Effective Emotion Recognition from Partially Occluded Facial Images Using Deep Learning

Smitha Engoor, Sendhilkumar Selvaraju, Hepsibah Sharon Christopher, Mahalakshmi Guruvayur Suryanarayanan, Bhuvaneshwari Ranganathan

Effective Emotion Recognition from Partially Occluded Facial Images Using Deep Learning

Smitha Engoor[1[0000-0001-7443-9962]], Sendhilkumar S.[2*[0000-0001-6006-1866]],Hepsibah Sharon C.[3[0000-0001-9117-5153]],Mahalakshmi G.S.[4[0000-0002-6286-2058]] and Bhuvaneshwari R.[5[0000-0002-1912-0120]]

[1,2] Dept. of Information Science & Technology, Anna University, Chennai, Tamil Nadu,INDIA.

[3,4,5]Dept. of Computer Science & Engineering, Anna University, Chennai, Tamil Nadu, INDIA

*Corresponding Author: ssk_pdy@yahoo.co.in

Abstract. Effective expression analysis hugely depends upon the accurate representation of facial features. Proper identification and tracking of different facial muscles irrespective of pose, face shape, illumination, and image resolution is very much essential for serving the purpose. However, extraction and analysis of facial and appearance based features fails with improper face alignment and occlusions. Few existing works on these problems mainly determine the facial regions which contribute towards discrimination of expressions based on the training data. However, in these approaches, the positions and sizes of the facial patches vary according to the training data which inherently makes it difficult to conceive a generic system to serve the purpose. This paper proposes a novel facial landmark detection technique as well as a salient patch based facial expression recognition framework based on ACNN with significant performance at different image resolutions.

Keywords: Emotion Recognition, Partial Occlusion, Facial features, ACNN

1 Introduction

Facial expression classifiers are successful on analyzing constrained frontal faces. There have been less reports on their performance on partially occluded faces [3]. In this paper,emotion recognition is performed with partially occluded faces by identifying the blocked region in the image and get information from unblocked region or informative region. Facial expressions inherently extend to other parts of the face producing visible patches. If the facial parts are hidden or blocked by hands or spectacles or any other means, symmetric parts of the face or the regions that contribute to the expression shall be involved for facial emotion recognition. Inspired by the intuition, Attention based Convolutional Neural Networks (ACNNs) automatically handles the occluded regions by paying attention to informative facial regions. Every Gate Unit of ACNN associates an importance based weight factor via thorough adaptive learning.

In this work, two versions of ACNN: patch-based ACNN (pACNN) and global– local-based ACNN (gACNN) are deployed. pACNN has a Patch-Gated Unit (PG-Unit) which is used to learn, weigh the patch's local representation by its unobstructedness that is computed from the patch itself.gACNN integrates local and global

representations concurrently. A Global-Gated Unit (GG-Unit) is adopted in gACNN to learn and weigh the global representation.

2 Related Work

2.1 Facial Occlusion Method

VGG Net has the image represented as feature maps. ACNN decomposes the feature maps into multiple sub feature maps to obtain local patches. The feature maps are also sent to gg-unit to identify occlusion. The pg-unit and gg-unit are concatenated and softmax loss is used to predict the final output.

Patch based ACNN (pACNN) is decomposed into two schemes: (i) region decomposition (ii) occlusion perception. In region decomposition 69 facial landmarks and selecting 24 points [5] which covers all information. In occlusion perception it deals with pg-unit. In each patch-specific PG-Unit, the cropped local feature maps are fed to two convolution layers without decreasing the spatial resolution, so as to preserve more information when learning region specific patterns. Then, the last set of feature maps are processed in two steps: vector-shaped local features and attention net that estimates the importance based scalar weights. The sigmoid activation of Attention net forces the output as [0,1], where 1 indicates the most salient unobstructed patch and 0 indicates the completely blocked patch

Global-local based ACNN (gACNN) is divide into two schemes: (i) integration with full face region (ii) global-gated unit. In integration with full face region the gACNN takes the whole face region on the one hand, the global-local attention method is used to know the local details and global context cues. The global representation is then weighed by the computed weight.The ACNNs rely on the detected landmarks. It cannot be neglected that facial landmarks will suffer misalignment in the presence of severe occlusions. The existing ACNNs are not sensitive to the landmark misalignment.

