Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 13, Issue 1, January 2022: 460-472

Non-intrusive Stress detection based on temporal emotion analysis in videos applying machine learning

[1]Fasiha Anjum Ansari , [2]Prasadu Peddi

[1]PhD Research Scholar, JJT University Jhunjhunu Rajasthan 333001

[2]Professor, CSE Department, JJT University Jhunjhunu Rajasthan 333001

Abstract—Stress has become an embedded part of daily life in modern work environment. Continuous exposure to stress affects physical and mental health of an individual. Early detection and treatment of stress is the only way to avoid long term health issues. This work proposes a non intrusive method for stress detection from videos captured in work environment. A model for correlating stress to the facial and posture clues is proposed. Visual features extracted from facial clues and posture features are mapped to emotion state using machine learning classification. From the sequence of emotion observed over a period of time, stress level is classified using deep learning LSTM classifier.

I. INTRODUCTION

Stress is the response of human to the physical and emotion events affecting him. Modern work environment has become very stressful due to gap between the expectations placed on the employees and their perceived ability to meet these expectations. Both physical and mental health is affected due to prolonged and rhythmic exposure to stress. It can result in abnormal cardiac rhythm, depression etc. Recent research, have found that stress also affect the human immune system leading to cancer. A survey [1] by American institute of stress inferred that 80% of employees experience job stress and nearly half of them needed assistance to manage the workplace stress. Early detection and medication is the solution to manage work place stress.

Questionnaire based stress evaluation is the conventional method adopted by physiologists. These questionnaires are qualitative, time consuming, conducted onsite and their answers are subjective. Emotional arousal does not necessarily reach the level of consciousness and the extent to which it does and whether one wants to share this information can vary from person to person. Therefore, the answers of employees to these questionnaires are not always a good representation of their wellbeing in the organization. A possible solution to this problem is provided in terms of biochemical and physiological indicators. But these methods are intrusive and the methods themselves become a source of stress. Research has focused on finding objective, non-intrusive, continuous and quantitative ways to detect stress.

This work aims to detect stress in non intrusive manner with observation being continuous and result being quantitative. To achieve this, this work uses the workplace videos. Nowadays most of work premises are monitored using surveillance cameras. The video feeds from these surveillance cameras are used for stress detection. In this work, the facial and posture clues captured from the surveillance videos are mapped to emotional state applying machine learning classification. From the sequence of

temporal emotional shifts, stress level of the person is detected. Following are, some of the contributions in this work

1. Analysis of a correlation model mapping facial and posture features to emotional state
2. A machine learning model mapping the visual facial features and salient posture features to emotion state.
3. A deep learning model to map the sequence of temporal emotional state of individual to the objective stress level.

II. RELATED WORK

Zhang et al (2020) proposed a two level stress detection framework. Face and action level representation is learnt from videos and both are fused in a weighted vector for stress detection. But the stress assessment is snapshot based and does not consider the spatial and temporal variations in the facial cues for stress detection. Han et al (2018) detected stress from speech signals using deep learning. Mel filter bank coefficients are extracted from the speech signals. These features are provided as input to long short term memory (LSTM) for stress classification. The accuracy of stress detection in this approach is about 66.4%.

