Journal of Engineering and Applied Sciences 14 (4): 1286-1294, 2019
ISSN: 1816-949X
© Medwell Journals, 2019

Automatic Recognition of Cognitive States Using Multimodal Approaches in e-Learning Environments

[1]H.S. Gunavathi and [2]M. Siddappa
[1]Department of CS and E, Jain University, Bengaluru, India
[2]Department of CS and E, SSIT, Tumkur, India

Abstract: Cognitive state recognition is one of the active science researches all around the world and it has grown spontaneously in recent years. However, most research focuses on posed expressions, near frontal recordings and they ignore eye gaze, head pose and considers hand occlusions as noise. It makes tough to tell how the existing methods perform underneath conditions where faces appear in a wide range of poses and occluded by hands. In this study, we propose multimodal approaches for building a real-time cognitive state recognition system in e-Learning environments by integrating hand-over-face gesture with facial expression. Our proposed system performs an average recognition rate of 90.51% with 15.8 fps is robust to variations in facial expressions, hand shapes and occlusions.

Key words: Cognitive states, compressive sensing, facial expressions, hand-over face gesture, robust, hand

INTRODUCTION

Emotion recognition is an estimation tool to recognize the human desires and emotions for predicting the present state of mind and future interactions. Besides, various emotions affect the behaviour patterns in the human beings and have to be recognized. The emotions of a person are depicted through the facial expressions and body languages that explain the cognitive states of mind and psychometric aspects of a person (Fernandes and Bala, 2017). Knowing the cognitive state can help us to understand and prevent many situations such as traffic accidents by detecting drowsiness, help teacher/trainer to understand student's interest level in the classroom, online training or meetings.

Further, the emotions are also recognised from one's speech, body language, facial expressions, biosignals, breathing patterns, etc. By combining such multiple emotion recognition modalities, a robust single recognition model is created to eliminate the ambiguity and uncertainty during decision making. The integration of the corresponding, redundant or contradictory data extracted from the multimodal emotional recognition model, gives a unique description of the natural emotions and also processes this information to predict the future behaviour of the person (Al-Tayyan *et al*., 2017). Hence, an efficient computing approach enhances the level of interaction, detection and modelling of the emotions to maximize the learning and entertainment (Liu *et al*., 2014).

Now researchers globally and social psychologists and body languages experts identified the face and the facial expressions to be the most powerful and highly influential and impactful in determining person reactions. Combining these facial expressions with hand over face gestures helps in determining the cognitive state of a person, especially in sitting positions where there is less body movement people tend to express their cognitive state through hands touching the face. The proposed approach can be highly useful in e-Learning environments or smart classrooms to detect whether the person is interested, thinking, unsure, bored and happiness.

Hand-over-face descriptor, hand action on the face and the region occluded by hand on the face are a significant channel of nonverbal communication and gives essential information about the cognitive state of a person. Figure 1 shows the heat map of cognitive states described by Mahmoud and Robinson (2011).

In this study, we will explore the combination of hand-over-face gestures with facial expressions to build a multimodal system for cognitive state recognition. We believe this problem can be divided into three independent tasks: coding and classification of hand-over-face descriptors, detection of basic facial expressions using facial landmarks and combining facial expressions and hand-over-face gesture for recognition of the cognitive state.

Literature review: Emotion plays a significant role in human-computer interaction, social interaction,

Corresponding Author: H.S. Gunavathi, Department of CS and E, Jain University, Bengaluru, India

