Cognitive behaviour analysis based on facial information using depth sensors

Montenegro, J.F., Villarini, B., Gkelias, A. and Argyriou, V.

An accepted author manuscript version of a paper published in Wannous H., Pala P., Daoudi M., Flórez-Revuelta F. (eds) Understanding Human Activities Through 3D Sensors. UHA3DS 2016. Lecture Notes in Computer Science, vol 10188. Springer, Cham.

It is available from the publisher at:

https://doi.org/10.1007/978-3-319-91863-1_2

Cognitive behaviour analysis based on facial information using depth sensors

Juan Manuel Fernandez Montenegro, Barbara Villarini, Athanasios Gkelias, and Vasileios Argyriou

Kingston University London, University of Westminster London, Imperial College London
Juan.Fernandez@kingston.ac.uk,B.Villarini@westminster.ac.uk,A.Gkelias@imperial.ac.uk,Vasileios.Argyriou@kingston.ac.uk

Abstract. Cognitive behaviour analysis is considered of high importance with many innovative applications in a range of sectors including healthcare, education, robotics and entertainment. In healthcare, cognitive and emotional behaviour analysis helps to improve the quality of life of patients and their families. Amongst all the different approaches for cognitive behaviour analysis, significant work has been focused on emotion analysis through facial expressions using depth and EEG data. Our work introduces an emotion recognition approach using facial expressions based on depth data and landmarks. A novel dataset was created that triggers emotions from long or short term memories. This work uses novel features based on a non-linear dimensionality reduction, t-SNE, applied on facial landmarks and depth data. Its performance was evaluated in a comparative study, proving that our approach outperforms other state-of-the-art features.

Keywords: cognitive behaviour, depth sensors, dimensionality reduction

1 Introduction

Cognitive behaviour analysis is considered of high importance with many innovative applications in a range of sectors including healthcare, education, robotics and entertainment. Particularly in healthcare, cognitive and emotional behaviour analysis helps to improve the quality of life of patients and their families. For instance, behavioural symptoms in patients with mild cognitive impairment can indicate an increased risk of progressing to dementia, while mild behavioural impairment in patients who usually have normal cognition indicates an increased risk of developing dementia. Based on these symptoms many research projects are focused on the detection of these cognitive handicaps that characterise different forms of dementia and possible Alzheimer disease. Moreover, it has been well established that changes in cognitive behaviour in many cases are closely related to anxiety and depression and that schizophrenic patients have neurocognitive deficits.

Amongst all the different approaches for cognitive behaviour analysis, significant attention has been drawn upon emotion analysis through facial expressions using RGB and depth data. Along the same lines, our work introduces an emotion recognition approach using facial expressions based on depth data and landmarks. A novel dataset was created that triggers emotions from long or short term memories. This work uses novel features based on a non-linear dimensionality reduction technique, i.e., t-SNE, which is applied on facial landmarks and depth data.

In literature, several approaches for automatic emotion recognition are focused on the variety of human interaction capabilities or biological data. For instance, the study of speech and other acoustic cues in [36], body movements in [5], electroencephalogram (EEG) in [17], facial expressions or combinations of previous ones, such as speech and facial expressions in [21] or EEG and facial expressions in [30]. The study of facial expression has been part of various disciplines since Aristotelian era but it was only in 1978 when the first automatic recognition study appeared [27, 2]. Since then, several techniques have been proposed to interpret facial expressions. One of the most well known approaches is the Facial Action Coding System (FACS) [8]. FACS describes facial expressions as action units (AU), where each AU corresponds to a facial configuration. Based on the computational methodology existing approaches can be classified as "spatial" or "spatio-temporal" and "appearance" or "shape based". The first approach differentiates between methodologies that work with single images or groups of successive frames. The second approach groups methods that use the appearance features of the face, such as pixel intensity, and methods that use a description of the face shape.