Yogesh et al (2017), extracted glottal and speech features from the speech signal and used it for classification of stress. Mean shift clustering method is used for detection of stress. Wristband physiological signals are used for stress detection by Sevil et al (2017). A significant change in blood volute pulse and skin temperature was observed in response to stress. This work used this observation to design a stress detection system. Wrist band signals were collected using Empatica E4 wrist band and signal energy value is thresholded to detect the stress. Giannakakis et al (2017) designed a stress detection system using the facial cues considering both non voluntary and semi voluntary actions. Features extracted from eye movements, mouth activity, head motion and heart rate are used for stress classification. Feature ranking using ranking transformation is done to improve stress detection accuracy. Pampouchidou et al (2016) analyzed the correlation between the characteristics of mouth activity and stress. Template matching with Eigen method is applied to extract mouth features. Among different mouth features, normalized opening rate per minute and opening intensity have higher correlation to the stress. Gao et al (2014) proposed a real time non-intrusive method for detecting the emotional state of the driver from the facial expressions. Discrete cosine transform (DCT) features and SIFT feature extractor are extracted from the face. SVM classifier is used to classify stress. Viegas et al (2018) used 17 facial action units with binary classifier to detect stress. Facial action units were taken from each frame of the video captured during different activities of the person. The facial action units were found to have strong correlation to the stress level of the person. Gavrilescu et al (2019) used facial expressions represented facial coding system for stress detection. They output of the system was a stress scale called as determine Depression anxiety stress scale (DASS). A three layer architecture based on machine learning was proposed in this work. Active appearance model with multi class SVM is used at first layer. Action unit features are constructed at second layer. Neural network is used at third layer to classify action features to stress levels. Prasetio et al (2018) detected stress from features of facial frontal image. Facial features around of pair of eyes, nose and mouth is extracted using DoG, HOG and DWT. The features are then classified using depth learning ConvNet to detect the stress. A facial feature based

stress recognition system is proposed by Tamura et al 2018. Gabor filter and HOG features are extracted from the region of interest around pair of eyes, nose and mouth. Features were classified to stress level using a hybrid SVM classifier. Facial features related to eye and mouth region are used to classify stress by Pediaditis et al 2015. Decision tree classifier was used to classify stress level from the facial features. McDuff et al (2016) used photoplethysmorgrpahic signals for stress detection.

Various Heart rate related features are extracted from the photoplethysmorgrpahic signals and used for detection of stress. Among multiple parameters, HRV is found to have higher correlation to stress. Deep learning multi modal stress detection system was proposed by Bara et al (2020). Convolutional auto-encoder and recurrent neural network are used to classify stress from multimodal input feeds of thermal images, physiological measurements and audio signals. A stress detection system using physiological signals was proposed by Can 2019. Smart wearable devices were used to collect the physiological signals in a non intrusive manner. Measurements were made during daily life routine of individuals. Heart activity, skin conductance and accelerometer signals features are used as input to machine learning classifiers to detect stress. Stress and boredom during playing of games is detected using facial cues in [17]. Seven facial features are extracted from 68 different landmarks. Euclidean distance between the features is used as landmark.

The facial features around mouth and eye regions are found to have higher correlation to the stress. Authors in [18] surveyed the use of body posture features for stress detection. The study documented the correlation between the various postures and stress.

III. PROPOSED SOLUTION

The proposed solution for stress detection involves following stages

1. Selecting key frames from the video
2. Feature extraction from the frames of interest
3. Mapping features to emotional state
4. Classification of stress level

The architecture of the proposed solution is given in Figure 1.

A. Frame of interest selection

The workplace environment is captured and provided as input for stress detection. Not all the frames in the video are significant for stress detection, so it is necessary to select the relevant frames in the video. The relevant frame selection involves following steps

1. Reading the frames from the video
2. Detect the presence of Face or upper body using Voila Jones detector
3. If Face or upper body is not detected, skip the frame
4. For the not skipped frame , compare the frame difference with previous salient frame
5. If the difference is less than threshold , skip the frame
6. If the different is greater than threshold, mark the frame as salient frame and pass it to next stages.

Threshold factor is to be configured by the user. When the threshold value is less, too many frames

are marked as relevant and computation time increases. When threshold is set high, some of emotional states can be missed and it affects the accuracy of stress calculation. The best value for threshold is found through trial and error.

