
Max-Pooling Layer	3 X 3	-
Fully convoluted layer	4096	-
Fully convoluted layer	4096	-
Fully convoluted layer	7	-

A. Automatic Caption Generation Process

The following Figure 1. clearly shows the process of automatic caption generation. It starts with reading an input facial image and detecting all frontal faces using Viola Jones Algorithm with maximum acceptable deviation angle range of -30^0 to $+30^0$. In case of tilted human faces the picture is rotated in such a way that eyes are aligned horizontally to x-axis which is also followed by human face detection using popular Viola-Jones Algorithm.

The overall system design is depicted in Figure 2. For frontal human faces detection Haar features are extracted and Viola Jones Algorithm is applied. In Haar features extraction for the input image, always eyes and nose regions are darker than cheeks and eyes regions respectively. Thus, 2-rectangle based Haar features are used for detecting the human faces. Non-frontal faces (i.e. above 30^0 variation) are converted to frontal faces based on eye positioning. Picture is rotated such that both left-eye and right-eye are aligned horizontally straight. For this, points are manually marked by the user and based on theta value is computed based on equation (1).

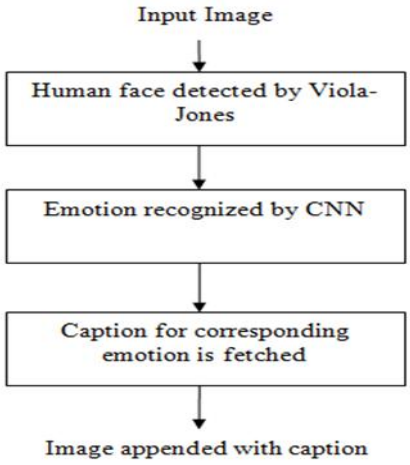

Fig. 1. Steps for Caption Generation Process

$$angle = \tan(\frac{rCenterY2 - lCenterY1}{rCenterX2 - lCenterX1}) \quad (1)$$

$$theta = \left(\frac{angle}{pi}\right) * 180^0 \quad (1)$$

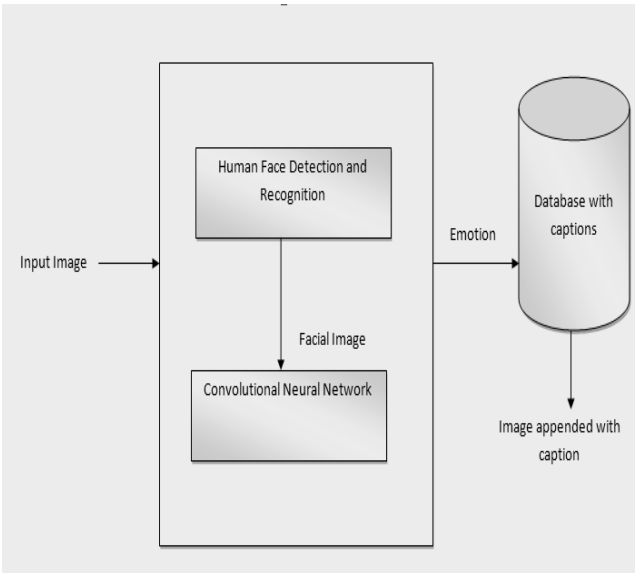

Fig. 2. System Design for Automatic Caption Generation Process

Fig. 3. shows the process of rotating and cropping the non-frontal face into frontal face. This is done by calculating theta value by using equation (1) and then rotated and cropped face was given as input to Viola Jones Algorithm for face detection.

Once the face is detected, emotion present in the face is detected by using our constructed CNN. Seven basic emotions such as Anger, Disgust, Fear, Sad, Surprise, Happy and Neutral are identified by the proposed CNN. Once the emotion was detected corresponding caption will be fetched from DB which already consists of more than 100+ captions for each emotion. Finally image appended with caption will be produced as output.