2.2 Detecting the Shape of Faces

Regression and Deep regression are proposed in the literature for face detection purposes [7]. Deep regression network aims at characterizing the nonlinear mapping from appearance to shape. For a deep network with $m - 1$ hidden layers, de-corrupt auto-encoders are used for recovering the occluded faces. Auto-encoder will tackle partial occlusion and de-corrupt auto-encoder will occlude the parts by partitioning the face image x into j components. Therefore, after partitioning the image there will be 68 facial points which has 7 components. The components cover all the information regions. Considering that the face appearance varies under different poses and expressions, it is nontrivial to design one de-corrupt auto-encoder network to reconstruct the details of the whole face.

The third approach, cascade deep regression with de-corrupts auto-encoder, concatenates both deep regression and de-corrupts auto-encoder to get the local patches. By learning de-corrupt auto-encoder networks and deep regression networks under a cascade structure, they can benefit from each other. On the one hand, with more accurate face shape, the appearance variations within each component becomes more con-

sistent, leading to more compact de-corrupt auto-encoder networks for better de-corrupted face images. On the other hand, the deep regression networks that are robust to occlusions can be attained by leveraging better de-corrupted faces

2.3 Localization of ROI

The eyes and nose localization is detected by Haar cascade algorithm. The Haar classifier returns the vertices of the rectangular area of detected eyes. The eye centers are computed as the mean of these coordinates. Similarly, nose position was also detected using Haar cascades. In case the eyes or nose was not detected using Haar classifiers [2], the system relies on the landmark coordinates detected by anthropometric statistics of face. In summary, Deep Regressive and De-corrupt Auto-encoders does not recover other type of deep architecture for the genuine appearance for occluded parts. In traditional CNN based approaches,registered facial images were handled and partial occlusion is not addressed.

3 Proposed Work: ACNN for Partial Facial Occlusion

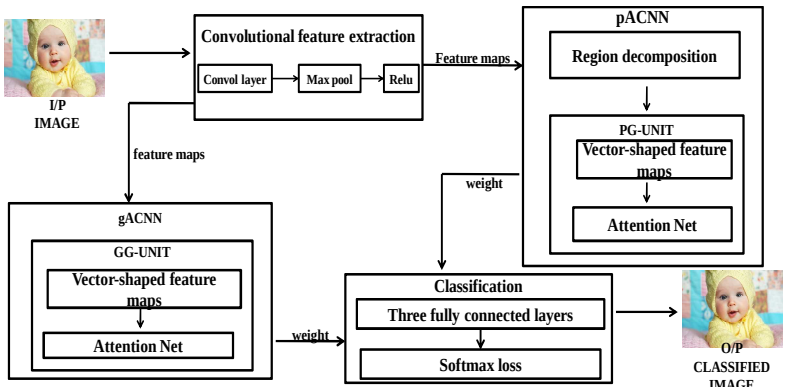


Figure 1. Working Model for ACNN based Partial Facial Occlusion Driven Emotion Detection

In this work, to address the occlusion issue, ACNN endeavors to focus on different regions of the facial image and weighs each region according to its obstructed-ness as well as its contribution to facial emotions. The CNN takes facial image as the input. The image is fed into a convolutional net as feature maps. Then, ACNN decomposes the feature maps of the whole face to multiple sub-feature maps to obtain diverse local patches. Each local patch is encoded as a weighed vector by a Patch-Gated Unit (PG-Unit). A PG-Unit computes the weight of each patch by an Attention Net, considering its obstructedness. The feature maps of the whole face are encoded as a weighed vector by a Global Gated Unit (GG-Unit). The weighed global facial features with local representations are concatenated and serve as a representation of occluded face. Two fully connected layers are followed to classify the facial emotions. ACNNs are optimized by minimizing the softmax loss (refer figure 1).

3.1 Convolution and Max-Pooling Layer

The work presented here is inspired by the techniques provided by the GoogLeNet and AlexNet architectures. Our network consists of two traditional CNN modules (a traditional CNN layer consists of a convolution layer and a max pooling layer). Both of these modules use rectified linear units (ReLU) which have an activation function.Using the ReLU activation function[1] allows us to avoid the vanishing gradient problem caused by some other activation functions. Following these modules, we apply the techniques of the network in network architecture and add two "Inception" style modules, which are made up of a 1×1, 3×3 and 5×5 convolution layers (Using ReLU) in parallel. These layers are then concatenated as output and we use two fully connected layers as the classifying layers.

