Learning Sequential Patterns Using Spatio-Temporal Attention Networks for Video Classification

Shubhada Deshmukh (shubhadadeshmukh@gmail.com)
Priyadarshini Institute of Engineering and Technology https://orcid.org/0000-0002-2587-3655
Manasi Patwardhan
TCS Research
Anjali Mahajan
Priyadarshini Institute of Engineering and Technology
Sadanand Deshpande
Priyadarshini Institute of Engineering and Technology

Research

Keywords: Attention Mechanism, Facial Expression Recognition, Video Processing

Posted Date: January 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-148673/v1

Learning sequential patterns using Spatio-temporal Attention networks for video classification

Shubhada P Deshmukh[1] ,Manasi S. Patwardhan [2] ,Anjali R. Mahajan[3] ,Sadanand B. Deshpande [4]

[1]*shubhadadeshmukh@gmail.com,Computer Science and Engg. Dept. PIET, Nagpur, MH, India.*

[2]*manasikelkar@gmail.com,TCS Research, Pune, MH, India.*

[3]*armahajan@gmail.com,Computer Science and Engg. Dept. PIET, Nagpur, MH, India.*

[4]*sbd119@gmail.com,Computer Science and Engg. Dept. PIET, Nagpur, MH, India.*

Abstract:

Extensive research effort has been focused on extracting temporal patterns from videos, to improve the accuracy of video classification using a deep neural network based approaches. In this paper, we show that long term dependency patterns may not be enough to achieve sufficient improved results. We propose the Attention-based Spatio-Temporal model (AST) for video classification, which is a self-attention model that learns to attend to spatial features using Convolutional Neural Network (CNN) and temporal features using attention mechanisms. We evaluate our model on motion dependent Action recognition (UCF-101) dataset, facial expression recognition (MMI) dataset, and micro-expression recognition (CASME2) dataset and generated real-life Facial Expression Recognition (FER) dataset and improved by 10%, 4.7% and 5.6% accuracy respectively as compared to state-of-art on the three standard datasets and a synthetic dataset as well.

In our research, we performed several experiments for detecting expressions and actions, the AST model plays a vital role in selecting the frames and carry the sequential context in the real-time application as well. We also experimented by extracting the features using the Active shape model (ASM) for FER and found the AST model surpasses other approaches.

Keywords: Attention Mechanism, Facial Expression Recognition, Video Processing

1. Introduction:

Deep neural network based algorithms have achieved great success in recent years as it can learn complex features and sequential patterns from video data using Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN). Automatic understanding of human behavior and its interaction with the environment have been an active research area in the last years due to its potential application in a variety of domains. To achieve such a challenging task, several research fields focus on human behavior under its multiple aspects (emotions, relational attitudes, actions, etc.)[1, 14, 16]. Facial Expression Recognition and Human Action Recognition are the most important problems that have received considerable attention from the computer vision community and pattern recognition in recent years.

Early research in facial expression recognition is majorly on still images by extracting spatial information [2]. Some studies [9, 10, 11, 12] tried to capture the temporal information by extracting

the geometric based features or deep neural network based algorithms like CNN for both, temporal and spatial based information [3] from a video. Traditional facial expression recognition in machine learning involves phases such as pre-processing, feature extraction and classification as shown in Fig. 1. There are many algorithms to extract facial features like the Gabor feature [4], local binary pattern (LBP) [4] and ASM [5] for emotion detection. In the machine learning approach, a set of features or attributes of each instance is extracted whereas in deep neural network based algorithms, raw data is stated to the model for the classification task. However, with deep neural network based methods, the algorithm extracts the features of the data automatically. In contrast, a set of features, requiring further processes such as feature selection and extraction, (such as PCA) in case of machine learning algorithms.

Figure 1: Typical Architecture of Facial Expression Recognition (FER)

The field of action recognition is one of the research fields attracting recent attention. It has varied applications in the field of video surveillance, human-computer interaction (HCI), healthcare, and sports analysis [6]. In recent years, there has been an increasing attention in deep learning models, in which it can learn the feature representation automatically by multiple layers of neural network and can build sequential representations using sequential models of the raw input.

Our focus is to apply the attention mechanism for capturing the context of sequential patterns from the motion dependent data. We have considered the dynamic data as a video of facial expressions and action recognition. The facial features in the video change dynamically and our task is to identify the pattern within the video for an emotion classification task. Since deep neural network-based algorithms have been proven to be effective at learning and generalizing data with high-dimensional feature spaces like the video [7], our approach is to identify the most contributing set of facial features in sequence for expression recognition (How?). The state-of-the-art approach has used CNN for extracting features and the set of features are input to the LSTMs [8] for detecting facial expression. We evaluate the proposed model for facial expression recognition on the standard dataset, namely CASME2 and MMI and action recognition tasks on a benchmark dataset, namely UCF101.

The proposed system resulted in performance gain over traditional baseline system by 10%, 5.6% and 4.7% accuracy on UCF101, CASME2 and MMI datasets. In this paper, we have designed the architectures to generalize the algorithm and identify the most contributing frames using the self-attention mechanism [9]. The proposed architecture achieves better accuracy as compared to state-of-art on the three standard datasets and a real-life dataset as well. The experimental results show that our method outperforms the traditional baseline systems.

The other part of the paper is arranged as follows. First, we present our motivation towards our research problem along with the similar work done in this area. Next in Section 2, we present our proposed architecture and algorithm for capturing context information followed by implementation details in Section 3. Finally, results and discussion are presented in Section 4 and conclusion at the last.