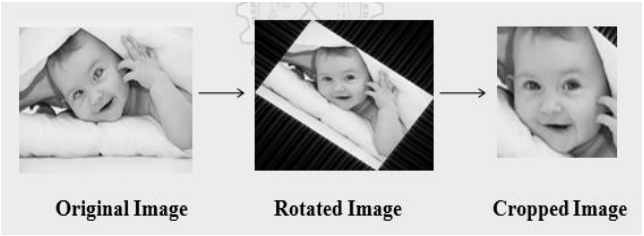

Fig. 3. Aligning non-frontal faces using Rotation and Cropping

IV. RESULTS AND DISCUSSION

In this chapter, the results of Automatic caption generation from images based on facial emotions are discussed for CK+ dataset. The constructed CNN was trained with 50% images in the CK+ dataset and remaining is used for testing the network. The sample images in the dataset for all emotion categories are shown in Fig. 4.

Fig. 5. (a) shows the sample input frontal face image given to the system, Fig. 5. (b) shows the output of face detection using Viola Jones Algorithm, Fig. 5. (c) shows the output of cropped image, Fig. 5. (d) depicts the emotion recognized output using CNN and Fig. 5. (e) display the final output of proposed system with caption attached to it. Fig. 6. (a) shows the sample input non-frontal face image given to the system, Fig. 6. (b) shows the eye points manually selected for determining theta value, Fig. 6. (c) displays the rotated human face and Fig. 6. (d) shows the cropped face after applying rotation operation using theta value and face was detected using Viola Jones Algorithm, Fig. 6. (e) depicts the emotion recognized output using CNN and Fig. 6. (f) Displays the final output of proposed system with caption attached to it.

A. Performance Measures

The parameters such as accuracy, misclassification rate, true positive rate, false positive rate, specificity and precision are calculated. Accuracy determines the measure of how often the classifier is correct. Misclassification or error rate denotes the measure of how often the classifier is wrong in predicting the result. True positive rate specifies the measure of how often the system predicts abnormal, when it's actually abnormal and vice versa. False positive rate discusses about the measure of how often it predicts abnormal, when it's actually normal and vice versa. The confusion matrix for the proposed system is shown in Fig. 7.

From the confusion matrix shown in Fig. 7, it was clear that the system attains an overall accuracy of 89.4%, with misclassification rate as 10.06%, true positive rate as 1 and false positive rate as 0.4. The work proposed in [1] achieved an overall accuracy of 64.6% for CK+ dataset and our proposed system achieves an overall accuracy of 89.4%. The individual emotion accuracy for seven emotions in CK+ data set for the proposed architecture is depicted in Fig. 8.

Fig. 4. Sample Dataset images of Human faces with various emotions

Fig. 5. (a) Sample Input frontal face image; (b) Frontal Face Detection using popular Viola Jones Algorithm; (c) Cropped face after detection (d) Emotion recognized and appended to the image (e) Caption was randomly selected from database for the identified emotion.

