RESEARCH REVIEW JOURNALS
Volume-04
Issue-05
May-2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com[UGC Listed Journal]

Expression-Invariant Tied Factor Analysis for Joint Face and Expression Recognition

[1]Harish Kamat & [2]Dr. A C Subhajini

[1]Research Scholar, Sri Satya Sai University, Sehore M.P. (India)
[2]Research Guide, Sri Satya Sai University, Sehore M.P. (India)

ARTICLE DETAILS

Article History
Published Online: 15 May 2019

Keywords
CNN, LSTMs, Face Recognition.

ABSTRACT

A wide number of hand-drafted highlights, for example, and Gabor highlights perform well on customary face and expression databases, yet they accomplish generally more unfortunate exhibitions on face and expression in the wild databases. As of late, models dependent on profound learning, specifically profound CNN have been proposed which yield surprising execution in item and picture order undertakings. These profound CNN structures have the ability to use the exhibition of face and outward appearance recognition on information procured in nature. In any case, as pointed out in a broad survey, current CNN methodologies have basic lacks: models have depended on huge preparing information and a colossal number of parameters have need to have been educated.

1. Introduction

The point of the work in this section is twofold: Firstly, we mean to utilize profound learning techniques to improve face recognition and outward appearance recognition exactness on wild information. Besides, we likewise intend to diminish the quantity of parameters of a profound model, the quantity of preparing tests required, just as the required preparing time. To accomplish this, we propose to utilize LSTM engineering rather than profound CNNs for facial investigation. The LSTM is a particular RNN engineering and it performs incredibly well on a huge assortment of genuine applications, and is presently generally utilized in the PC vision network. LSTMs are adaptable models to deal with a variable-length successive information in PC vision applications with lower calculation cost. Moreover, they are an incredible asset for facial investigation with key clarifications of their capacity to catch successive examples. Besides, significant investigations have guaranteed that LSTMs are more compelling than ordinary CNNs for a few order assignments.

Focusing to improve the characterization execution of facial examination in the wild applications, we propose a different profound system learning model by consolidating LSTM systems. We make the accompanying noteworthy commitments in this section:

- The improvement of a LSTM model for video-based face confirmation in the wild that accomplishes check exactness that outflanks cutting edge results on the as of late presented testing face video database (YouTube faces) .The point of this commitment is to examine the consecutive examples of video-based faces in the wild through LSTM models;

- The advancement of a consolidated profound CNN model and LSTM model design to acquire improved unconstrained expression execution exhibited on the difficult outward appearance dataset. The point of this commitment is to break down the successive coherency of unconstrained static picture through LSTM models.

2. Feature Extraction Model

We follow Alex Net ne-tuning on top of FER spontaneous large dataset. Alex Net system is pre-trained deep CNN architecture which has been trained on over 1.2 million images. However to establish a recognition it is highly required to enrich the input data. This raises the requirement of using a feature descriptor to encapsulate the features of each significant point. In our work we used ne-tuned Alex Net features for dealing with this challenge. The aim of this ne-tuning scenario is to extract more advanced and significant expressive facial features from the input image. Since Alex Net features are based on the Image Net data, in order to obtain emotion specific features we ne-tuned Alex Net model on expressive data.