B. Feature Extraction

On each relevant frame, facial and body posture features are extracted. The proposed solution extracts following features

1. Euclidean Facial features
2. Scalable invariant feature transform (SIFT) features in body
3. Deep features from body parts

Euclidean Facial Features

Euclidean facial features are extracted from the facial landmarks as given below

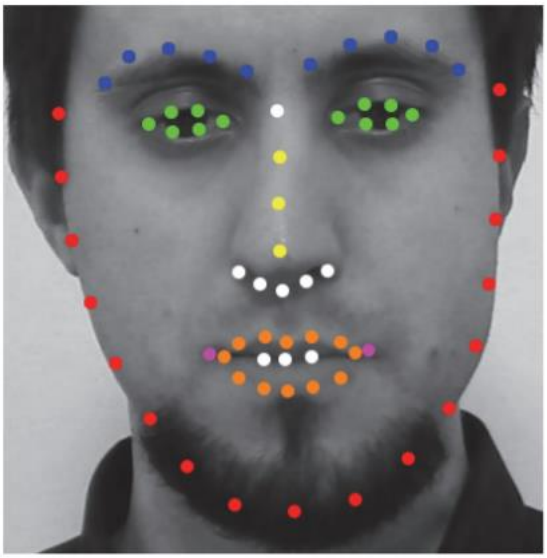


Figure 1 Facial landmarks

Following seven features are extracted in terms of Euclidean distance.

F1	Sum of the Euclidean distance between the mouth contour landmarks and the anchor landmarks. I
F2	Sum of the Euclidean distance between the mouth corner landmarks and the anchor landmarks.
F3	Area of the regions bounded by the closed curves formed by the landmarks in contour of the eyes
F4	Sum of the Euclidean distance between eyebrow landmarks and the anchor landmarks
F5	Area of the region bounded by the closed polygon formed by the most external detected landmarks
F6	Average value of the Euclidean norm of a set of landmarks in the last frames. It describes the total distance the head has moved in any direction in a short period of time.
F7	Average value of all detected landmarks. It describes the overall movement of all facial landmarks

SIFT Features from body

Previous study [18] has found a relationship between emotion state and body postures. The relationship between upper body parts and the emotion is summarized below

Fear	Leg and arm crossing. Movement. Hand or arms clenched.
Anger	Body spread. Closed hands or clenched fists.
Sadness	Body dropped, shrunken body, bowed shoulders, trunk leaning forward
Happiness	Arms open, eye contact relaxed and lengthened.
Disgust	Hand covering the neck, one hand on the mouth, one hand up, hands close to the body, orientation towards side, hands covering the head.

From the body postures, SIFT features are extracted and mapped to emotions.

Deep features from face and body parts

Deep features are extracted using a modified VGG 16 convolutional neural network

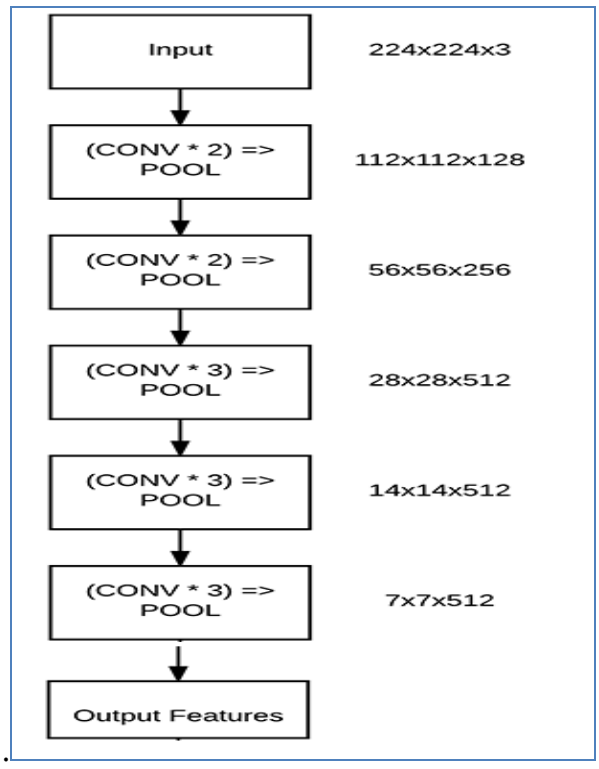


Figure 2 VGG Net for Feature extraction

When a image is passed to modified VGG16 network, features are extracted and collected at the activation layer. The output volume of 7 x 7 x 512 is flattened to a feature vector of dimension 21,055.