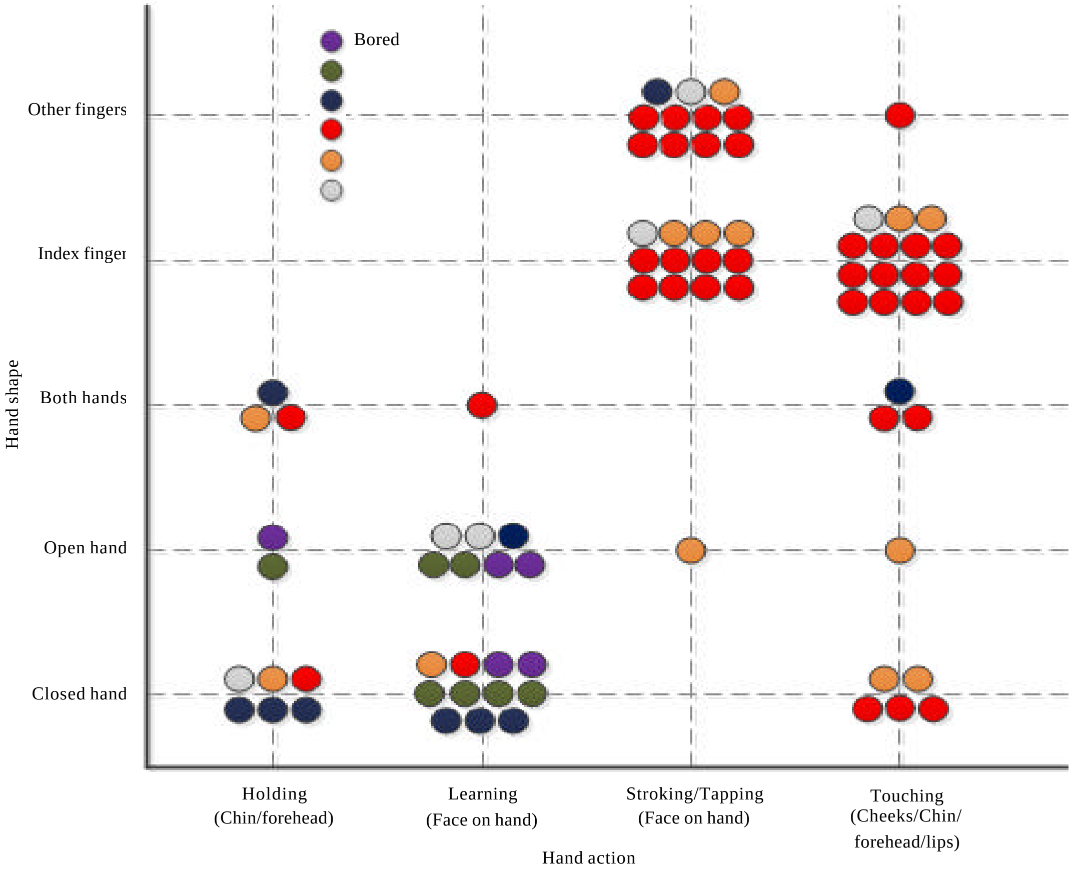

Fig. 1: Encoding of hand-over-face descriptor and action in different cognitive states (Mahmoud and Robinson, 2011)

perception, educational scenarios, e-Learning environments, etc. Emotion is expressed by a person in many ways such as speech, text, facial expression, biosignals, etc. Automatic detection of human emotion or cognitive state of a person can help in providing the human-computer interface with humanity and achieve a pleasant interaction. Potential applications based on multimodal cognitive state recognition are Affective robot, smart classroom, e-Learning, smart home, etc.

In remaining part of this study, the recent techniques which are employed in recognition of emotions based on various multimodal approaches such as facial expressions and hand-over-face gestures are discussed in detail.

Emotion recognition using facial landmark detection and tracking: Castellano *et al.* (2008) developed a novel method by considering the facial expression comprehensiveness to visualize the Facial Action Coding System (FACS). This technique is used to understand the different capability of facial muscle movements in terms of predefined Action Units (AUs). In this research, action units are found to be numerically coded and facial expressions were noticed to correspond with one or more action units. These approaches were used for the detection of emotions by employing FACS and are used to describe facial muscle activation in spite of the underlying cause. Consequently, several other techniques were developed by considering the FACS such as the Emotional Facial Action Coding System (EMFACS) (Friesen and Ekman, 1983).

Hickson *et al.* (2017) proposed a novel technique for facial landmark and the emotion recognition based on the data collected from a particular region of the face such as the eye region. The method captures the facial marker and the skeleton data of a human while interacting with the automated machines. This technique provided an accurate analysis of the emotion recognition that predicted the human behaviour and the emotional state of a being.

Tsai and Chang (2013) presented a novel technique for Facial Expression Recognition (FER) by executing in the support vector machine. The model consists of Self-quotient Image (SQI) filter, to overcome the insufficient light while detecting the face regions.

Also, the model designs the features using the schemes, Discrete Cosine Transforms (DCT), Angular Radial Transform (ART) and Gabor Filter (GF).

The classification techniques used in facial expression recognition systems are categorized based on the static images (Shan *et al*., 2009) and with dynamic image sequences (Byeon and Kwak, 2014). The temporal information such as feature vector comprises information about the current input image was not incorporated into static images. However, the uses of temporal information of images were included in sequence-based methods to identify the expression captured from one or more frames. Liu *et al*. (2014) developed a technique for facial expression recognition based on Boosted Deep Belief Network (BDBN) which was quite promising when tested on CK+ and JAFFE but didn't obtain proper results when the cross-database configuration was considered.