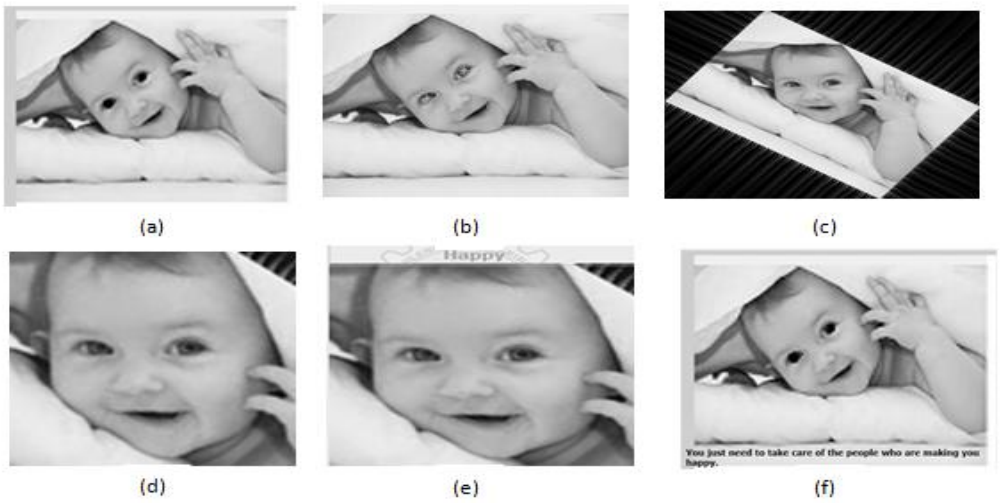

Fig. 6. (a) Sample Input non-frontal face image; (b) Eye points plotted image; (c) Rotated human image using theta value; (d) Cropped face image after rotation; (e) Emotion recognized and appended to the image; (f) Caption was randomly selected from database for the identified emotion.

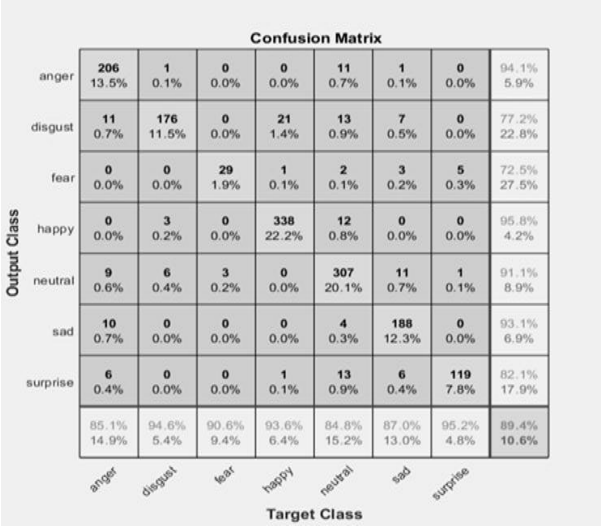

Fig. 7. Confusion Matrix of the proposed system using CK+ dataset.

Table-II: Individual Emotion accuracy for proposed system in CK+ dataset

SL.NO	Emotion	Accuracy
1	Anger	85.1%
2	Disgust	94.6%
3	Fear	90.6%
4	Neutral	84.8%
5	Happy	93.6%
6	Sad	87.0%
7	Surprise	95.2%

V. CONCLUSIONS AND FUTURE WORKS

We automated this caption generating process, so people will be feeling free to share images. In the proposed system, catchy captions will be automatically generated for the posted pictures in social media websites. In this work, a convoluted neural network is proposed to extract features from the faces and to classify the facial images based on emotions. Based on the emotion predicted, captions are generated automatically. We consider only still images for emotion classification. We first use face detection algorithm to detect face from an image which is followed by emotion prediction. We automated this detection process for images having frontals faces. In images having non-frontal faces we manually plot the eye points for rotating the face in such a way that algorithm can detect faces easily. Our work concludes with 89.4% accuracy for identifying the correct emotions.

In future, this work can be further improved by integrating this process in all social media applications like Instagram, Snapchat, etc., Finally users will be feeling delighted by saving time for generating catchy captions with the help of this project. Once it is integrated, users can upload their photos and can get catchy captions as per their wish. If the user is not satisfied with the generated caption, he/she can select some other caption too. This project can further be extended to generate new captions using sentimental analysis so that user satisfaction level may get increased. Also automatic alignment of non-frontal face can be taken into consideration for future enhancement.

ACKNOWLEDGMENT

We are very much thankful to our College Management and Principal for allowing us to do this project. A special thanks to our fellow students S.N. Roobini, M.Harishma and B. Swarnalakshmi.

