International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April 2020

CNN-BLSTM Joint Technique on Dynamic Shape and Appearance of FACS

Nazmin Begum, A Syed Mustafa

Abstract— Facial recognition is a process where we can identify or verify a person from digital image or videos and is used in ID verification services , protecting law enforcement ,preventing retail crime etc. Past work on automatic analysis of facial expression focuses on detecting the facial expression and exploiting the dependencies among AU's. But, spontaneous detection of facial expression depending on various factors such as shape, appearance and dynamics is very difficult. Joint learning of shape , appearance and dynamics is done by a deep learning technique.This includes a convolutional neural networks and bidirectional long short term memory(CNN-BLSTM). This combination of CNN-BLSTM excels the modeling of temporal information. FERA2015 dataset achieves the state of art.

Index Terms—: Bidirectional LSTM,Convolutional Neural Networks, Face Recognition.

I. INTRODUCTION

One of the strongest indications for emotion is our face. Most of the muscles are triggered by one single nerve called facial nerve. Ekman and Friesen promotethe Facial Action Coding System(FACS) provides a scientific and comprehensive system, by in lieu of bycombination of individual muscle actions mentioned as Action Units (AU). Recognition of facial expression is one of the primary challenging problems. Peculiar features of person, different pose, low intense expressions captured spontaneously cannot be identified easily. In addition to this, the co-effects caused by co-occurring Action Units is also a factor , which increases the difficulty level in training the data.. Facial features are classified intoappearance feature and geometric based features. Appearance based feature captures local and global appearance changes, for example Harr feature Local binary pattern, Gabor wavelets, Canonical appearance. Geometric feature based includes direction or magnitude of skin surface and salient feature points. Image, Text or data classification can be performed using Deep Learning techniques. Artificial recurrent neural network called LSTM is used for classifying, processing and making predictions by taking time series. Quick access to the data, increased computing power and high performance are the three factors which increases the accuracy.

Revised Manuscript Received on March, 29 2020.
* Correspondence Author
Mrs. Nazmin Begum*, Assistant Professor, Department of Computer Science & Engineering, Dayananda Sagar University, Bengaluru, Karnataka , India.
Dr. Syed Mustafa, Professor and Head of the Department, Department of ISE, HKBK College of Engineering, Bengaluru, Karnataka , India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Different kinds of objects can be trained with freely available dataset such as Image Net and PASCAL VoC. Deep learning also trains huge amount data by reducing training time. To perform new recognition task retraining of models is done .This process is called Transfer Learning and is accurate only on smaller dataset. One such model is Alex Net which was trained 1.3 millions high resolution images 1000 different objects. Dynamics, appearance and face shape are the prominent three factors for facial AU recognition. Detection of Action units can be enhanced by Deep Learning techniques in images. One such technique used in this paper is CNN. CNN approach learns all the key features jointly.In addition to that BLSTM can also be used, as it stores information for extended time intervals. Multiplicative gates in BLSTM control the access for memory blocks. BLSTM can also be used to predict emotion from multimedia, which yielded promising results.

II. RELATED WORK

- Facial recognition using CNN was first proposed by Fasel [3]. This CNN architecture consist of 6 layers for sorting 7 facial expressions which includes two convolutional layers, two sub sampling layer and two FC (fully connected layers). Initially it consists of 2 versions. In version1 the filter size was 5*5 in first convolutional layers. In version 2 features were extracted using 3 different size filters such as 5*5, 7*7, 9*9. This corresponds to three scales of different filters which are joined to each other at FC (Fully connected layer of network).
- Learning to localize certain dynamic parts and encode them for classification in videos was proposed by Liu et al. in [2].
- Gudi et al.[1] implemented a deep CNN (Convolution Neural Network) which includes three layers of convolution , one layer of sub-sampling one layer of fully connected layer in order to resolvethe occurrence and intensity of Facial AU's using a softmax objective function.
- To know the temporal appearance features for facial expressions recognition,Jung et al.[4] also considered a deep CNN technique. From the detected facial landmarks they also learnt the geometrical features. For this they used a deep Neural Network.
- Tang[7] achieves State of art byreplacing softmax function with L2SVM.This can be achieved in two stages of network. In first stage a network is trained in a supervised manner. In second stage the output of first stage is given to SVM.The initial global contrast normalization is used as a preprocessing step.