From the key frames, human upper body is segmented using Voila Jones. The segmented part is passed to VGG16 network to extract the features.

C. Emotion state mapping

The features extracted from face and body posture is mapped to one of following emotion labels using machine learning classifiers of happiness, sad, surprise, fear, disgust and anger.

The emotion is classified from features using multi class SVM.

D. Stress level classification

From the sequence of emotion labels observed over a period of time, stress level must be classified. Long Short Term Memory (LSTM) is used in this work for stress level classification from the sequence of emotion labels.

LSTM is an extension of RNN (Recurrent Neural Networks) with gating mechanisms to allow learning and forget learnt information. The structure of LSTM is given below

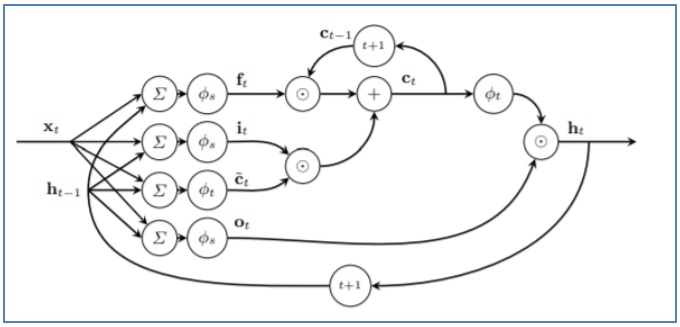


Figure 3 LSTM Structure

An LSTM node take the input vector x and the last hidden state information as inputs for processing. It calculates the cell activation c as weighted sum of inputs and bias b. Hyperbolic tangent activation function is applied onto the weighted sum as below

$$c_t = \emptyset_t(W_c x_t + U_c h_{t-1} + b_c)$$

c_t is the candidate cell activation. x_t is the input vector. W and U are the weight matrices. h_{t-1} is the hidden state vector at the previous time step and b_c is the bias. The level of actions to be retained or forgot is controlled by the input gates. Hidden state is calculated by the final gate.

$$f_t = \emptyset_s(W_f x_t + U_f h_{t-1} + b_f)$$

$$i_t = \emptyset_s(W_i x_t + U_i h_{t-1} + b_i)$$

$$o_t = \emptyset_s(W_o x_t + U_o h_{t-1} + b_o)$$

f_t is the forgot gate vector. i_t is the input gate vector. o_t is the output gate vector.

In this work, LSTM is used to predict the stress label based on past sequence of emotion labels and the current observation of emotion label. The dataset can be represented as $X = \{x_1, x_2, \ldots x_N\}$ where N is the total number of events reported. It is multivariate irregular time series data and each event x_k consists of sequence of emotion labels. $x_k = \{x_k^1, x_k^2, .. x_k^T\}$ where x_k^t represent the emotion label at time t. For each x_k , there is event label $y_k = \{y_k^1, y_k^2, .. y_k^T\}$. The value of y_k^t is one of K values of event labels. In this work LSTM is trained to predict the event label at time t+1 called y_k^{t+1} based on the past observations till time t. We design event label prediction framework based on LSTM. The design is given below

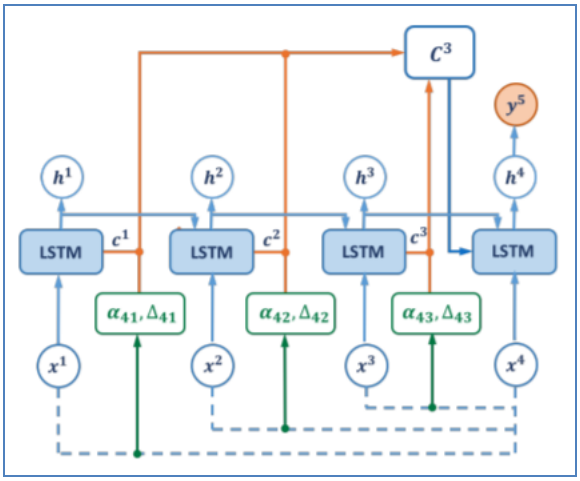


Figure 4 Stress Classification Framework

IV. NOVELTY IN PROPOSED SOLUTION

The proposed solution is different from existing solutions is following aspects

1. An integrated model combining facial and posture features is proposed for modeling the emotion state of an individual
2. Frame of interest selection algorithm is proposed to select the frames with high correlation to stress assessment from the videos.
3. Different from snapshot based stress classification, deep learning LSTM based continuous stress evaluation from temporal variations in emotional state is proposed.