Hand-over face gestures: Hussain *et al*. (2012) developed a non-intrusive student mental state prediction system in which they coded six different gestures of hand-over-face using SLBP, force field features and classified the mental states using the three-layered Bayesian networks.

Dominio *et al*. (2014) proposed a real-time multi-class SVM classifier based on multimodal hand posture recognition technique. The classifier trains the data by two methods namely user training and general training. The dataset was extracted from the hand gestures like fingertips, hand position and curvature, shape and position of palm, etc. The model showed more accurate validation in the hand-over face gesture recognition.

Mahmoud *et al*. (2014) proposed a technique for analyzing natural hand-over-face gestures by using spatial and temporal features. They were reasonably successful in extracting hand from the face in varying lighting conditions and results were quite promising.

Prabhu and Jayagopi (2017) developed an emotional intelligent system to determine the approaches used by humans to interact with machines. The method used in this research is not only to interpret human affective states but also to respond in real time during assistive human to device interactions. A Multimodal Emotion Recognition System (MERS) is postulated to integrate face cues and hand over face gestures to operate in near real time with an average frame rate of 14 fps. Presently, there are several techniques used as an emotion recognition systems using facial landmarks. This proposed research is considered for the inclusion of hand over face gestures that are commonly expressed during emotional interactions.

Although, there are considerable numbers of methods which can measure emotion recognition through the face, there are few of them which include hand over face gesture and none of them combines facial expressions with hand over face gesture to detect the cognitive state in real time. So, this manuscript focuses primarily on combining the facial expressions with hand over face gestures for building cognitive state recognition system.

MATERIALS AND METHODS

We break up the problem into three parts. First finding the gestures through hand-over-face descriptors, second detection of basic facial expressions using facial landmarks and finally we combine facial expressions and hand-over-face gesture for recognition of the cognitive state. Figure 2 shows the architecture of our proposed system for multimodal cognitive state recognition.

Coding of hand-over-face gesture descriptor: For categorizing the hand-over-face gesture, we have selected publicly available Cam3D corpus database (Mahmoud *et al*., 2011) from Cambridge University. The Cam3D corpus has 192 audio/videos segments that comprise of natural complex mental states and it has been publicly labelled using crowd sourcing and licensed under a creative commons attribution non-commercial share alike 3.0 license. Out of 192, we have selected around 73 videos containing hand-over-face gestures. We coded the handover face descriptor in terms of hand action and shape and facial region occluded. We have used similar coding as described by Mahmoud *et al*. (2014) for hand action detection and proposed a novel approach for detecting the hand gestures and hand occlusions. Labelling of videos was done manually using following guidelines.

Hand action: For the full video hand action is coded as one label (static or dynamic). Labels are the stroking/tapping (if the hand is dynamic) or touching (If the hand is static).

Hand occlusion: Describes hand occlusion present or not present and coded as one per frame.

Hand gestures/Hand occlusion descriptor: Coded as one label per frame based on novel binary vectors created as described in next session.

Hand gesture detection and localization: Handover face gesture recognition is a complex problem in image

Table 1: Hand occlusion descriptor coded as binary vectors

Hand occlusion descriptor	Binary codes				
HG1 (Hands on cheek/Chin rest)	11010000	01100100	11110100	01100000	11000000
HG2 (Eye rubbing)	00000001	00000010	00000011		
HG3 (Lip touch)	01001000	01000000			