© Copyright: All rights reserved.

- In paper[8] "Supervised descent methods and its application to face alignment", the author proposed a non linear least square (NLS) function. In the learning period the SNM identifies chain of direction that reduces the mean of NLS at different points. During testing SDM minimizes the NLS objectives using learnt decent directions.
- In paper[1], author added a region relation modeling block with the help of an improved graph convolution network to simulate the relations among different facial regions for facial expression restoration method.This method is important in the analysis of facial expression in challenging environments such as low resolution and occlusion**.**
- Most of the above methods [1, 3,5, 6],b do not implement the facial features like face outline. Appearance altogether. Techniques implementing all those features together do not learn jointly. Those which consider all three of them do not learn them jointly. In order to get the best grouping of these features, it is essential to mold them jointly because it allows us to find the best possible combination of these features. CNN methods implement the fixed time window to know the temporal information. This restricts the right of entry to temporal information in a stipulated time window . We can overcome the difficulty of a fixed time window in CNN by using Bidirectional LSTM for learning long term temporal dependency.

III. METHODOLOGY

To learn the appearance of image and shape, we use binary image masks which consist of number of rectangular images regions. For modeling the dynamics, series of images are used . The transformed images and binary images are then used to train CNN. These trained images are then used for guidancein the BLSTM neural network to learn temporal characteristics over lengthy and uneven time windows. The output of BLSTM neural network serves as final occurrences of an AU.

The facialimageryis preprocessed by self trackof facial landmarks and aligning face images. The location of facial landmarks is same as facial expression and hence suitable for face alignment. The set of facial points are then utilized to describe rectangular region that are chosen by the experts to aim ActionUnit. The average of facial point is considered midof area of fixed width w and height h. The rectangular region is then prune from original image f.

To compute binary image mask bi, the facial points are joined together setting all the values of fi inside a polygon to 1 and outside the polygon to 0.This is mainly used for better alignment and to automatically encode the different parts of face.

Temporal information can be encoded by extracting corresponding binary mask [bi] {bt−n, ..., bt+n} and image regions {ft−n, ..., ft+n} from 2n+1 consecutive frame at current frame t .The resulting sequence of image is transformed to a sequence A={Ai} and a sequence of binary mask images are transformed to a sequence S={Si}.The resulting A and S are used as an input to deep convolutional network. This dynamic encoding enable us to learn only short term temporal information.This transformation of image sequence by subtracting from current fame makes it easier to learn dynamics in a CNN network.

$$Ai = fi, \quad if\, i = tand \ldots\ldots\ldots(1)$$
$$fi - ft \quad otherwise$$

$$Si = bi, \quad if \quad i = tand \ldots\ldots\ldots(2)$$
$$bi - bt otherwise$$

Figure1 shows the image region A and binary mask S are the two input stream in CNN architecture. These are fed as an input to convolutional layer 1 (conv1) having 32 filters. Later it is allowed for max pooling which is of size 3*3*1. The outputs of both streams are then combined into a single stream. This combinationhelps in extracting the featuresby passing through 2 more convolution layers and 1 fully connected layer .

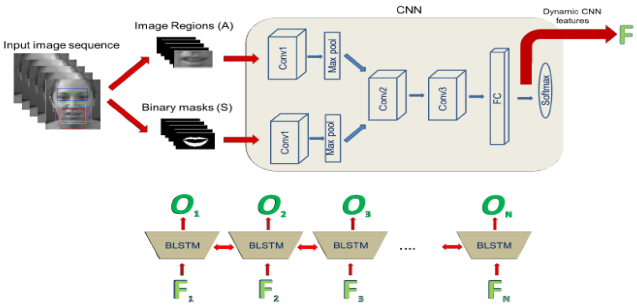

Figure1: A graphical overview of training pipeline

The initial convolutional level after merging in (Conv2) has 64 filters with size 5x5x64. The next convolution layer in the Conv3 consists of 128 filters of size 4x4x64. In completely connected layer (F C) the input from the other layer is compressed and send to change the output into number of classes as desired by the network. It includes3072 units and uses dropout with a probability of 0.2.The output level has two units consisting of positive classes and negative classes.The activation function used is Rectified Linear Unit (ReLU)

During CNN training, at conv1 layer only single feature map is computed at each filter.The network here is fully connected because Conv1 filters in the temporal direction is equal to size of A and S, and by using mini-batch gradient descent method it is trained with logarithmic loss function The trained information is then normalized.This will give a zero mean and standard deviation equal to 1 for each pixel.