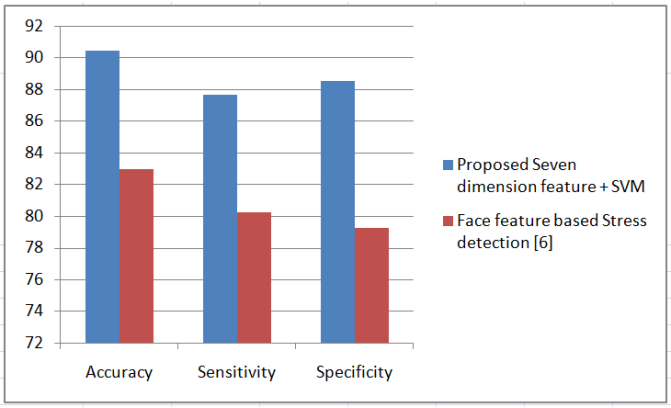
V. RESULTS

The dataset for the testing the proposed solution was collected from IT companies in Manyata Tech Business park located in Bangalore. A total of 122 volunteers were selected

body postures, a set of 100 images were manually tagged with emotions as per the procedure given in [20] and this 100 images are used as training dataset for emotion classification from body postures. Accuracy, sensitivity, Specificity were the parameters used measuring the effectiveness of emotion classification.

Proposed facial features of seven dimensions is extracted and classified to emotions using multi class SVM classifier. The performance is compared to “Facial features based stress detection” approach proposed in [6].The result is given below

Method	Accuracy	Sensitivity	Specificity
Proposed Seven dimension feature + SVM	90.23	87.19	88.25
Face feature based Stress detection [6]	82.18	80.17	79.12

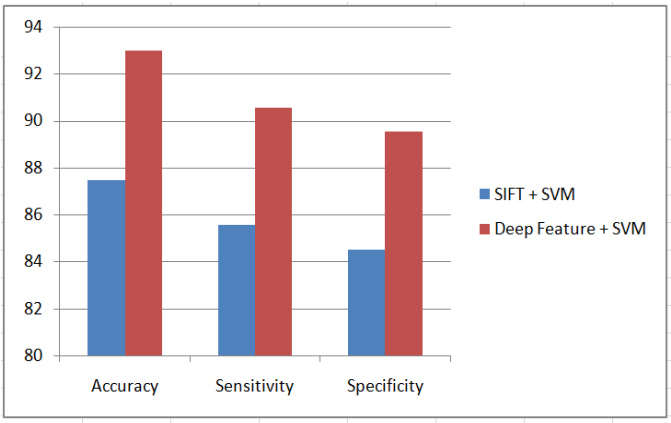


The results are for the study with 58 males and 64 females in age group of 25 to 45. CCTV video feeds were collected from the working sites of these 122 volunteers. All the volunteers signed the participant consent form before the study. Video feeds of these volunteers were collected every day for a total duration of 1 month. The stress levels of these volunteers were measured using standard questionnaire analysis every week once and at the end of the study. Out of 122 volunteers, 80 were used as training dataset for stress classification and rest 42 is used as test dataset.

For the case of the emotion classification from face image, Extended Cohn-Kanade dataset [19] was used. For the case of emotion classification from accuracy in the proposed solution is higher than [6] by 7.48%. The sensitivity in proposed solution is higher by 7.42% compared to [6]. The specificity is the proposed solution is higher by 9.27% compared to [6].