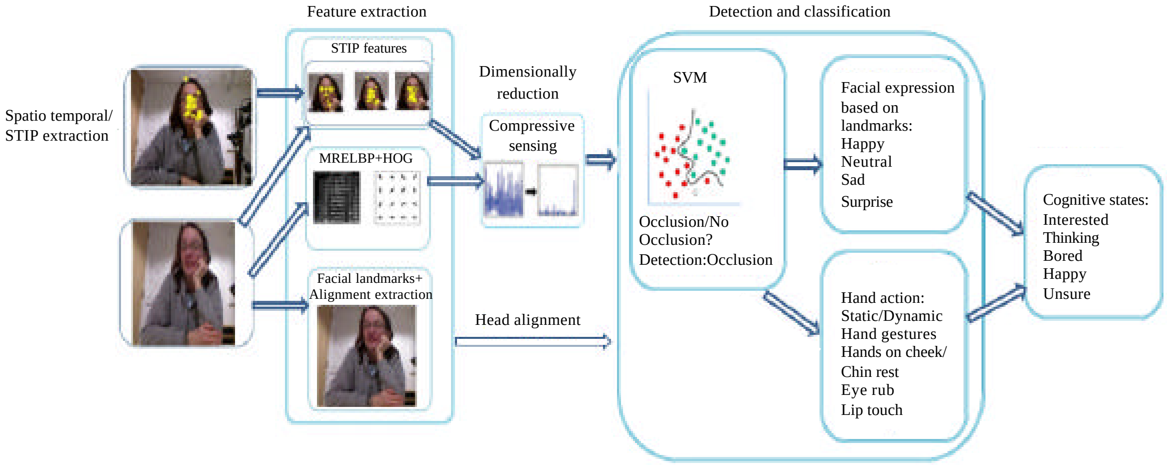

Fig. 2: Architecture of multimodal cognitive state recognition system

processing and pattern recognition. It is challenging to segment hands from faces because of similar colour and texture of skin. So, we assume near frontal faces and fewer variations in head movements. First, we extract the MRELBP (Liu *et al*., 2016) of the initial frame and compute feature histogram of the frame and next, we compute the MRELBP based histogram for each frame and subtract it from original frame to find the delta H(δH). When the hand occluded the face the difference in the value of δH is enhanced. In Eq. 1:

$$H(i, j) = \sum_{I_c \in R_j} f\{MRELBP(I_c)\}, i = 0, ..., n\text{-}1, j = 0, ..., m\text{-}1 \quad (1)$$

Where:

I_c = The central pixel of MRELBP coded image
H(i, j) = The ith value of MRELBP based histogram of jth region in the image

We have divided the image into eight regions as shown in Fig. 3 and assign value ‘1’ to a region if the region is occluded by hand otherwise assign a value ‘0’. Based on assigned values we create a binary vector of 8 bits and a binary code is generated for each frame and labelled as one gesture as described in Table 1. The initial value of H is computed using near frontal and neutral face and subsequent H’ is computed for next frame. Using difference between H and H’ the delta H(δH) is computed for each region and compared against the threshold value. If the difference δH is more than the threshold value, we consider it has hand occlusion. Location of the hand in different frames is coded as a binary vector as shown in Table 1 and these codes are used for training the network.

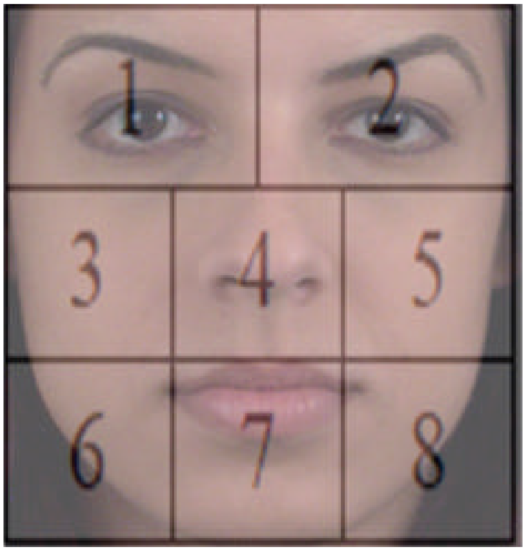

Fig. 3: Picture showing how we partition face into eight regions

Feature extraction for hand over face occlusion descriptor: Feature extraction is a significant step in cognitive state recognition. In our proposed method, we consider Space-Time Interest Points (STIP) that combined spatial and temporal features for hand action recognition. For spatial features, we have used Median Robust Extended Local Binary Pattern (MRELBP) (Liu *et al*., 2016) and Histogram of Oriented Gradient (HOG). After feature extraction, we have used compressive sensing to reduce the feature dimension.

Space-time interest point (STIP): In recent times local space-time features are widely employed for action recognition (Poppe, 2010). Song *et al.* (2013) used local space-time feature to encode the face and body micro expressions for emotion detection. STIP seizures significant visual patterns in space-time and reflect interesting events by augmenting local spatial image descriptor to space-time domain. To extract the space-time interest point's features, we have used a technique developed by Song *et al.* (2013). Acquiring local space-time features has two stages. First detection of spatiotemporal interest points and second extraction of STIP features. Wang *et al.* (2009) proposed a technique in which he used Harris3D interest point detector followed by a combination of HOG and HOF descriptor. This combination provides excellent performance. Keeping good performance in mind, in our proposal, we used Harris 3D interest point detector followed by MRELBP based HOG feature descriptor for extracting local space-time features. Further to improve the performance we restrict STIP features extraction for face area and discard any STIP points outside face area.

To capture micro expressions high dense local space-time features are extracted. Because of dense feature vectors, the computation cost is very high and it limits us using it in the real-time applications. So, we propose to minimize feature space dimensionality using compressive sensing which will lead to the decrease in computation cost and increase in performance.