REFERENCES

1. E. Dellandrea , N. Liu , L. Chen , Classification of affective semantics in images based on discrete and dimensional models of emotions, in: Proceedings of the International Workshop on Content-Based Multimedia Indexing (CBMI) IEEE (2010) 1–6.
2. M.M. Bradley, M. Codispoti, D. Sabatinelli, P.J. Lang, Emotion and motivation II: sex differences in picture processing (2001) Sep;1(3):300-19.
3. D. Keltner P.C. Ellsworth and K. Edwards Beyond simple pessimism: Effects of sadness and anger on social judgement. Journal of Personality and Social Psychology (64) (1993) 740-752.
4. J.S. Lerner and D. Keltner, Beyond valence: Toward a model of emotion-specific influences on judgement and choice, Cognition and Emotion, 14(4) (2000) 473-493.
5. J.S. Lerner and D. Keltner, Fear, anger and risk, J Pers Soc Psychol. (2001) 81(1):146-59.
6. Traian Caramihale, Dan Popescu and Loretta Ichim, Emotion Classification Using a Tensorflow Generative Adversarial Network Implementation, Symmetry (2018) 10(9), 414.
7. https://www.marketsandmarkets.com/PressReleases/facial-recognition.asp .
8. J.A. Mikels , B.L. Fredrickson , G.R. Larkin , C.M. Lindberg , S.J. Maglio,P.A. Reuter-Lorenz , Emotional category data on images from the international affective picture system, Behav. Res. Methods 37 (4) (2005) 626–630.
9. Shasha Mo, Jianwei Niu, Yiming Su, Sajal K. Das, A novel feature set for video emotion recognition, Neurocomputing, (2018) (291) 11-20.
10. Tao Chen, Damian Borth, Trevor Darrell, Shih-Fu Chang, DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks, Computer Vision and Pattern Recognition (2014).
11. V. Campos, A. Salvador, B. Jou, X. Giro-i-Nieto, Diving Deep into Sentiment: Understanding Fine-tuned CNNs for Visual Sentiment Prediction, 1st International Workshop on Affect and Sentiment in Multimedia, Brisbane, Australia: ACM (2015).
12. Xuanyu He, Wei Zhang, Emotion recognition by assisted learning with convolutional neural networks, Neurocomputing, 291 (2018) 187-194.

AUTHORS PROFILE

G. Priyanka, received the Bachelor of Engineering degree in Computer Science and Engineering from Anna University, Tirunelveli, India in 2011 and the Master of Engineering in Computer Science and Engineering from Anna University, Chennai, Tamilnadu, India, in 2013. She is in teaching and research for the past 5 years and 1 year respectively and currently, she is working as Assistant Professor (Senior Grade) in Computer Science and Engineering Department at Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India. Her area of research interests includes Video Analytics, Machine Learning & Deep Learning techniques, and Data Analytics. She has authored or co-authored about 10 publications in International Journal/Conference level. She is a life member of Computer Society of India (CSI) and Indian Society for Technical Education (ISTE).

T. Revathi, completed her B.E Electrical and Electronic Engineering in the year 1986 and M.E Computer Science and Engineering in the year 1995 followed by that received Ph.D in Computer science and Engineering from Manonmaniam Sundaranar University, Tamilnadu, India in the year 2008.

Her area of interest is pattern recognition, Big data analytics and networking. She has published many technical articles in reputed International Journals and

conferences. She has produced around six Ph.D research scholars in her domain of interest. She is life member of Computer Society of India (CSI), Indian Society for Technical Education (ISTE) and Institute of Electronics and Telecommunication Engineers (IETE).

Muneeswaran Karuppiah received the bachelor of engineering degree in Electronics and Communication engineering from Madurai Kamarajar University, Tamilnadu, India in 1984 and the master of engineering in computer science and engineering from Bharathiyar University, Tamilnadu, India, in 1990. In 2006, he received the Ph.D. degree in computer science engineering from M.S. University, Tamilnadu, India. He is in teaching and research for the past 34 years and 17 years respectively and currently, he is working as senior professor in Computer Science and Engineering Department at Mepco Schlenk Engineering College, Tamilnadu. His research interests are image analysis, machine learning techniques, and Data Analytics. He has authored or co-authored about 125 publications in journal/conference level and one book on compiler design with Oxford University. He is currently working as Senior Professor in the Department of Computer Science and Engineering at Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India and he is a life member of Computer Society of India (CSI), Indian Society for Technical Education (ISTE) and Institute of Electronics and Telecommunication Engineers (IETE).