With SIFT and deep features extracted from upper body, the classifier accuracy is measured and the result is given below

Method	Accuracy	Sensitivity	Specificity
SIFT + SVM	87.45	85.57	84.50
Deep Feature + SVM	92.97	90.55	89.54



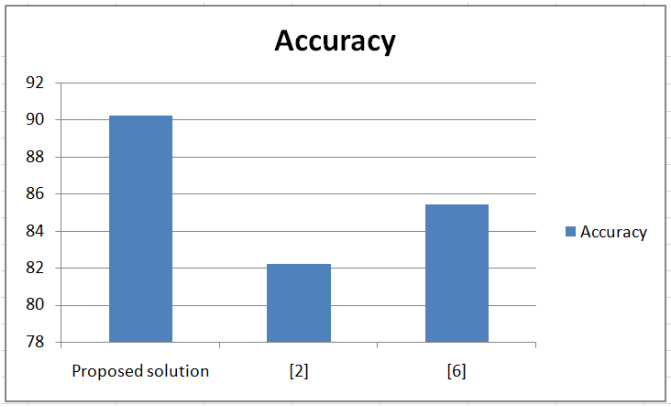
Compared to SIFT+SVM, deep features with SVM have higher accuracy of 5.52%. The sensitivity in the Deep Feature with SVM is higher than SIFT with SVM by 4.98%. The specificity in the Deep Feature with SVM is higher than SIFT with SVM by 5.04%. Deep feature in combination with SVM Classifier has higher accuracy compared to SIFT feature with SVM Classifier.

The performance of proposed LSTM based Stress level classification system is compared with that of Facial features based stress detection proposed in [6] and Deep learning based stress detection proposed in [2]. The performance is measured in terms of following metrics

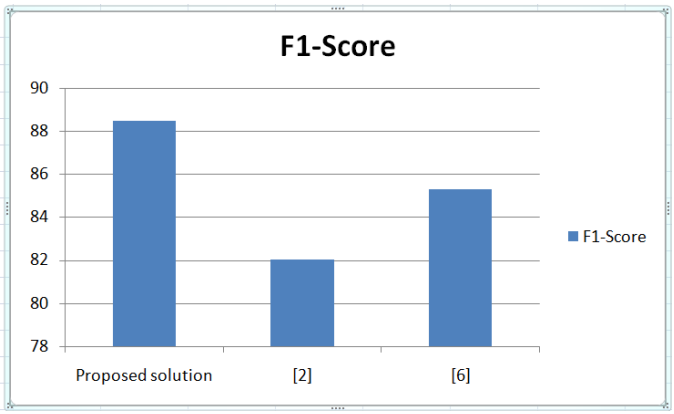
1. Accuracy
2. F1-Score
3. Precision
4. Recall
5. Classification time

Method	**Accuracy**	**F1-Score**	**Precision**	**Recall**	**Classification time (seconds)**
Proposed solution	90.2	88.48	91.21	87.24	47
[6]	82.20	82.04	82.05	82.31	48
[2]	85.42	85.28	85.32	85.53	87

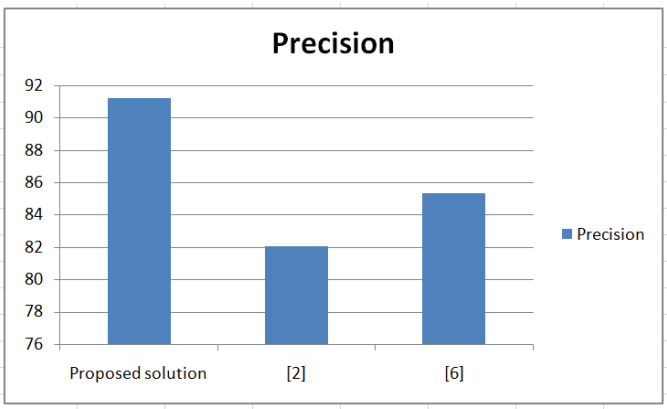
The accuracy in proposed solution is higher than [6] by 8% and higher than [2] by 4.78%.