MRELBP based HOG: Liu *et al.* (2016) proposed Median based Extended Local Binary Pattern (MRELBP) to overcome the limitation of LBP such as sensitivity to image noise and not able to capture macrostructure texture information. Median based extended LBP compares regional image medians rather than raw image intensities. MRELBP is highly robust to image noise including salt-and-pepper noise, Gaussian blur, Gaussian noise and random pixel corruption.

HOG descriptors are highly useful for detection of textured objects including distorted shapes. They are used efficiently for facial landmark detection (Zhu and Ramanan, 2012) and pedestrian detection. In our proposed approach, we are going to divide an image into a number of small connected regions and compute MRELBP. We generate the histogram of gradient directions for all pixels in the region and the whole image.

Compressive sensing: Compressive sensing is a way of an efficiently acquiring a signal with prior knowledge that the signals of interest are sparse and rebuilding the signal by finding solutions to underdetermined linear systems.

Compressive sensing relies on two main principles. Natural signals/images have a sparse representation in specific basis or transformation domain such as discrete cosine transform, Haar, wavelet domain, etc. The measurement matrix Φ is incoherent (poorly correlated) with the signal basis matrix Ψ.

We propose to use STIP, MRELBP with HOG and compressive sensing technique. These techniques have many advantages like illumination invariant, operating on local cells, allowing for contrast normalization and being invariant to photometric transformations and geometric orientation. Because of these natural benefits, we use MRELBP and HOG for feature extraction. After identifying the features, we reduce the dimensionality of feature space using compressive sensing which will lead to a reduction in computation cost. For classifying these features, we use multiclass SVM classifiers.

Facial expression recognition using facial landmark (Action units): In recent years, facial expression recognition is studied widely as part of which we have many publicly available databases. We have considered CK+database for our training purpose as it has emotion, image data, action unit's labels and tracked landmarks. We are concentrating on four facial expressions viz. neutral, sad, happy and surprise which is relevant to cognitive state recognition. For this purpose, the CK+database was considered and physically separated out based on facial expression labels.

In the research, we have chosen to use dlib and openCV, since, it had 1 msec pose estimation with an ensemble of regression trees (Kazemi and Sullivan, 2014). The faces are detected using Viola and Jones (2004) method.

Table 2: The production rules of combining the facial expressions, hand occluded region descriptor and hand action

R_i	P	Q
R_1	(Happy, hands on cheek/chin rest, static)	Interested
R_2	(Happy, hands on cheek/chin rest, dynamic)	Thinking
R_3	(Neutral, hands on cheek/chin rest, static)	Interested
R_4	(Neutral, hands on cheek/chin rest, dynamic)	Thinking
R_5	(Sad, hands on cheek/chin rest, static)	Bored
R_6	(Sad, hands on cheek/chin rest, dynamic)	Thinking
R_7	(Surprise, hands on cheek/chin rest, static)	Other
R_8	(Surprise, hands on cheek/chin rest, dynamic)	Other
R_9	(Happy, eye rub, static)	Happy
R_{10}	(Happy, eye rub, dynamic)	Happy
R_{11}	(Neutral, eye rub, static)	Unsure
R_{12}	(Neutral, eye rub, dynamic)	Bored
R_{13}	(Sad, eye rub, static)	Bored
R_{14}	(Sad, eye rub, dynamic)	Bored
R_{15}	(Surprise, eye rub, static)	Other
R_{16}	(Surprise, eye rub, dynamic)	Other
R_{17}	(Happy, lip touch, static)	Interested
R_{18}	(Happy, lip touch, dynamic)	Thinking
R_{19}	(Neutral, lip touch, static)	Unsure
R_{20}	(Neutral, lip touch, dynamic)	Thinking
R_{21}	(Sad, lip touch, static)	Unsure
R_{22}	(Sad, lip touch, dynamic)	Thinking
R_{23}	(Surprise, lip touch, static)	Other
R_{24}	(Surprise, lip touch, dynamic)	Other

Combining facial expression and hand-over-face gesture: For combining facial expression, hand action and hand occluded region descriptor, we adopt a decision making strategy based on production rules. Production rules are commonly used in artificial intelligence and cognitive modelling as a simple expert system. Through the production rules, the four facial expressions (Happy, sad, neutral and surprise), three hand occluded region descriptors (Hands on cheek/chin rest, eye rub and lip touch) and two hand actions (Static or dynamic) are combined to cognitive state recognition. The production rules contain IF part (a condition or premise) and then part (an action or conclusion). The form of production rules is defined as Eq. 2:

$$R_i = \text{IF P THEN Q} \quad (2)$$

Where:

- R_i = The rule i, P is the antecedent of rule i and is formed by (r_1-r_3)
- r_1 = The facial expression
- r_2 = Hand occluded region descriptor (Hands on cheek/chin rest, eye rub, lip touch)
- r_3 = Hand action (Static/dynamic)
- Q = The latter of rule i that is a cognitive state (Interested, thinking, bored, unsure, happy)

Table 2 describes the production rules for cognitive state recognition by combining facial expressions, hand occluded region descriptor and hand action.