Motivation:
Video surveillance cameras are capable of capturing images and videos. Video surveillance is the process of monitoring behavior and activities of people using video cameras. Cameras are constantly capturing images of a person or person's face on cell phones, webcams, etc. Applications tracking user's faces and human behavior by a video camera are growing exponentially.
Spatiotemporal data mining refers to the process of discovering patterns and knowledge from spatiotemporal data. Among many kinds of spatiotemporal data, video data (i.e., data about facial expression or human action) are especially important. The field video classification is a vital research topic focused on sequential long term dependencies.

Our goal is to automatically learn sequential patterns in long term dependencies that occur in motion dependent applications. As per our knowledge, we observed in motion dependent real-time application there is no algorithm to detect the context of sequential patterns, in general, using an attention mechanism. We hypothesize that alone self-attention mechanisms can improve the results for sequential long term dependencies. In our paper, LSTM could be replaced Attention mechanism by averaging networks using the context vector. The self-attention model stores all previous representations in memory. Using LSTM rather is inefficient as it stores the set of features of every frame in a video, most of the feature vector of frames in video does not change frame-to-frame, so storing each frame representation is a waste of memory. The attention unit is preventing correlated data to be stored.

2. Related Work

Previous efforts have used deep neural network based algorithms to analyze emotion from image datasets like JAFFE and CK+ database [10]. In the paper [11], authors investigated for facial expression recognition using DCNN features and resulted in 97.08% on JAFFE. The paper [9] is about the multimodal approach for video-based emotion recognition in the wild. These features include deep convolutional neural network (CNN) based features obtained via transfer learning. The CNN pre-training and fine-tuning stages are carried out before processing of the target emotional video corpus. The CNN-LSTM approach [12] can predict the expression with partial sequences of expression and learned Spatio-temporal features with very good accuracy.
The paper [30] combines the approach of spatial and temporal aspects using CNN pre-trained model on a public dataset of facial images with (1) a spatial attention mechanism, to select the most vital parts of the face for a given emotion, and (2) temporal softmax pooling. The approach in the paper was to compute the softmax on the class probabilities and the video frames jointly to select the most prominent frames of the given video. Our approach was to select the most contributing frames in real-time with respect to the current frame using the context vector in attention mechanism. In paper [13], LSTM show gradually promising outcomes for an application like Forecasting Financial Time Series pursues to assess if attention mechanisms can further improve performance. The paper proposed that attention can help to prevent long-term dependencies experienced by LSTM models and tested using LSTM with attention.
Likewise, previous work shows challenging research carried in Human Action Recognition with deep neural network based algorithms as well. Authors in the paper [14] learn to focus on spatial and temporal parts of the video frames and classifies videos with the deep model of multi-layered Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units. The model

essentially learns which parts in the frames are significant for the task at hand and are higher importance to them with 84.96% accuracy. The paper [15] learn video feature vector using neural networks with long-term temporal convolutions (LTC). They demonstrate that LTC-CNN models improve the accuracy of action recognition by 95.4%. Experimental results show that Action recognition task on the dataset, namely UCF101 for shuttleNet, which is composed of both feedforward and feedback connections, outperforms remarkably as compared to LSTMs and GRUs by simply adding shuttleNet into a CNN-RNN network. The paper [16] proposes a model in which the deep visual features are extracted using Long-term Recurrent Convolutional Network (LRCN) that can learn to recognize and create temporal features for tasks involving sequential data. In the paper [31], the author has proposed the Spatio-temporal attention mechanism by extracting the features through CNN and applied the attention model on every frame to select the important feature map and then this final feature map is given as an input to the ConvLSTM which results in 87.11% accuracy. Our model has extracted the context features using attention mechanisms in real-time and improved the accuracy by 10.4%.

3. Method- Attention-based Spatio-Temporal model

The proposed architectures shown in Figure 2, detect the context behind sequential patterns to detect emotions in a video using different deep neural network based algorithms. The state of art [17, 18] showed that the architectures of CNN with Long Short-Term Memory (LSTM) can improve recognition performances compared to conventional CNN. Convolutional and pooling layers can act as effective feature extractors, which are able to take out local features from different parts of an image. The spatial features are extracted with the CNN layer and self-attention mechanism to improve the performance in terms of accuracy.