IV. EXPERIMENTAL RESULTS

To evaluate various aspects of methods the various to compare performance , we conduct the various set of experiments which includes effectiveness of different CNN features, exploring effect of CNN architecture, finding effect of adding BLSTM and performances.

© Copyright: All rights reserved.

To find the effectiveness of different CNN features in training method, we use SEMAINE dataset to carryout task number of base line models are used. In baseline model one the input used is a complete picture of face defined by face bounding box and temporal window parameter n=0, Since n=0 there is no picture and no binary mask and does not use any temporal information. Second baseline method uses temporal information with window parameter n=2. In third baseline method same CNN architecture is used with input parameter n=2.Fourth baseline takes image region and binary shape mask as input and temporal window n=2.

Performance measures can be defined by Alternative forced choice (2AFC). Then comparison between $CRM_{n=2}$ and supplementary baseline on SEMAINE dataset is performed. Here we conclude that performance of $CRM_{n=2}$ is greater than $CR_{n=0}$ which prove that encoding temporal information leads to improved performance. During second set of experiment, we calculate the affect of accumulating max pooling (mp) level and varying dropped out factor (dp) in fully connected layer.

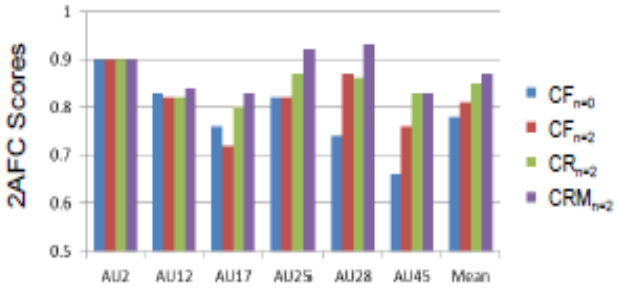

Figure2: Comparison of theprevious approaches and the proposed approach$CRM_{n=2}$.

We again use SEMAINE factor as dataset for performance evaluation and 2 AFC for performance measure. Also, the probability factor was increased from 0.2 to 0.5 and did not give any significant change. For calculating the effect of adding Bidirectional LSTM, we compute the performance of Bi-directional LSTM beside three other baseline with our proposed approach..

First baseline CRMLn=0 has not implemented time consideration since n=0 and for this reason doesnot use any dynamics.After CNN training,BLSTM is employed in second methodCRMLn=0. Third methodis CRMn=2 thatimplements a 2n+1=5 frames temporal window and not implements BLSTM.

The proposed work of CRMLn=0and CRMLn=0indicates BLSTM allows to know the dynamics without any temporal window which leads to improved performance. Similarly CRMn=2 improves the performance without using BLSTM.

Best result can be obtained with CRMLn=2with frames =5and employ BLSTM with feature extracted from CNN.

Table1: Performance (F1 score) evaluation on SEMAINE dataset

AU	LGBP	DLE	GDNN	$CRML_{n=2}$
2	0.75	0.66	0.67	0.80
12	0.63	0.76	0.63	0.74
25	0.40	0.61	0.77	0.32
28	0.01	0.26	0.31	0.32
45	0.21	0.35	0.55	0.85
17	0.07	0.25	0.14	0.33
Average	0.33	0.48	0.51	0.60

Table2: Performance (F1 score) evaluation on BP4D dataset

AU	LGBP	DLE	GDNN	$CRML_{n=2}$
1	0.18	0.25	0.33	0.28
2	0.16	0.17	0.25	0.28
4	0.22	0.28	0.21	0.34
6	0.67	0.73	0.64	0.70
7	0.75	0.78	0.79	0.78
10	0.80	0.80	0.80	0.81
12	0.79	0.78	0.78	0.78
14	0.67	0.62	0.68	0.75
15	0.14	0.35	0.19	0.20
17	0.24	0.38	0.28	0.36
23	0.24	0.44	0.33	0.41
Average	0.44	0.51	0.48	0.52