In yet another approach, bidirectional warping of Active Appearance Model (AAM) and a Supervised Descent Method (SDM) called IntraFace to extract facial landmarks is proposed, however further work could consider improving the landmark recognition in order to extract more accurate faces. IntraFace uses SIFT features for feature mapping and trains a descent method by a linear regression on training set in order to extract 49 points. These points are used to register faces to an average face in an affine transformation. Finally, a fixed rectangle around the average face is considered as the face region.Once the faces have been registered, the images are resized to 48×48 pixels for analysis. Even though many databases are composed of images with a much higher resolution testing suggested that decreasing this resolution does not greatly impact the accuracy, however vastly increases the speed of the network. To augment the data, we shall extract 5 crops of 40×40 from the four corners and the center of the image and utilize both of them and their horizontal flips for a total of 10 additional images.

3.2 Analysing Facial Image Patches

Facial expression is distinguished in specificfacial regions, because the expressions are facial activities invoked by sets of muscle motions. Localizing and encoding the expression related parts is of benefit to recognize facial expression. To find the typical facial parts that related to expression, we first extract the patches according to the positions of each subject's facial landmarks. We first detect 68 facial landmark points and then, based on the detected 68 points, select or re-compute 24 points that cover the informative region of the face, including the two eyes, nose, mouth, cheek, and dimple. The selected patches are defined as the regions taking each of the 24 points as the center. It is noteworthy that face alignment method in is robust to occlusions, which is important for precise region decomposition.The patch decomposition operation is conducted on the feature map from convolution layers rather than from the original image. This is because sharing some convolutional operations can decrease the model size and enlarge the receptive fields of subsequent neurons. Based on the 512×28×28 feature maps as well as the 24 local region centers and get a total of 24 local regions, each with a size of 512×6×6.

In PG-CNN [4], the idea was to embed the PG-Unit (refer Figure 1) to automatically percept the blocked facial patch and pay attentions mainly to the unblocked and informative patches. In each patch-specific PG-Unit, the cropped local feature maps are fed to two convolution layers without decreasing the spatial resolution, so as to

preserve more information when learning region specific patterns. Then, the last 512×6×6 feature maps are processed in two branches. The first branch encodes the input feature maps as the vector-shaped local feature. The second branch consist an attention net that estimates a scalar weight to denote the importance of the local patch. The local feature is then weighted by the computed weight.In PG-Unit, each patch is weighted differently according to its occlusion conditions or importance. Through the end-to-end training of overall PG-CNN, PG-Units can automatically learn low weights for occluded parts and high weights for unblocked and discriminative parts.

4 Experimental Results & Discussion

4.1 Dataset

In this work, Facial Expression Dataset with Real Occlusion (FER-RO) is used. The occlusions involved are mostly real-life originating with natural occlusions limited to arising from sunglasses, medical mask, hands or hair (refer Table 1).

TABLE I. TOTAL NUMBER OF IMAGES FOR EACH CATEGORY IN FER-RO

Dataset	neutral	anger	disgust	fear	happy	sad	surprise
Total	50	53	51	58	59	66	63
Training	22	17	18	27	25	33	22
Testing	28	36	33	31	34	33	41

4.2 Results

4.2.1 Convolutional feature extraction

The input to convolutional layer is of fixed size 224 x 224 RGB image. The image is passed through a stack of convolutional layers, where the filters were used with a very small receptive field: 3×3 (which is the smallest size to capture the notion of left/right, up/down, center). In one of the configurations, it also utilizes 1×1 convolution filters, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1-pixel for 3×3 conv. layers. Spatial pooling is carried out by five max-pooling layers, which follow some of the convolutional layers, not all the conv. layers are followed by max-pooling, which is performed over a 2×2 pixel window, with stride 2.