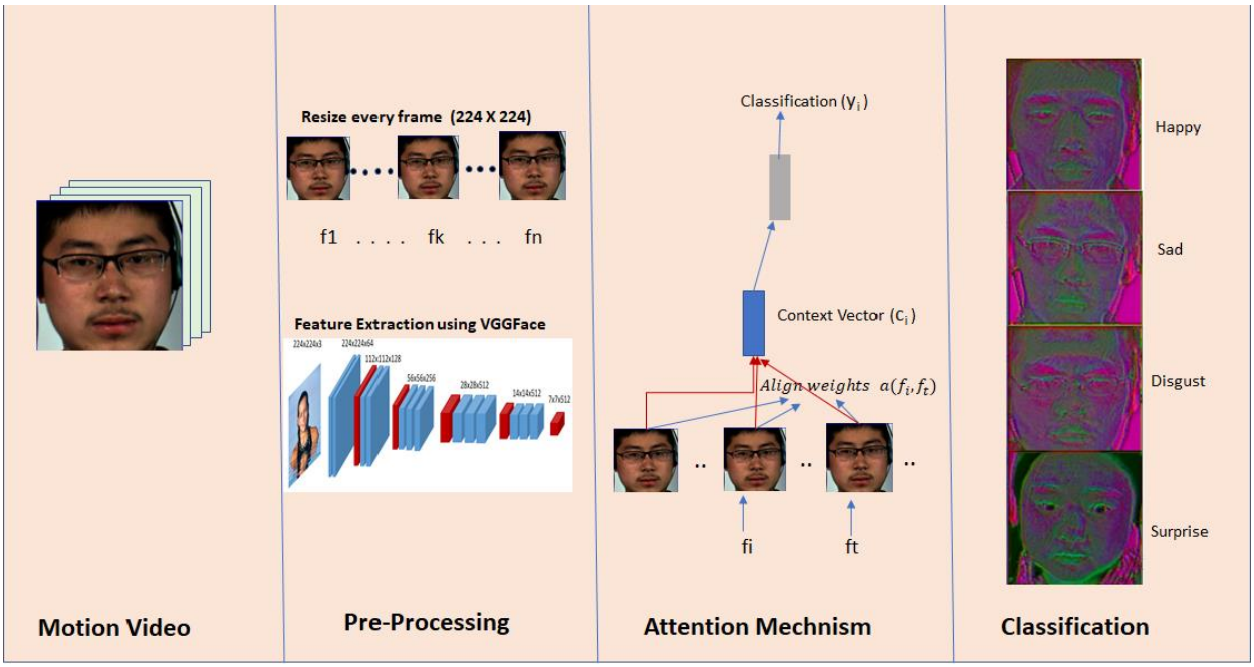


Figure 2: Proposed Architecture using VGGFace and Attention Mechanism for FER

Feature Map:

1. VGGFace feature map

Convolution Neural Network: CNN is a biologically-inspired model and firstly proposed by LeCun et al [19] Shown in Figure 3 is a general structure of a CNN.

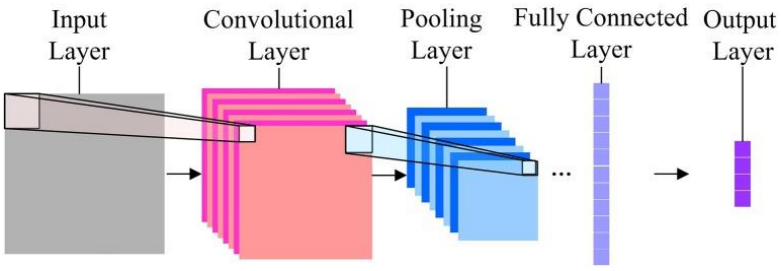


Figure 3: The General structure of a CNN [13].

In Figure3, our proposed architecture input layer receives normalized images with identical size as 224X224. A set of units in the input layer will be processed by a convolution kernel to form the unit in a feature map of the subsequent convolutional layer.

Transfer learning method was applied to reduce the risk of overfitting [15], many studies used pre-trained model (e.g., AlexNet [20], VGG [21] and VGG-face [22]) to train deep neural networks with comparatively little data. The VGG-Face network [21] is composed of 5 convolutional/pooling and three fully connected layers and was trained for face recognition using 2.6 million images of 2622 celebrities. With such a large amount of face data, VGG-Face's convolutional layers are expected to produce suitable features for faces out of the box [23].

2. Active Shape Model (ASM) for facial features

Face detection in the video is the first step for emotion recognition in the traditional system. Face in the video frames is detected by using the Haar classifier [24]. If the face is located, the computational complexity of the facial feature extraction can be significantly reduced. In this approach, the facial features are extracted using the Active Shape Model (ASM). ASM [25] is a geometric based algorithm and used to extract 77 facial features. For emotion recognition, using pair-wise action point distances our algorithm generates facial feature vector consisting of (77X77)/ 2) -77 features per video frame, avoiding self and redundant distances.

Temporal Self-Attention Mechanism:

The attention mechanism has been proposed to improve the performance of machine translation in [26] as a sequence-to-sequence model. The mechanism makes the decoder not only depend on the context vector from encoder but allows the decoder model to learn relevant parts over source sequence to predict the target. In our approach encoder reads a sequence of input vectors as features

of images over input time t = 1 . . . N as $f_i = (f_1, \ldots, f_N)$, into a "context vector" at encoder c_i. The context vector here represents the stored state and information of the encoder.

The context vector c_i depends on a sequence of annotations $(f_1, \cdots, f_N)$ to which an encoder maps the input images. Each annotation c_i contains information about the whole input sequence with a strong focus on the parts surrounding the i^{th} image of the input sequence.

The context vector c_i is, then, computed as a weighted sum of these annotations f_i:

$$c_i = \sum_{i=1}^{t} \alpha_{ti} f_i$$

The weight α_{ti} of each annotation f_i is computed by

$$\alpha_{ti} = \frac{\exp\,(e_{ti})}{\sum_{i=1}^{t} \exp\,(e_{ti})}$$

Where $e_{ti} = a(f_i, f_t)$

is an alignment model which, scores how well the inputs around position i and the query frame at position t match, where i < t.

Proposed Algorithm Flow

Initially the sequence of frames are extracted with the size 224 X 224 in the motion dependent video. Then from the frame, features are extracted using VGGFace pre-trained model with feature vector 7 X 7 X 512 of each image and reads a sequence of input vectors as features of images over input time t = 1 . . . N as $f_i = (f_1, \ldots, f_N)$.

Proposed model is applied to the sequence of the feature vector. Self-attention by averaging networks is influenced by a context vector and the model stores all previous representation in memory and gives a global context vector. The context of every previous frame is computed with the query frame i, where i is current frame over time t = 1 . . . N. Classifier with a fully connected layer and a softmax activation function is used for classification. Based on the experimental results we have analyzed that the proposed method is better in performance compared with the state-of-the-art.

Implementation details

In this section, we will introduce the detail of datasets and their corresponding evaluation schemes. Then, we will give details of the implementation of our model. We finally show the experimental results by comparing our approach, AST with other approaches (LSTM and GlobalAvgPooling) with different feature extraction methods. We also compared our results with the state-of-the-art methods to demonstrate its superior performance.