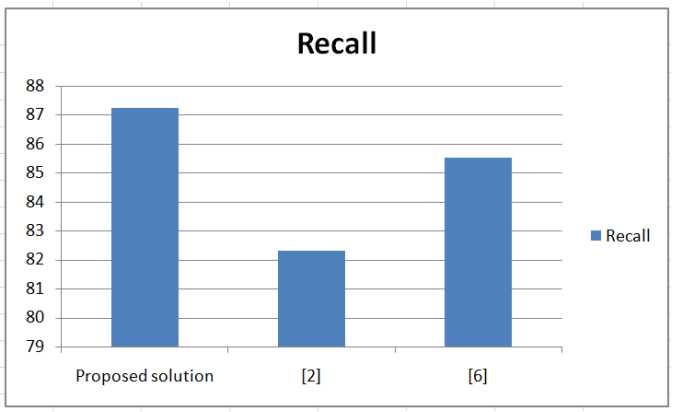
The F1-score in the proposed solution is higher than [6] by 6.44% and higher than [2] by 3.2%.



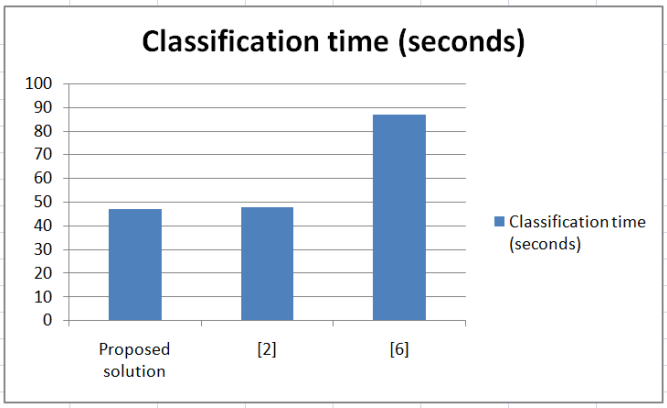
The precision in proposed solution is higher than [2] by 9.16% and higher than [6] by 5.89%



The recall in the proposed solution is higher than [2] by 4.93% and higher than [6] by 1.71%.



The classification time taken in proposed solution is lower than [2] by 2.1% and lower than [6] by 85.10%.



The accuracy of the proposed integrated facial and upper body posture features is higher than existing solutions by minimum 4.78% with peak accuracy of 90.2%. The reason for the higher accuracy in the proposed solution is due to use of integrated features of facial and upper body postures. Deep learning based features from upper body postures performs better than SIFT based features due to inherent learning ability of Deep learning.

The proposed solution has reduced the classification time by 85.10 % in comparison to deep learning approach. This is due to use of deep learning only at the feature extraction stage.

VI. CONCLUSION

A novel stress level classification system is proposed in this research work. An integrated feature combining facial features and upper body postures is used for emotion classification. Seven novel features are extracted from 64 different landmarks in the face image. Upper body postures are extracted using Resnet based deep learning method. Based on the sequence of emotions over a continuous period of time, LSTM is used in this work for stress level classification. The proposed solution achieved an accuracy of 90.2% for stress level classification with an increment of 4.78% over the existing solution.