RESULTS AND DISCUSSION

To evaluate our proposed approach, we have considered a labelled subset of the Cam3D dataset. Table 3 describes the Cam3D dataset which we considered for training our network. We have considered 73 videos which have 13268 frames.

Table 3: Data set consideration for training (Cam3D)

Hand action	Static/Dynamic	Whole video	73 videos (13268 frames)
Hand occlusion	Occluded/Not occluded Binary vector for hand gesture (If hand occluded)	Whole video one frame	73 videos (13268 frames)

Table 4: Hand over face detection

Methods	STIP (%)	MERLBP+HOG (%)	Fusion (%)
Hand occlusions (1367 frames)	78.1	87.1	91.46
Hand action (748 frames)	72.3	84.1	93.72

In a pre-processing step, we remove noise and carry out face alignment on videos considered from Cam3D database and face detection was done using Viola and Jones method followed by fine-tuning and tracked using dlib landmark detector. The image was scaled to the resolution of 160×120 pixels. Some of the videos were removed as we were not able to detect the face because of extreme head rotation or full hand occlusion. STIP features are extracted as discussed in this study. The features which are not associated to facial regions are discarded by using facial landmarks. For our experiment, we extracted the MRELBP based HOG features from normalized 160×120 pixel image of a face. The 8×8 pixel cells are used with eighteen gradient orientations and block size of 2×2 cells. The length of MRELBP based HOG feature vector in the research is 9576 and we reduced the dimensionality using compressive sensing to 937 dimension vector per frame. For classification, we experimented on both unimodal and multimodal fusion of feature vectors using the standard library, i.e., Liblinear for SVM.

Hand occlusion and hand action detection: Our first step in testing is to detect the hand occlusion. The face was divided into eight regions and it's considered to be occluded if one or more region of the face is occluded. For hand occlusion, the data was labelled one per video describing whether it's occluded or not occluded and for hand action, it was labelled one per frame describing the hand action as a static or dynamic. We employed binary classifier to classify whether hand action is static or dynamic. Table 4 shows the hand occlusion and hand action accuracy results comparing unimodal and multimodal feature fusion.

Hand occluded region descriptor recognition: Our second step is to detect the hand shapes or region of face occluded using the binary codes. The binary vector is generated based on the regions of face occluded and based on binary vector we will detect the hand gestures. The occlusion of every region is considered as a different binary classification problem and a threshold of 25% is

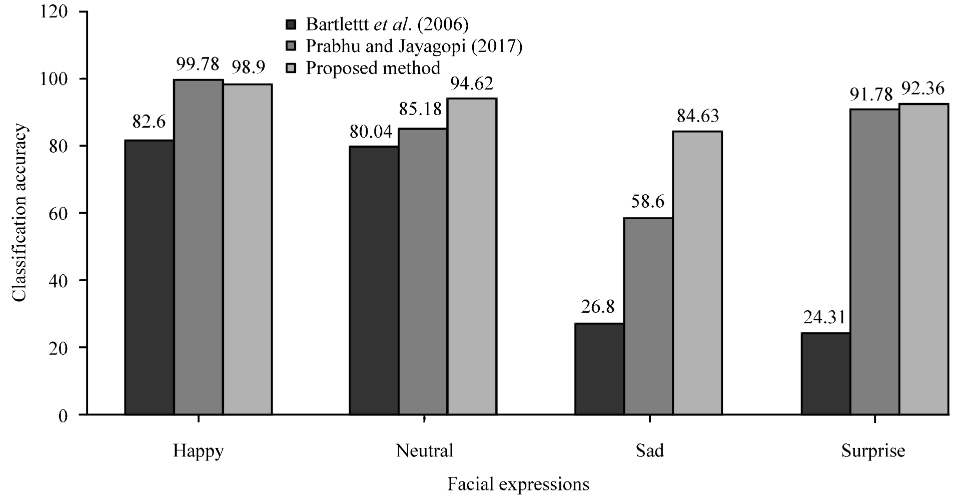

Fig. 4: Facial expression recognition accuracy comparison between Bartlett *et al*. (2006), Prabhu and Jayagopi (2017) and proposed method