CASME II (micro-expressions dataset) contains 247 spontaneous micro-expressions from 26 subjects, categorized into five classes: happiness (33 samples), disgust (60 samples), surprise (25 samples), repression (27 samples) and others (102 samples). The micro-expressions are recorded at 200 fps in well-controlled laboratory environment [27]. MMI dataset includes 29 videos for 6 basic facial expressions: neutral (3 samples), smile (8 samples), yawn (4 samples), sleep (4 samples), surprise (6 samples) and angry (4 samples). The expression sequences were recorded at a temporal resolution of 24 fps [28]. The UCF101 dataset is one of the most popular action recognition benchmarks. It contains 13,320 video clips (27 hours in total) from 101 action classes and there are at least 100 video clips for each class [29]. The real-life dataset (FER) is generated for various facial gestures. Each expression sequence of the dataset was labeled with one of the six basic expression classes (i.e., angry, disgust, fear, happy, sad, and surprise). We have experimented on 69 videos for four classes: neutral (17 samples), smile (17 samples), sleep (18 samples) and yawn (17 samples).

Our implementation is based on Keras using Python requires higher memory since it needs to process all the frames and find out most contributing frames using the attention model. In this case, we used a VM optimized for performance and has 112 GB RAM. With a lesser configuration like 16 GB of RAM, we can process only 20 videos with intermediate frames. All the frames in the videos datasets were resized to 224×224 resolution and fed to a VGG-Face model trained on the ImageNet dataset. The last convolutional layer of size $7 \times 7 \times 512$ was used as an input to our AST model. For both training and testing our model takes 20 frames at a time sampled at fixed fps rates. We compute class predictions for each for the entire video clip.

4. Results and discussion:

We have experimented on three standard datasets (i.e. UCF-101, CASMEII, and MMI) and one real-life generated dataset. Finally, we compared the AST, VGGFace-LSTM and VGGFace-GlobalAvgPooling approaches with the state-of-the-art algorithm on UCF-101, CASMEII, MMI and Generated dataset. The attention-based model (AST) beats the LSTM model in both applications. The results are summarized in Table 1.

Table 1: Accuracy of the different approaches with features extracted from VGG-Face (pre-trained model)

Method	UCF-101	MMI	CASMEII	FER
AST	97.56	83.33	66.6	95.16
LSTM	97.56	83.33	59.02	85.46
GlobalAVGpooling	95.12	50	58.2	90.91
State-of art	87.11	78.61	60.98	NA

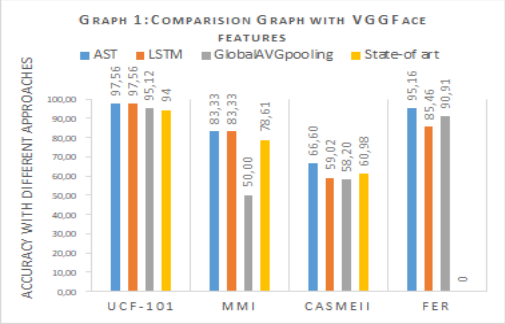
In Table 2, the attention-based model is compared with the LSTM approach with different input feature vector using the ASM algorithm. The Attention-based model showed that LSTM can be replaced by averaging networks with a context vector in Graph 2, which shows our model is independent of application and feature map. Using LSTM rather is inefficient as it stores the

feature vector of every frame in a video, most times the feature vector does not change frame-to-frame immediately.

Table 2: Accuracy of the different approaches with facial features extracted from ASM algorithm

Method	MMI	CASMEII	FER
Attention Model	66.6	90.79	35.71
LSTM	33.33	73.68	28.57

The attention unit in our model is preventing correlated data to be stored and improves the accuracy as compared to the LSTM model as shown in Graph 1 and Graph 2. In the Graph 1, the features of the frames are extracted using VGG-Face model and all the three approaches are compared. In Graph 2, the facial features are extracted using the ASM algorithm and compared with other facial expression recognition dataset. The attention unit in our model outperforms in terms of accuracy with the LSTM approach and baseline method, which attains our hypothesis to be true.



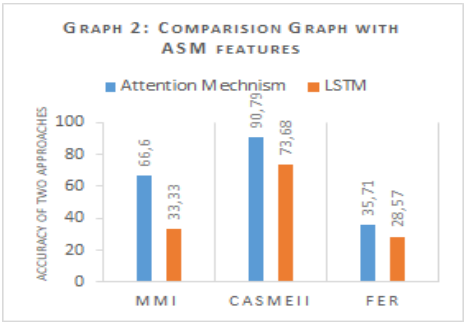


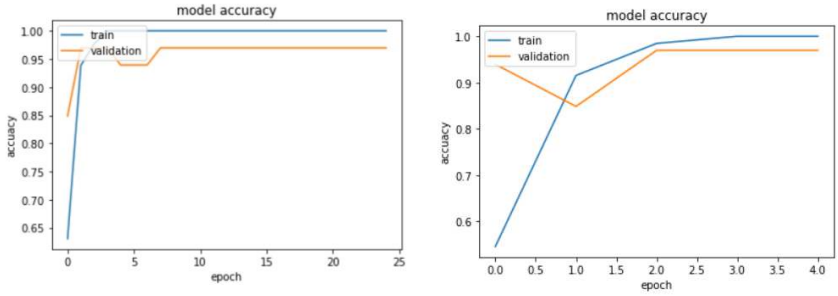
Table 3: Accuracy of the different approaches with features extracted from the CNN layer.