4.2.2 Patch-based ACNN

After identifying the 68 facial landmark points we select or recompute 24 points that cover the informative regions of the face, including the eyes, nose, mouth, cheeks. Then we extract the patches according to the positions of each subject's facial land-marks. The selection of facial patches follows the procedure below:

- Pick 16 points from the original 68 facial landmarks to cover each subject's eyebrows, eyes, nose, and mouth. The selected points are indexed as 19, 22, 23, 26, 39, 37, 44, 46, 28, 30, 49, 51, 53, 55, 59, 57.
- Add one informative point for each eye and eyebrow. We pick four point pairs around the eyes and eyebrows, then compute midpoint of each point pair as delegation. It is because we conduct patch extraction on convolutional feature maps rather than on the input image, adjacent facial points on facial images will coalesce into a same point on feature maps.

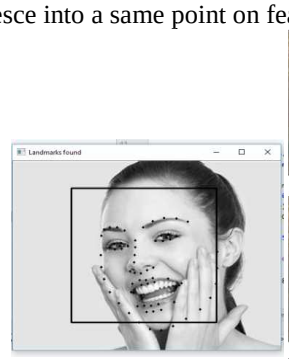


Figure 2. Facial LandMark Detection for images from FFE [6]

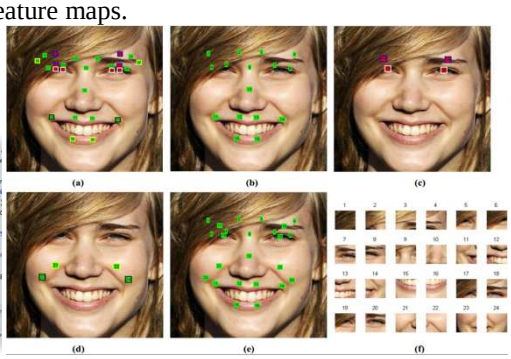


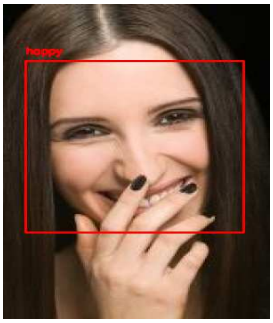
Figure 3. Patch Identification by Region Decomposition

Based on the 512×28×28 feature maps as well as the 24 local region centers (refer figure 2), we get a total of 24 local regions, each with a size of 512×6×6. Following this, we embed the Patch-Gated Unit in the pACNN (refer figure 3). Under the attention mechanism in the proposed Gate-Unit, each cropped patch is weighed differently according to its occlusion conditions or importance.

4.2.3 Global local-based ACNN

gACNN takes global face region into consideration. Global-Local Attention method helps to infer local details and global context cues from image concurrently. On the other hand, gACNN can be viewed as a type of ensemble learning, which seeks to promote diversity among the learned features. We further embed the GG-Unit in gACNN to automatically weigh the global facial representation.

4.2.4 Classification



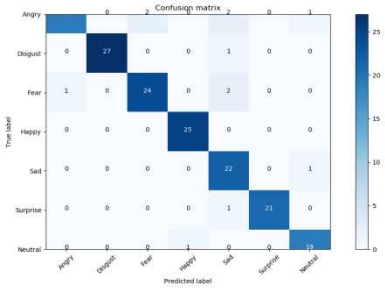


Figure 4. Emotion Classification for facial image with partial occlusion

Figure 5. Confusion Matrix

The output from the convolutional layers represents high-level features in the data. While that output could be flattened and connected to the output layer, adding a fully-connected layer is a deep learning framework. After feature extraction we need to classify the data into various classes, this can be done using a fully connected neural network. The weight of patch-based ACNN and global local-based ACNN is integrated to find the emotion on the given face.The emotions can be classified into seven categories such as: happy, sad, disgust, fear, neutral, anger and surprise. pACNN is capable of discovering local discriminative patches and is much less sensitive to occlusions. Among all the different network structures, pACNN and gACNN are capable of perceiving occlusions and shifting attention from the occluded patches.The proposed approach arrives at a precision of 0.929 for partially occluded faces (refer figure 4 & figure 5).