On the other hand, several datasets are available for emotion recognition focusing on different applications. For example, DEAP dataset provides EEG and face recordings of participants while they watch musical videos just for the analysis of human affective states [12]. SEMAINE database aims to provide voice and facial information to study the behaviour of subjects interacting with virtual avatars [18]. MAHNOB-HCI database was created for the study of emotions while humans are watching multimedia, supplying several data, such as audio, an RGB video and five monochrome videos of the face, EEG, ECG, respiration amplitude, skin temperature and eye-gaze data [31]. Finally, CASMEII dataset studies facial micro-expressions for security and medical applications, requiring cameras of higher frame rate and spatial resolution [38]. All of them face the same challenges, such as head-pose and illumination variations, occlusions, registration errors and identity bias. Most of these impairments/conditions are excluded in most of the available databases which makes these data and the corresponding approaches inappropriate in real environment conditions.

The purpose of this work is to introduce human cognitive behaviour analysis based on face expression recognition. Our novel dataset contains recordings of the participants' reactions when specific images related and unrelated with their personal life stories are shown. The classification of different reactions related to the images displayed is performed using different data features included in our

dataset, such as depth and EEG data, as input to supervised learning approaches. Thus our classification is based on the expected emotional behaviours according to the images displayed during the test. Furthermore, advanced techniques for dimensionality reduction are utilised providing a better representation of the input data, which corresponds to our feature vectors.

The remainder of this paper is organized as follows: In section 2 previous related works on behaviour and emotion recognition are summarized. Section 3 introduces and analyses the proposed novel methodology while in section 4 the details of the evaluation process and the obtained results are presented. Finally, conclusions are drawn in Section 5.

2 Previous Work

This section summarizes the state of the art approaches to facial and EEG based emotion recognition.

2.1 Facial Emotion Recognition Approaches

The most common approach to emotion recognition has been the utilization and processing of images and video sequences of faces. The majority of existing works are based on the Facial Action Coding System (FACS) [8] that enables the description of basic human emotions, such as happiness, sadness, surprise, fear, anger or disgust, as a combination of basic facial configurations/expressions known as action units (AU). Another popular approach is the the use of the three dimensional Schlosberg Model [10] (i.e.,"pleasantness-unpleasantness", "attention-rejection" and "sleep-tension") which attempts to conceptualize human emotions by defining where they lie in two or three dimensions. Other approaches abandon the path of specific emotions recognition and focus on emotions' dimensions, measuring their valence, arousal or intensity [37, 21, 22, 32, 34].

The facial emotion recognition methods can be classified into two main groups according to the approaches used during the recognition stages: (i) face registration and (ii) features selection [27, 2].

Face registration: Three different approaches have been used for face registration: whole face, facial parts and facial points registration. These registration approaches are usually based on Active Appearance Models (AAM) [6, 1], i.e., a method that matches facial images to a statistical facial model in order to extract specific face areas and facial landmarks. Whole face approaches extract the features from the whole face. For instance, Littlewort et al [16] extracted image based features of the whole face, such as Gavor Wavelets, in order to detect AUs for pain recognition. On the other hand, facial parts approaches use specific face areas, such as the eyebrows and the mouth, which contain the maximum amount of information related to face expressions. Nicolle et al [22] proposed a multimodal method for emotion recognition (valence, arousal, expectancy and power) using a combination of the whole face, facial parts, facial points and audio features. This approach gets patches of the face on regions of interest and uses

the log-magnitude Fourier spectra and other measures as features. Finally, facial points based approaches use fiducial points for shape representation. Michel et al [19] used a tracker to get 22 fiducial points and calculate the distance of each point between a neutral and a peak frame. These distances are used as features in a Support Vector Machine (SVM) algorithm that classifies the emotions. Neutral and peak frames are automatically detected when the motion of the points is almost zero. Valstar et al used Particle Filtering Likelihoods [24] in order to extract 20 fiducial points, however, the initial positions of these points were selected manually. These points were normalised with respect to a neutral point (i.e., tip of the nose) and a scale transformation was also applied. The distances between certain points were used as features to recognise specific AUs using SVM.