Method	UCF-101	MMI	CASMEII	FER
Attention Model	95.24	33.33	41.18	78.57
LSTM	95.24	33.33	41.94	71.4

Table 3, illustrates a comparison of the Attention mechanism and LSTM with input as a frame with size 224 X 224 and extracted features with the CNN layer followed by attention unit or LSTM network. The accuracy is less as simple CNN layer gives insufficient spatial information compared with the pre-trained model. In this case, also the attention unit carries the context of the sequential pattern and improve the accuracy as compared with LSTM.

Graph 4: Train and Validation Accuracy w.r.t. sequential AST Architecture

Graph 5: Train and validation accuracy w.r.t. VGGFace-LSTM architecture



Graph 4 and Graph 5 depict the comparable between AST and LSTM model on the action recognition dataset. We can see that the AST model learns the sequential context and fine-tunes the parameter in very few iterations, unlike the LSTM model.

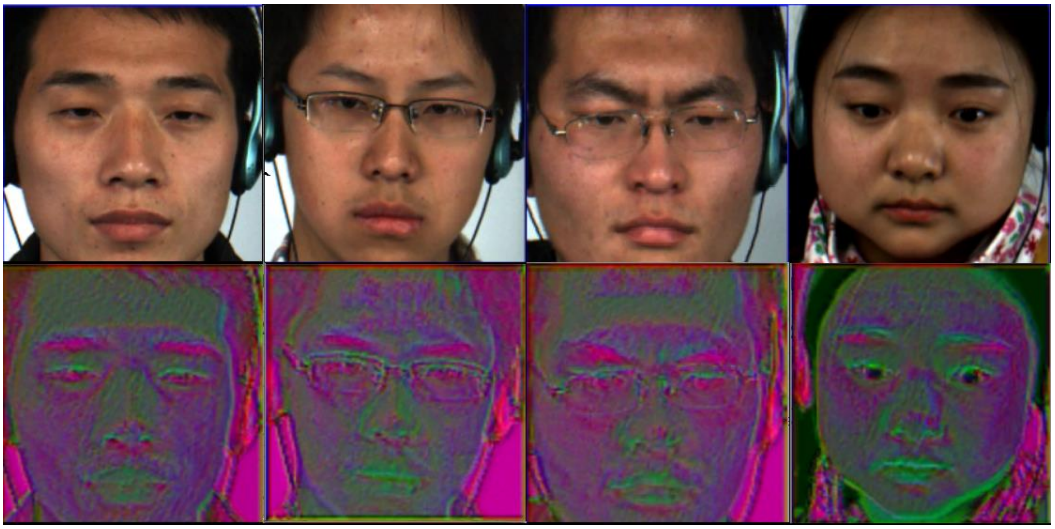


Figure 4: Heatmap visualization of the AST model Happiness, Sadness, Disgust and Surprise for CASMEII dataset

Figure 4 describes the heatmap of the AST model and shows the prominent features extracted from attention mechanism, like lip corners , the eyebrows, area near the eyebrows and eyes with respect to class Happiness, Sadness, Disgust and Surprise are best part on the face for classifications.

Conclusion:

We found that our model AST is learning the context of long term dependencies is one of the robust indications for emotion recognition and action recognition. We presented the general architecture for extracting features in various methods like ASM, VGG-face model from the image which gives spatial information by selecting the indicative parts of frames in a video and attain temporal information by considering the context of that frames for the recognition task. Our experiments in the result section have shown that our model achieves better accuracy compared to the LSTM model and State-of-art.

5. Abbreviations:

FER: Facial Expression Recognition

AST: Attention Spatio-Temporal

ASM: Active Shape Model

RNNs: Recurrent Neural Networks

LSTM: Long Short-Term Memory

CNN: Convolutional Neural Networks

DCNN: Deep Convolutional Neural Networks

6. Declarations section

Availability of data and materials:

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests:

The authors declare that they have no competing interests.

Funding: Not applicable

Authors' contributions:

Our contributions in this paper were that the first author (Shubhada Deshmukh) participated in the research, designing of the scheme, code design, the experiments and drafted the manuscript. The second author (Manasi Patwardhan) participated in research and designing of the scheme. All authors read and approved the final manuscript.

Acknowledgements:

Thanks to the My guide, second author, (Manasi Patwardhan) for reviewing and for her constructive suggestions to help improving this paper.

Authors' information:

Shubhada Deshmukh received the B.E. degree in computer engineering from SRTM University, and the master's degree of computer science from Pune University, India. She is doing Ph.D. and

has submitted the Ph.D. thesis at Nagpur University, India. She is a Data Scientist at Ericsson, Sweden. She has published 2 papers in journal and conference, and she has been funded for her research by AICTE, India during 2015-2018.

Manasi Patwardhan received the B.E. degree in computer engineering from Pune University, India, and the master's degree of computer science from University of Tulsa, Unites States. She received the Ph.D. degree from University of Tulsa, Unites States. She is a Senior Scientist at TCS Research, Pune, India. She has published more than 40 papers in journals and conferences.

Anjali Mahajan received the Ph.D. degree from Nagpur University, Nagpur, India. She is a HOD in Govt. Polytechnic college, Nagpur, India. She has published more than 60 papers in journals and conferences.

Sadanand B. Deshpande received the Ph.D. degree from Nagpur University, Nagpur, India. She is a HOD in Govt. Polytechnic college, Nagpur, India. She has published more than 5 papers in journals and conferences.