4.2.5 Discussion

Sparse representation classifiers were previously applied for partial facial occlusions [8][9]]. However, Li et al [5] attempts to apply ACNN for non-occluded and occluded face images. These occluded images attempted by Li et al [5] are full occlusions. Wang et al [10] have explored region based attention models for robust expression recognition, however , in this work [10] patches were not involved. Yet another work on the lines of using attention models are discussed in Wang et al [11]. Here, ACNNs are applied over unconstrained facial expressions but which were shot in the wild. Neither of the above approaches are applied over partial facial occlusions. Retail consumer segment has great potential to leverage into facial expression recognition. Since partial facial occlusion is very natural and common in retail, the proposed work records the usage of ACNN for partial facial occlusions captured with better clarity, which is reportedly first work in this direction. In addition, the proposed work uses less images for training for two reasons: 1. The ACNN model used with patch based approaches has the capability of learning from least inputs 2. The images fed to training were peak images for respective expressions which is enough to govern the handling of partially occluded faces.

5 Conclusion

This paper proposes CNN with attention mechanism (ACNN) for facial expression recognition in the presence of occlusions. The proposed work shall be applied successfully for customer interest detection [12], human behavior analysis [13]. In order to make it more real-time, handling video facial data is essential. For real-time facial emotion recognition with occlusion, the losses have to be very minimal [14].Further, to maintain a balance [15] between rich and poor facial feature classes, feature augmentation and feature normalization have to be addressed.

6 Acknowledgements

This Publication is an outcome of the R&D work undertaken in the project under the Visvesvaraya Ph.D. Scheme (Unique Awardee Number: VISPHD-MEITY-2959) of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

1. Mollahosseini, A., Chan, D., Mahoor, M. H.:. Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter conference on applications of computer vision (WACV), pp. 1-10. IEEE.(2016, March)
2. Happy, S.L., Routray, A.: Automatic facial expression recognition using features of salient facial patches. IEEE transactions on Affective Computing, vol.6, no. 1, pp.1-12.(2014)
3. Kotsia, I., Buciu, I., Pitas, I. : An analysis of facial expression recognition under partial facial image occlusion. Image and Vision Computing, vol. 26, no. 7, pp. 1052-1067.(2008)
4. Li, Y., Zeng, J., Shan, S., Chen, X.: Patch-Gated CNN for occlusion-aware facial expression recognition. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 2209-2214. IEEE.(2018, August)
5. Li, Y., Zeng, J., Shan, S., Chen, X.: Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2439-2450.(2018)
6. Michael J Lyons, Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba, Julien Budynek.: The japanese female facial expression (jaffe) database. In: FG, pp. 14–16. (2005)
7. Zhang, J., Kan, M., Shan, S., Chen, X.: Occlusion-free face alignment: Deep regression networks coupled with de-corrupt autoencoders. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3428-3437. (2016)
8. S. F. Cotter, Sparse representation for accurate classification of corrupted and occluded facial expressions, Proc. ICASSP, pp. 838-841, Apr. (2010).
9. S. F. Cotter, Weighted voting of sparse representation classifiers for facial expression recognition, Proc. Signal Process. Eur. Conf., pp. 1164-1168, (2010).
10. Wang, K., Peng, X., Yang, J., Meng, D. and Qiao, Y., Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 29, pp.4057-4069, (2020).
11. Wang, C., Hu, R., Hu, M., Liu, J., Ren, T., He, S., Jiang, M. and Miao, J., Lossless Attention in Convolutional Networks for Facial Expression Recognition in the Wild. arXiv preprint arXiv:2001.11869. (2020).
12. Gozde Yolcu, Ismail Oztel,Serap Kazan, Cemil Oz, Filiz Bunyak, "Deep learning-based face analysis system for monitoring customer interest", springer 2019.
13. Muhammad Sajjad, Sana Zahir, Amin Ullah, Zahid Akhtar and Khan Muhammad "Human Behavior Understanding in Big Multimedia Data Using CNN based Facial Expression Recognition", in springer. 2019.
14. Wei, X., Wang, H., Scotney, B. and Wan, H., 2020. Minimum margin loss for deep face recognition. Pattern Recognition, 97, p.107012.
15. Wang, Pingyu, Fei Su, Zhicheng Zhao, Yandong Guo, Yanyun Zhao, and Bojin Zhuang. "Deep class-skewed learning for face recognition." Neurocomputing 363 (2019): 35-45.