Feature representation: Feature representation methods can be divided into spatial and spatio-temporal. Spatial approaches include shape representations, low-level histograms or Gabor representations amongst others. For instance, Huang et al [9] proposed a spatial shape representation using groups of three fiducial points (triangular features) as input to a neural network classifier. Sariyanidi et al presented in [28] a low-level histogram representation using local Zernike moments for emotion recognition based on kNN and SVM classifiers. On the other hand, spatio-temporal approaches extract the features from a range of frames within a temporal window, detecting more efficiently emotions that cannot be easily differentiated in spatial approaches. Zhao et al [39] proposed a method that uses spatio-temporal local binary patterns as features and SVM for classifying facial expressions. Once the features are selected, dimensionality reduction techniques, such as PCA are usually used before classification in order to reduce illumination variation, registration errors and identity bias.

Unfortunately, the results in many of the aforementioned approaches are not always reliable since many of them are tested on posed datasets such as CK [11] and MMI [23]. Due to the limitations of these datasets, the obtained results are not directly applicable in real conditions that include illumination, head-pose variations and nature of expressions. Nevertheless, there are non-posed datasets to test naturalistic expressions such as SEMAINE [18] or MAHNOB-HCI [31]. In these cases the illumination and head-pose variation problems have been taken into account depending on the aim of the study.

Similar approaches are used for behaviour analysis based on the whole human body. The majority of them [26, 3, 4, 7, 15, 14, 13] for activity recognition focus on classifying the action as soon as possible using pre-segmented sequences and combine depth and skeleton data. These approaches achieve reasonable accuracy but manual pre-segmentation simplifies the task of detection which inflates accuracy and limits real-world applications.

2.2 EEG Emotion Recognition Approaches

EEG based techniques have been less popular compared to the use of facial or speech data as a source for emotion detection. Considering that these sources are easy to fake [17] amongst other problems, EEG provides an extra source that

solves problems such as falseness, illumination or speech impaired subjects. On the other hand, EEG signal faces another set of challenges, such as noise and biological and non-biological artefacts [30, 20], such as electrocardiogram (ECG), electromyogram (EMG) and electrooculogram (EOG). Nevertheless, these biological artefacts are also affected by emotions and have been shown to provide extra information to EEG signals for emotion recognition [30].

Two types of descriptors can be used for EEG signal analysis: (i) simple descriptors, such as frequency and amplitude, and (ii) more complex ones such as asymmetry metrics, time/frequency analysis, topographic mapping, coherence analysis or covariation measures. The use of these descriptors is usually dictated by the application/area of study. For example, asymmetry metrics are usually applied in cognitive neuroscience [20]. In particular, asymmetric hemispheric differences were used for emotion recognition [29, 25]. Furthermore, state of the art techniques, such as Independent Component Analysis (ICA), have been used to remove some artefacts and extract different features (such as amplitude or spectral power), and use them in classifiers such as k-Nearest Neighbour (kNN) or SVM. For instance, Vijayan et al [35] used DEAP data (data captured using 32 sensors) through a 50Hz filter to remove noise, then the Gamma band was extracted from the signal and auto-regressive modeling was used to obtain the features which were passed to an SVM classifier.

The combination of EEE and facial information has been also used to recognise and classify human emotions. In [30], authors considered both types of data using the MAHNOB-HCI database [31]. Regression models were used in the whole set of data for emotion detection. As a result, better results have been obtained using the facial data while the good performance of the EEG results are due to the facial artefacts present in the EEG signal.