7. References:

1. Moez B., Franck M., Christian W., Christophe G., and Atilla B.: Sequential Deep Learning for Human Action Recognition", *Int. Conf. Human Behavior Unterstanding*, pp: 29-39 (2011).
2. Byoung, C. K.: A Brief Review of Facial Emotion Recognition Based on Visual Information, *Sensors*, 18(2), 401 (2018).
3. Xianzhang, P., Wenping, G., Xiaoying, G., Wenshu, L., Junjie, X., Jinzhao, W.: Deep Temporal–Spatial Aggregation for Video-Based Facial Expression Recognition, *Symmetry — Open Access Journal*, (2019).
4. Happy, S.L., George, A., Routray, A.: A real time facial expression classification system using local binary patterns, In: *IHCI*, pp:1–5 (2012)
5. Shbib, R., Zhou, S.:Facial expression analysis using active shape model ,Int. J. Signal Process. Image Process. Pattern Recognit. 8, (1), pp: 9–22 (2015).
6. Shugang Z., Zhiqiang W., Jie N., Lei H., Shuang W., and Zhen L.: A Review on Human Activity Recognition Using Vision-Based Method, *Int. J. Healthcare Engg.*, pp:31 (2017)
7. Chan, W. L., Kyu, Y. S., Jihoon, J., Woo, Y. C.: Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data, *Workshop on Human Multimodal Language* (2018).
8. Essentials of Deep Learning: Introduction to Long Short-Term Memory: *https://www.analyticsvidhya.com/*
9. Heysem, K., Furkan, G., Albert, A. S.: Video-based emotion recognition in the wild using deep transfer learning and score fusion, *Image and Vision Computing*, pp: 66-75 (2017).

10. Abir, F., Lotfi, A., Ali, D.: Facial expression recognition via deep learning, *Computer Systems and Applications* (2017).
11. Veena, M., Radhika, M. P., Manohara, P. M. M.: Automatic Facial Expression Recognition Using DCNN, *International Conf. on Advances in Computing & Communications*, pp: 453 – 461 (2016).
12. Wissam, J. B., Yong, M. R.: Learning Spatio-temporal Features with Partial Expression Sequences for on-the-Fly Prediction, *Association for the Advancement of Artificial Intelligence* (2018).
13. Thomas, H., Antoine, V., Seung, E. Y.: A Comparison of LSTMs and Attention Mechanisms for Forecasting Financial Time Series, In: *arXiv* (2018).
14. Shikhar, S., Ryan, K., Ruslan, S.: Action Recognition using Visual Attention, *Neural Information Processing Systems (NIPS) Time Series Workshop* (2015).
15. Yemin, S., Yonghong, T., Yaowei, W., Wei, Z., Tiejun, H.: Learning long-term dependencies for action recognition with a biologically-inspired deep network, In: *ICCV* (2017).
16. Gul, V., Ivan, L., Cordelia, S.: Long-term Temporal Convolutions for Action Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6), pp: 1510 – 1517(2018).
17. B., Allaerta, IM., Bilascoa, C., Djerabaa: Advanced local motion patterns for macro and micro facial expression recognition, *Pattern Recognition* (2018).
18. Dae, H. K., Wissam, B., Jinhyeok, J., Yong, M. R.: Multi-Objective based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition, *IEEE Transactions on Affective Computing* (2017).
19. Yemin, S., Yonghong, T., Yaowei, W., Wei, Z., Tiejun, H.: Learning long-term dependencies for action recognition with a biologically-inspired deep network, In: *ICCV* (2017).
20. Alex, K., Ilya, S., Geoffrey, E. H., Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, pp: 1097–1105 (2012).
21. Karen, S., Andrew, Z.: Very deep convolutional networks for large-scale image recognition, In: *arXiv preprint arXiv:1409.1556* (2014).
22. Omkar, M. P., Andrea, V., Andrew, Z.: Deep face recognition, In: *BMVC*, 1(3), pp: 6 (2015).
23. Justus S., Esam G., Enrique H., Stylianos A.: High-performance and lightweight real-time deep face emotion recognition, *Int. workshop on Semantic and Social Media Adaptation and Personalization (SMAP)* (2017).
24. Wilson, P.I., Fernandez, J.: Facial feature detection using HAAR classifier, Journal of Comput. Sci. Colleges, 21(4), pp: 127–133 (2006).
25. Shbib, R., Zhou, S.: Facial expression analysis using active shape model, *Signal Process. Image Process. Pattern Recognit.*, 8(1), pp: 9–22 (2015).
26. Dzmitry, B., KyungHyun C., Yoshua, B.: Neural machine translation by jointly learning to align and translate, In: *ICLR* (2015).

27. Wen-Jing, Y., Xiaobai, L., Su-Jing, W., Guoying Z., Yong-Jin, L., Yu-Hsin, C., Xiaolan, Fu.: CASME II: An improved spontaneous micro-expression database and the baseline evaluation, *PLoS One*, 9(1), p. e86041 (2014).
28. M. Pantic, M. Valstar, R. Rademaker, and L. Maat: Web-based database for facial expression analysis, In: *ICME*, pp: 317-321 (2005).
29. Yemin, S., Yonghong, T., Yaowei, W., Wei, Z., Tiejun, H.:Learning long-term dependencies for action recognition with a biologically-inspired deep network, *Computer Vision and Pattern Recognition*, pp: 4321-4330 (2017).
30. Masih Aminbeidokhti, Marco Pedersoli, Patrick Cardinal, and Eric Granger : Emotion Recognition with Spatial Attention and Temporal Softmax Pooling, *Image Analysis and Recognition*, pp: 323-331(2019).
31. Lili Meng, Bo Zhao, Bo Chang, Gao Huang: Interpretable Spatio-temporal Attention for Video Action Recognition, *Computer Vision and Pattern Recognition*, last revised 3 Jun 2019.