Table 5: Classification results of hand occlusion and action detection

Hand gestures	Unimodel classification		Fusion (%)
	MRELBP+HOG (%)	STIP (%)	
Hands on cheek/Chin rest	84.21	62.12	95.72
Eye rub	89.84	67.32	91.23
Lip touch	86.12	65.82	96.47

Table 6: Confusion matrix of cognitive state recognition

Facial expression	Interested (%)	Thinking (%)	Bored (%)	Happy (%)	Unsure (%)
Interested	92.41	1.3	0.0	5.88	0.41
Thinking	3.8	94.34	0.7	1.16	0.0
Bored	0.86	0.12	82.62	0.0	16.4
Happy	1.1	0.0	0.0	98.9	0.0
Unsure	0.8	1.48	13.46	0.0	84.28

considered for to recognize whether the region is occluded or not. The classification results obtained by unimodal and multimodal feature fusion are presented in Table 5.

From results, we found that multimodal feature fusion had higher accuracy and performed better compared to unimodal in the detection of every occluded facial region.

Facial expression through facial landmarks: We have considered four facial expressions such as happy, neutral, sad and surprise which are essential for recognition of cognitive states. The Cohn-Kanade dataset was manually separated based on facial expressions. Firstly, we detect all the faces in the image using Viola and Jones method and crop the facial region and save the cropped images to disk. About 68 facial landmarks are detected using dlib and then landmark points are aligned on detected faces. Based on these landmarks MRELBP based HOG features are generated. For single face, the length of feature vector is 4556 as we used length and slope of the line segment from every point to every other point. We have used multiclass SVM classifier for classification task as it was computationally less intensive compared to the neural network. Both One-vs.-One (OVO) and One-vs.-All (OVA) techniques are used for training the network. One-vs.-one is less affected by the problem of imbalance dataset but it's computationally expensive and one-vs.-all is more straightforward to get the probability values corresponding to each facial expressions. We have used the one-vs-all technique for classification as it was computationally less expensive. We have compared our proposed approach against approach proposed by Bartlett *et al*. (2006), Prabhu and Jayagopi (2017) in terms of facial expression with landmark solutions. Figure 4 gives accuracy details of the proposed system for facial expression recognition. Based on the experimental results we found that our proposed framework is better in terms of accuracy precisely for cases like neutral, sad and surprise.

Cognitive state recognition: Our final aim was to detect the cognitive state by combining the facial expressions with hand over face gesture. Our final system is able to detect five cognitive states by combining four facial expressions with three hand-over-face gestures. To improve the performance of the system we skipped certain frames on the assumption that cognitive state will not change in six frames, i.e., 1/5th of a second.

Based on production rules defined in Table 2 by combining the facial expressions, hand occluded region descriptor and hand action, we have developed an expert system to classify the cognitive state. Table 6 shows the classification accuracies of cognitive states detected on our test data set. We obtained an average cognitive state recognition rate of 90.51% with 15.8 fps.

CONCLUSION

In this research work, we presented an automatic approach for recognizing the cognitive state by combining facial expression, hand-over-face descriptor and hand action. We divided the problem into three parts: classifying hand-over-face gesture cues, facial expression recognition using facial landmarks, combining facial expression and hand-over-face gesture cues to recognize the cognitive state. We showed that multimodal feature fusion performs better compared to uni-modality classification results. We have validated our proposed system on the Cam3D dataset and our results show improved performance and classification of cognitive states. Going forward, we can add more hand-over-face descriptors to make the accuracy of the system better and add more cognitive states. Also, we can add voice modality and other upper body features to improve the prediction results for cognitive state recognition.

REFERENCES

Al-Tayyan, A., K. Assaleh and T. Shanableh, 2017. Decision-level fusion for Single-view gait recognition with various carrying and clothing conditions. Image Vision Comput., 61: 54-69.

Bartlett, M.S., G. Littlewort, M. Frank, C. Lainscsek and I. Fasel *et al*., 2006. Fully automatic facial action recognition in spontaneous behavior. Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition FGR, April 10-12, 2006, IEEE, Southampton, UK., pp: 223-230.

Byeon, Y.H. and K.C. Kwak, 2014. Facial expression recognition using 3D Convolutional neural network. Intl. J. Adv. Comput. Sci. Appl., 5: 107-112.