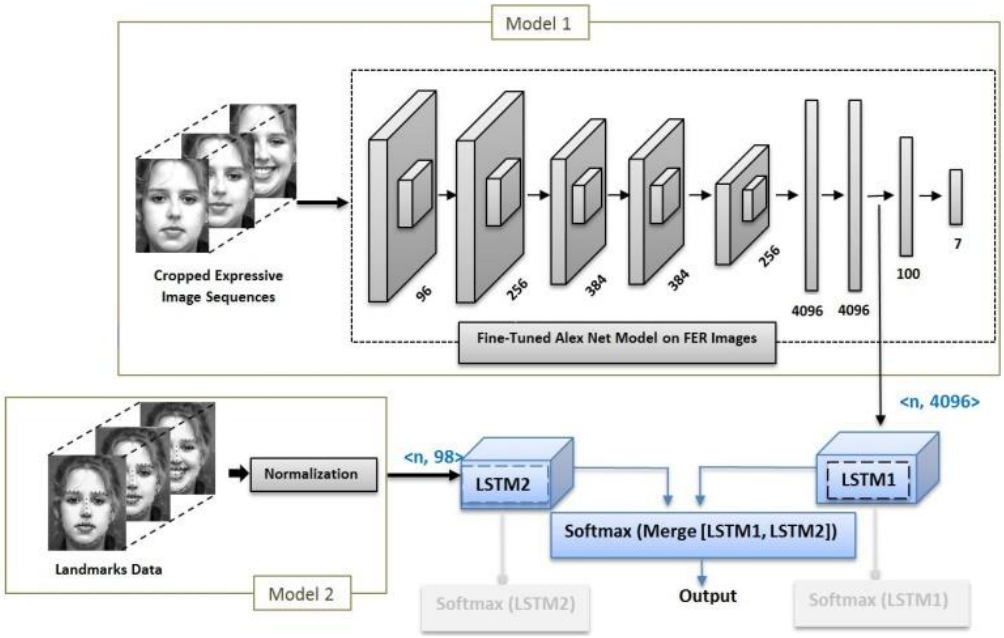

Figure 1: Our final model obtains an image sequence (Model 1) and landmarks data (Model 2) as inputs. Finally outputs of these two model pass through two separated LSTMs and then two LSTM models are concatenated using a merge layer concept. We use softmax for prediction.

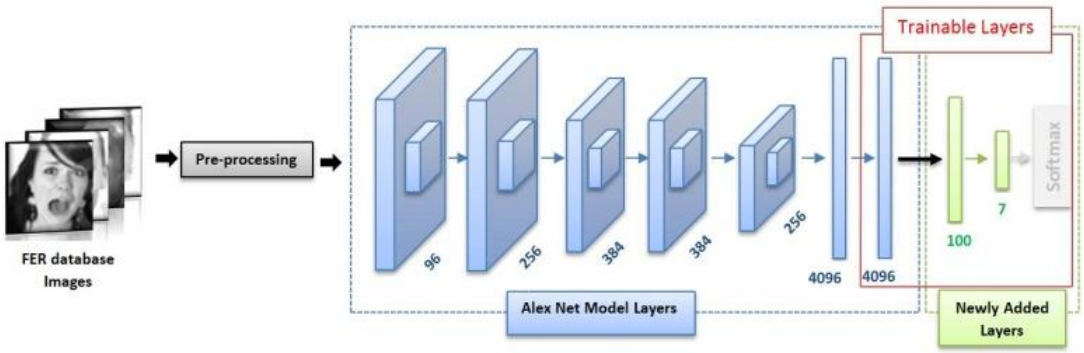

Figure 2: Pre-trained Alex net Model is ne-tuned on top of FER database images. Before this remove last two layers from Alex net model and add two new dense layers (last dense layer with softmax activation). The red color block represents the trainable layers of the model.

First each input image of FER dataset is pre-processed and resized to the size of 3 227 224. Next all training FER images are fed to pre-trained AlexNet model to ne-tune on top of FER images. In order to proceed this, remove third dense layer and softmax activation layer from the original AlexNet model and add dense layer (100, activation=relu), dropout (0.5), and dense layer (7, activation=softmax) as shown in Figure 4.16. We run experiments on the FER training set (28709 images), validation set (3589 images). We ne-tuned AlexNet model on FER data and achieved 0.72 accuracy with validation set Further, we compared ne-tuned AlexNet model result with other pre-trained CNN models (VGG19 and VGG16) in the state-of-the-art as well. For that, we also ne-tuned VGG19 and VGG16 models on FER data. Finally, ne-tuned AlexNet model is achieved the best performance for spontaneous facial expression recognition against the baseline methods and other pre-trained CNN models as shown in Table 1. As a result, as shown in Figure 1, features are extracted from the second fully connected layer of AlexNet model after ne-tuning.

Method	Accuracy
DLSVM [82]	0.694
MNL [94]	0.7
Ours (AlexNet)	0.72
Ours (VGG19)	0.684
Ours (VGG16)	0.681

Table 1: Performance comparison of validation set with the benchmarking approaches on FER2013.

3. Image Sequence-Based LSTM Model

LSTM (LSTM1) inputs the sequences of ne-tuned features of second fully con-nected layer as shown in Figure 4.2 to capture temporal changes of appearance.

We only stack 2 LSTM layers on top of each other and only rst LSTM layer returns its full output sequence. The output of the second LSTM layer is passed through the merge layer. We model our LSTM1 in simple manner, in order to reduce the number of parameters and to avoid the over fitting

4. Landmarks Data-Based LSTM Model

LSTM (LSTM2) receives the facial landmarks data points as input to capture temporal changes of geometry. We stack 2 LSTM layers on top of each other and only first LSTM layer returns its full output sequence. After the last LSTM layer, we add Batch Normalization in order to overcome over- tting with data. The output of the Batch Normalization layer is passed through the merge layer as shown in Figure 4.15.