Figures

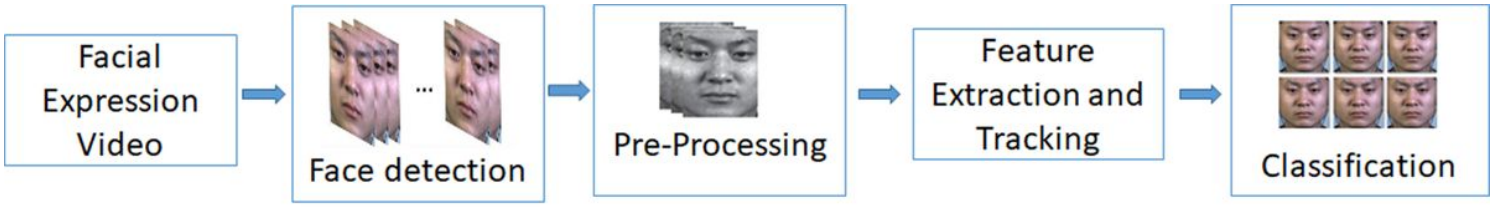


Figure 1

Typical Architecture of Facial Expression Recognition (FER)

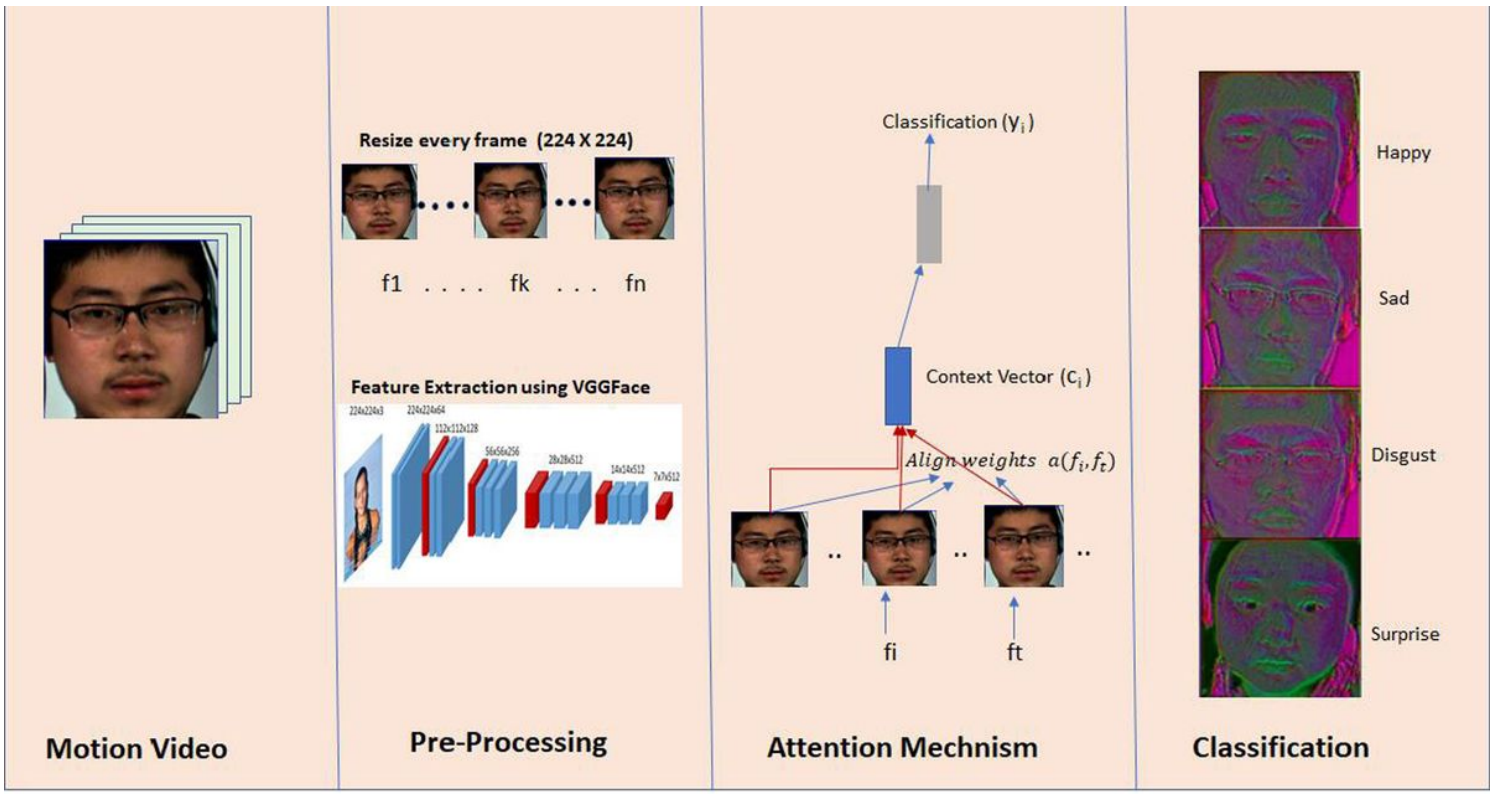


Figure 2

Proposed Architecture using VGGFace and Attention Mechanism for FER