In the final conduct ofexperiment, we judge against performance of the proposedapproach with three existing approaches on SEMAINE and BP4D dataset.F1score is used as a performance measure tocompare with performance of literature survey. We also compared the performance using a deep learning technique GDNN in which we used four hidden layers t totrain the deep neural network.We conclude that performance of our approach is higher than SEMAINE dataset.Hence we are able to effectively find the spontaneous detection of facial expression with deep learning technique using convolutional and Bidirectional LSTM neural networks (CNN-BLSTM) that allows us to acquire knowledge about shape appearance and dynamics.

This combination of CNN-BLSTM excels the modeling of temporal information. FERA2015 dataset achieves the state of art.

© Copyright: All rights reserved.

V. CONCLUSION

In this paper we learnt the spontaneous detection of facial expression depending on various factors such as shape, appearance and dynamics.A new state of art is achieved on FERA2015 dataset by learning a new novel based approach called CNN-BLSTM which collectivelytrains shape appearance and dynamics. The dynamics features are taken from time window CNN and BLSTM, shape and appearance from local image regions and binary mask.

REFERENCES

1. A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis.Deep learning basedfacs action unit occurrence and intensity estimation. *In Facial Expression Recognition and Analysis Challenge, in conjunction with IEEE Int'l Conf. on Face and Gesture Recognition*, 2015.
2. M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In D. Cremers, I.Reid, H. Saito, and M.-H. Yang, editors, Computer Vision – ACCV 2014, volume 9006 of Lecture Notes in Computer Science, pages 143–157.Springer International Publishing, 2015.
3. B. Fasel. Head-pose invariant facial expressionrecognitionusing convolutional neural networks. In Multi-modalInterfaces, 2002. Proceedings. Fourth IEEE International Conference on, pages 529–534, 2002.
4. H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J.Kim.Deeptemporal appearance-geometr network for facial expression recognition.
5. M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep net--works for facial expression recognition. In Automatic Face and Gesture Recognition.FG 2013. IEEE, pages 1–6. IEEE, 2013.
6. M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. Deeply learning deformablefacial action parts model for dynamic expression analysis. In D. Cremers,Reid, H. Saito, and M.-H. Yang, editors, Computer Vision – ACCV 2014.
7. Yichuan Tang, "Deep learning using linear support vectormachines," in Workshop on Challenges in Representation Learning, ICML, 2013.
8. X. Xiong and F. De la Torre Frade. Supervised descent method and its applications to face alignment. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), May 2013
9. B. Fasel and J. Luettin. Automatic facial expression analysis:A survey. PATTERN RECOGNITION,
10. R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Richfeature hierarchies for accurate object detection andsema--nticsegmentation. 2014
11. T. R. Almaev and M. F. Valstar. Local gabor binary patterns from three orthogonal planesfor automatic facialExpression recognition. In Affective Computing and Intelligent Interaction,2013

AUTHORS PROFILE

Mrs. Nazmin Begum is a Research Scholar in Computer Science & Engineering at Visvesvaraya Technological University,Belgavi with research title" Multiconditional Joint Facial Action Unit Detection-Neural Networks". I have 8 years of teaching experience.Presently working as an Assistant Professor in Computer Science & Engineering in Dayananda Sagar University. I completed my Mtech from East West college of Engineering ,Bangalore in the year 2012.My area of interest includes Image Processing & Analysis,Facial Recognition, Pattern Recognition and Computer Vision.

Dr. Syed Mustafa A, holds an Engineering Bachelor's degree in Computer Science & Engineering from the prestigious Bangalore University (1999) , Master's degree in Computer Science & Engineering from Visvesvaraya Technological University (2005) and Doctorate degree in Computer Science & Engineering from Sathyabama University, Chennai. He currently works as Professor and Head of the Department at Department of ISE, HKBKCE. Strong leadership, team work, passion and commitment towards work are his traits. His area of research includes Web Services and Web Engineering. He had contributed to four articles in International and National peer-reviewed journals. Also he had presented four International and National conferences including IEEE international Conferences.

© Copyright: All rights reserved.