Since xy coordinates are not normalized, before feeding landmarks data to the LSTM model, we follow the normalization process for these xy data points between (-1, 1) range. We de ne Landmarks data matrix of j^{th} image of i^{th} image sequence, L_{ij} as follows. Number of facial landmarks point is equal to p.

$$L_{ij} = [x^1{}_{ij} \quad y^1{}_{ij}; x^2{}_{ij} \quad y^2{}_{ij}; ...; x^p{}_{ij} \quad y^p{}_{ij}].$$

We normalized L_{ij} matrix as follows ans obtain normalized matrix, N orm_{ij}. Thevector representation of the normalized matrix is V_{ij}

$$Norm_{ij} = -1 + 2 * [L_{ij} - min(L_{ij}(:))]/[max(L_{ij}(:)) - min(L_{ij}(:))]$$

Normalized landmarks data output vector of i^{th} image sequence of t-frames, V_i can be explained as,

$V_i = (V_i^1; V_i^2; V_i^3; :::; V_i^t)$:

Then the LSTM output of i^{th} image sequence ($h^2{}_i$) can be de ned as,

$h^2{}_i$ = LST M(V_i):

5. Model Concatenation

The outputs from two LSTM models are passed through the merge layer as shown in Figure 4.15. Then two outputs are integrated using concatenation operation. These results can be explained as below.

$$M_i = Merge(concat[h^1{}_i, h^2{}_i]),$$

$$y_i = Softmax(M_i).$$

Where M_i is the output of the i^{th} image sequence after the merge operation and y_i is the output expression label of the i_{th} image sequence from the softmax classifier.

During the training process of our model, we use optimization function and categorical cross entropy loss function

To evaluate the effectiveness of our proposed approach, we conducted facial expression recognition on CK+, MMI, and BP4D databases. The below explains the datasets, model architecture and results.

6. Databases

- ***CK+***

The CK+ database has 593 image sequences from 123 subjects. We selected 327 image sequences from 100 subjects that can be labeled as one of six expression sequences: surprise, sad, fear, anger, disgust, and smile. We randomly divided all individuals into ten group and followed leave-one-group-out cross validation. All images were segmented from the background and processed through illumination normalization process, in order to keep same lighting condition. In CK+ database, all sequences start from a neutral expression and continue gradually to the apex

- ***MMI***

MMI database has 205 image sequences 30 subjects. This database is different from the other two databases. Here, all sequences start from a neutral expression, then move gradually to the apex, and end with the neutral expression. The 10-fold cross validation was performed when we followed the experiments with this dataset.

- ***BP4D***

This dynamic spontaneous facial expression database [95, 96] has 41 subjects with 8 expression labels. BP4D database is mainly focused on analysis of 3D model sequences. But here we only used its 2D texture sequences for testing purposes as our model is implemented using combination of 2D appearance features and geometry data. As CK+ database, this database also starts with the neutral and ends with the peak expression.

7. Conclusion

It demonstrated the capability of recognizing sequential patterns of facial expressions via two LSTM models with different feature sets. By avoiding over- tting, having reduced training data for expressions, exploiting strengths of LSTMs to model sequential temporal dependencies in expression data, and our model representation is developed. According to the model performance, it is clearly proven that appearance and geometry features can be separately modeled and captured through LSTMs for recognizing facial expressions. Since managing the ratio-temporal nature of the facial expressions through two different feature sequences is valuable fact within the computer vision applications, our contribution of this chapter is a new benchmark point.

References

1. Ming-Jung S., Valaparla D., and Asari V.K., 2003, Neural network based skin color model for face detection, Proceeding of Applied Imagery Pattern Recognition, pp.141–145.

2. Ti-Qiong X., Bi-Cheng Li, and Bo Wang, 2003, Face detection and recognition using neural network and hidden Markov models, IEEE International Conference on Neural Networks and Signal Processing, Vol.1, pp. 228–231.

3. Jaeger H., and Haas H., 2004, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, Science, Vol.304, pp.78-80.
4. Chen C., and Prakash E.C., 2005, Personalized cyberface: A novel facial modeling approach using multi-level radial basis function, IEEE International Conference on Cyberworlds, pp.474-482.
5. Bartlett M.S., Movellan J.R., and Sejnowski T.J., 2006, Face modeling by information maximization, IEEE Transactions on Neural Networks, Vol.13, Issue 6, pp.219-253
6. Bojkovic Z., and Samcovic A., 2006, Face detection approach in neural network based method for video surveillance, IEEE Neural Network Applications in Electrical Engineering, pp.44–47