REFERENCES

1. Global Organization for Stress on stress facts. http://www.gostress.com/stress-facts. Accessed:

2020-27-02.
2. Zhang, H., Feng, L., Li, N., Jin, Z., & Cao, L. (2020). Video-Based Stress Detection through Deep Learning. Sensors (Basel, Switzerland), 20(19), 5552. https://doi.org/10.3390/s20195552
3. Han, H.; Byun, K.; Kang, H. A deep learning-based stress detection algorithm with speech signal. In Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia, Seoul, Korea, 22–26 October 2018; pp. 11–15.
4. Yogesh, C.; Hariharan, M.; Yuvaraj, R.; Ruzelita, N.; Adom, A.; Sazali, Y.; Kemal, P. Bispectral features and mean shift clustering for stress and emotion recognition from natural speech. Comput. Electr. Eng. 2017, 62, 676–C691.
5. Sevil, M.; Hajizadeh, I.; Samadi, S.; Feng, J.; Lazaro, C.; Frantz, N.; Yu, X.; Br, T.R.; Maloney, Z.; Cinar, A. Social and competition stress detection with wristband physiological signals. In Proceedings of the 2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN), Eindhoven, The Netherlands, 9–12 May 2017; pp. 39–42
6. Giannakakis, G.; Pediaditis, M.; Manousos, D.; Kazantzaki, E.; Chiarugi, F.; Simos, P.G.; Marias, K.; Tsiknakis, M. Stress and anxiety detection using facial cues from videos. Biomed. Signal Process. Control. 2017, 31, 89–101
7. Pampouchidou, A.; Pediaditis, M.; Chiarugi, F.; Marias, K.; Simos, P.; Yang, F.; Meriaudeau, F.; Tsiknakis, M. Automated characterization of mouth activity for stress and anxiety assessment. In Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST), Chania, Greece, 4–6 October 2016; pp. 356–361.
8. Gao, H.; Yüce, A.; Thiran, J. Detecting emotional stress from facial expressions for driving safety. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 5961–5965
9. Viegas, C.; Lau, S.; Maxion, R.; Hauptmann, A. Towards Independent Stress Detection: A Dependent Model Using Facial Action Units. In Proceedings of the International Conference on Content-Based Multimedia Indexing (CBMI), La Rochelle, France, 4–6 September 2018; pp. 1–6
10. Gavrilescu, M.; Vizireanu, N. Predicting Depression, Anxiety, and Stress Levels from Videos Using the Facial Action Coding System. Sensors 2019, 19, 3693
11. Prasetio, B.H.; Tamura, H.; Tanno, K. The Facial Stress Recognition Based on Multi-histogram Features and Convolutional Neural Network. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 7–10 October 2018; pp. 881–887
12. Prasetio, B.H.; Tamura, H.; Tanno, K. Support Vector Slant Binary Tree Architecture for Facial Stress Recognition Based on Gabor and HOG Feature. In Proceedings of the 2018 International Workshop on Big Data and Information Security (IWBIS), Jakarta, Indonesia, 12–13 May 2018; pp. 63–68.
13. Pediaditis, M.; Giannakakis, G.; Chiarugi, F.; Manousos, D.; Pampouchidou, A.; Christinaki, E.; Iatraki, G.; Kazantzaki, E.; Simos, P.G.; Marias, K.; et al. Extraction of facial features as indicators of stress and anxiety. In Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 3711–3714
14. McDuff, D.J.; Hernandez, J.; Gontarek, S.; Picard, R.W. Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 7–12 May 2016;

pp. 4000–4004

15. Bara, Cristian-Paul & Papakostas, Michalis & Mihalcea, Rada. (2020). A Deep Learning Approach Towards Multimodal Stress Detection.
16. Can, Yekta & Chalabianloo, Niaz & Ekiz, Deniz & Ersoy, Cem. (2019). Continuous Stress Detection Using Wearable Sensors in Real Life: Algorithmic Programming Contest Case Study. Sensors. 19. 10.3390/s19081849.
17. Fernando Bevilacqua, Henrik Engström, Per Backlund, "Automated Analysis of Facial Cues from Videos as a Potential Method for Differentiating Stress and Boredom of Players in Games", International Journal of Computer Games Technology, vol. 2018, Article ID 8734540, 14 pages, 2018. https://doi.org/10.1155/2018/8734540
18. H. Zacharatos, C. Gatzoulis, and Y. L. Chrysanthou, "Automatic emotion recognition based on body movement analysis: A survey," IEEE Computer Graphics and Applications, vol. 34, no. 6, article no. 106, pp. 35–45, 2014
19. Lucey, Patrick & Cohn, Jeffrey & Kanade, Takeo & Saragih, Jason & Ambadar, Zara & Matthews, Iain. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010. 94 - 101. 10.1109/CVPRW.2010.5543262.
20. C. Corneanu, et al.,"Survey on Emotional Body Gesture Recognition" in IEEE Transactions on Affective Computing, vol. , no. 01, pp. 1-1, 5555. doi: 10.1109/TAFFC.2018.2874986