Castellano, G., L. Kessous and G. Caridakis, 2008. Emotion Recognition Through Multiple Modalities: Face, Body Gesture, Speech. In: Affect and Emotion in Human-Computer Interaction, Peter, C. and R. Beale (Eds.). Springer, Berlin, Germany, ISBN:978-3-540-85098-4, pp: 92-103.

Dominio, F., M. Donadeo and P. Zanuttigh, 2014. Combining multiple Depth-based descriptors for hand gesture recognition. Pattern Recognit. Lett., 50: 101-111.

Fernandes, S. and J. Bala, 2017. A comparative study on various state of the art face recognition techniques under varying facial expressions. Intl. Arab J. Inf. Technol., 14: 254-259.

Friesen, W.V. and P. Ekman, 1983. EMFACS-7: Emotional Facial Action Coding System. University of California at San Francisco, San Francisco, California, USA.,.

Hickson, S., N. Dufour, A. Sud, V. Kwatra and I. Essa, 2017. Eyemotion: Classifying facial expressions in VR using Eye-tracking cameras. J. Inf. Software Eng., 1: 1-10.

Hussain, A., A.R. Abbasi and N. Afzulpurkar, 2012. Detecting and interpreting Self-manipulating hand movements for student's affect prediction. Hum. Centric Comput. Inf. Sci., 2: 2-18.

Kazemi, V. and J. Sullivan, 2014. One millisecond face alignment with an ensemble of regression trees. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, IEEE, Columbus, Ohio, USA., ISBN:978-1-4799-5118-5, pp: 1867-1874.

Liu, L., S. Lao, P.W. Fieguth, Y. Guo and X. Wang *et al*., 2016. Median robust extended local binary pattern for texture classification. IEEE. Trans. Image Process., 25: 1368-1381.

Liu, P., S. Han, Z. Meng and Y. Tong, 2014. Facial expression recognition via. a boosted deep belief network. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, IEEE, Columbus, Ohio, USA., ISBN:978-1-4799-5118-5, pp: 1805-1812.

Mahmoud, M. and P. Robinson, 2011. Interpreting Hand-over-face gestures. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, October 9-12, 2011, Springer, Berlin, Germany, ISBN:978-3-642-24570-1, pp: 248-255.

Mahmoud, M., T. Baltrusaitis, P. Robinson and L.D. Riek, 2011. 3D corpus of spontaneous complex mental states. Proceedings of the International Conference on Affective Computing and Intelligent Interaction, October 9-121, 2011, Springer, Berlin, Germany, ISBN:978-3-642-24599-2, pp: 205-214.

Mahmoud, M.M., T. Baltrusaitis and P. Robinson, 2014. Automatic detection of naturalistic hand-over-face gesture descriptors. Proceedings of the 16th International Conference on Multimodal Interaction, November 12-16, 2014, ACM, New York, USA., ISBN:978-1-4503-2885-2, pp: 319-326.

Poppe, R., 2010. A survey on vision-based human action recognition. Image Vision Comput., 28: 976-990.

Prabhu, M.K. and D.B. Jayagopi, 2017. Real time multimodal emotion recognition system using facial landmarks and hand over face gestures. Intl. J. Mach. Learn. Comput., 7: 30-34.

Shan, C., S. Gong and P.W. McOwan, 2009. Facial expression recognition based on local binary patterns: A comprehensive study. Image Vision Comput., 27: 803-816.

Song, Y., L.P. Morency and R. Davis, 2013. Learning a sparse codebook of facial and body Microexpressions for emotion recognition. Proceedings of the 15th ACM on International Conference on Multimodal Interaction, December 9-13, 2013, ACM, New York, USA., ISBN: 978-1-4503-2129-7, pp: 237-244.

Tsai, H.H. and Y.C. Chang, 2018. Facial expression recognition using a combination of multiple facial features and support vector machine. Soft Comput., 22: 4389-4405.

Viola, P. and M.J. Jones, 2004. Robust real-time face detection. Int. J. Comput. Vision, 57: 137-154.

Wang, H., M.M. Ullah, A. Klaser, I. Laptev and C. Schmid, 2009. Evaluation of local Spatio-temporal features for action recognition. Proceedings of the International Conference on British Machine Vision BMVC, September 7-10, 2009, BMVA Press, London, UK., pp: 124-1-124-11.

Zhu, X. and D. Ramanan, 2012. Face detection, pose estimation and landmark localization in the wild. Proceedings of the 2012 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 16-21, 2012, IEEE, Providence, Rhode Island, USA., ISBN: 978-1-4673-1226-4, pp: 2879-2886.