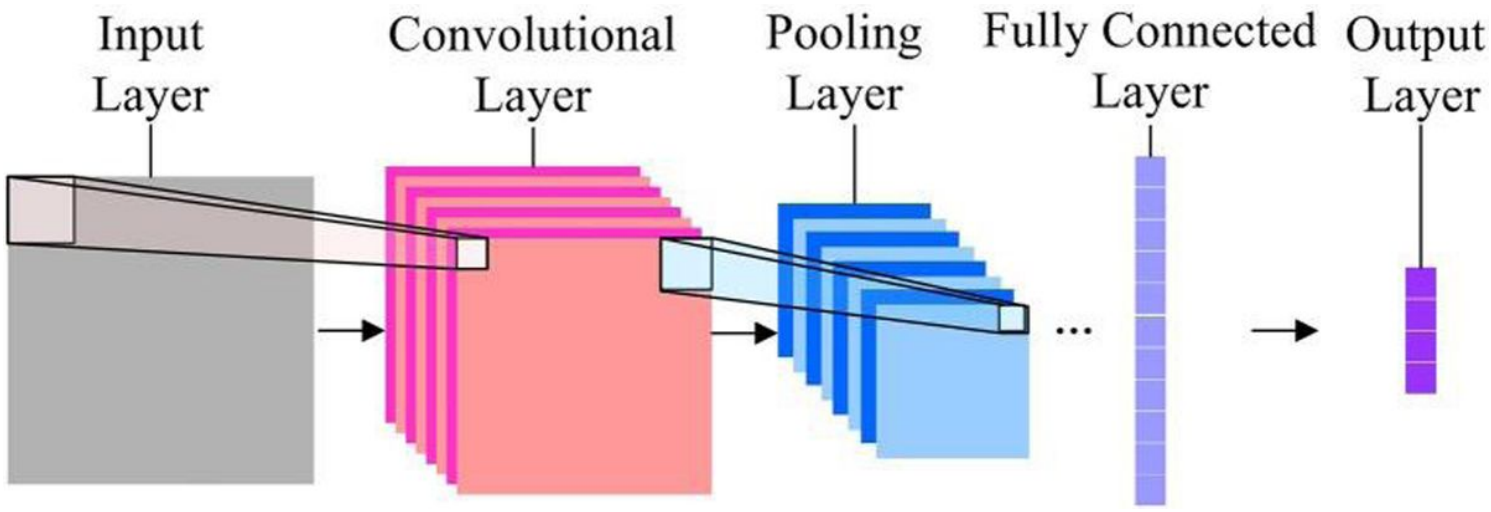


Figure 3

The General structure of a CNN [13].

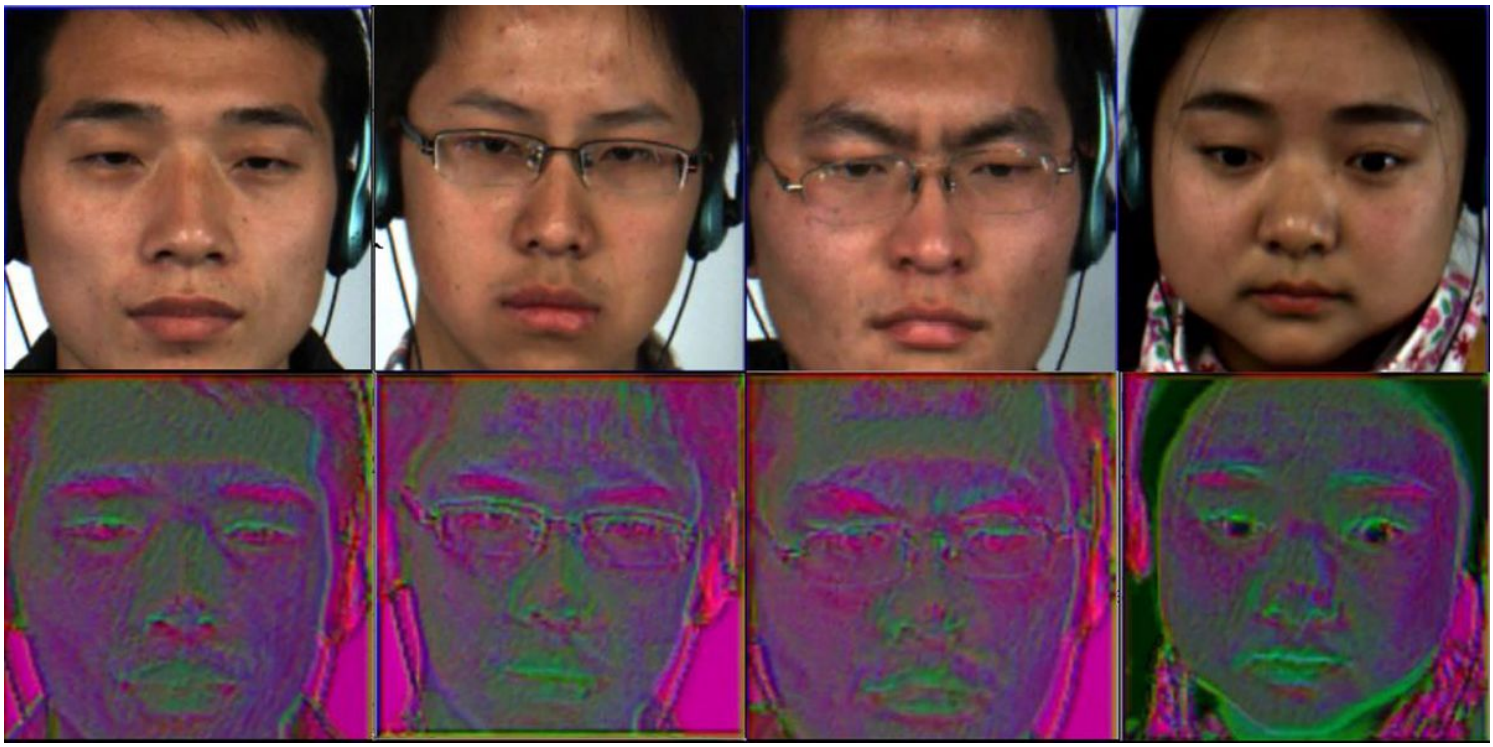


Figure 4

Heatmap visualization of the AST model Happiness, Sadness, Disgust and Surprise for CASMEII dataset