In this work a novel multimodal non-posed dataset is introduced. Due to the nature of our study, the environment where the RGB video is recorded is controlled avoiding illumination variations and occlusions. Moreover, the head-pose variations are minimal since the video sequences are recorded while the participants are looking at the screen in front of them. Using this novel multimodal dataset a method based on expected emotions is presented. These emotional behaviours are not defined as specific standard emotions, therefore our approach does not use FACS or any other emotion coding system. The facial modality presented uses geometric based spatio-temporal features. For the depth data and the landmarks, a new feature is introduced based on the t-Stochastic Neighbour Embedding (t-SNE) technique. These modalities are studied and compared with with state of the art methods.

3 Proposed methodology for spontaneous cognitive behaviours analysis

In this section we present the proposed method for recognition of spontaneous behaviours to specific visual stimulus. A novel multimodal database is introduced

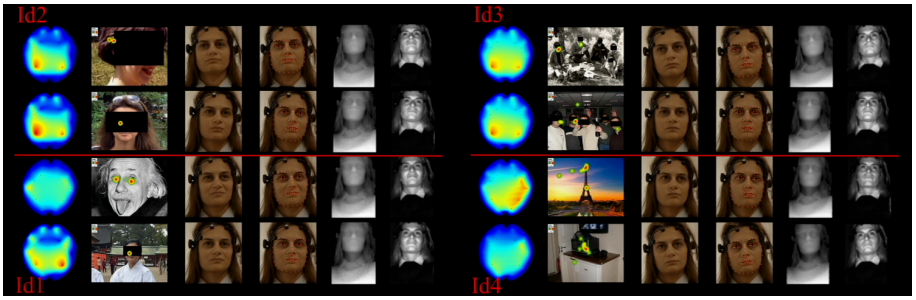

Fig. 1. Data modalities contained in the database and the related classes analysed in our approach (see Table 1 for the emotion definitions). The left figure shows from top to bottom, images of people from distant vs recent past; and famous vs unknown people. The right figure shows from top to bottom images of group of people from distant vs recent past; and famous vs unknown places. The different modalities from left to right in each case are EEG, gaze tracked heat map, RGB, facial landmarks, depth and IR.

and the proposed features utilized for the recognition of cognitive behaviour are described.

3.1 Spontaneous Emotion Multimodal Database (SEM-db).

SEM database is a novel non-posed dataset that can be utilized for recognition of emotional reactions to autobiographical and non-autobiographical visual stimulus data. The main novelty and contribution of SEM dataset is the use of personalized images for each participant. In fact, the dataset contains multimodal information of participants instinctive reactions when images related and unrelated with their personal life stories are shown. For each participant photos of themselves or their friends and relative from recent and distant past have been collected. In order to obtain a spontaneous and instinctive response, the participants did not know that those images were about to be used during the tests. Additionally, photos of famous persons and places, and photos of unknown to the participant persons and places were shown (see Fig. 1).

In total nine participants, aged between 30 and 60 years old of different educational background, were involved in this cognitive test. Each participant was asked to observe 80 images. In details, we have 8 different classes of images corresponding to an expected spontaneous cognitive behaviour: images of faces of relatives and themselves in the distant past (10 images), images of faces of relatives and themselves in the recent past (10 images), images of group of relatives including themselves, in the distant past (10 images), images of group of relatives including themselves, in the recent past (10 images), images of famous people (10 images), images of unknown persons (10 images), images of famous places or objects (10 images), and images of unknown places or objects (10 images).

The different reactions related to the above images were captured in the SEM database using different modalities: High Definition (HD) RGB, depth and

Infrared (IR) frames of the face, EEG signal and eye gaze data. In order to record the data, 4 different devices has been used: a 30 fps HD RGB camera, Kinect IR/Depth sensors, Tobii eye tracker and Emotive Headset (EEG sensors). The data have been recorded in a controlled environment, such as an office. The participants were provided with the instructions of the test before the start and they were asked to wear the EEG headset and to sit on a chair in front of the test screen, the RGB camera, the eye tracker and the Kinect sensors (see fig. 2). The test begins when the eyes tracker detects the eyes and the EEG sensors acquire good quality signal. A red image is displayed at the beginning of each test for synchronization purposes.

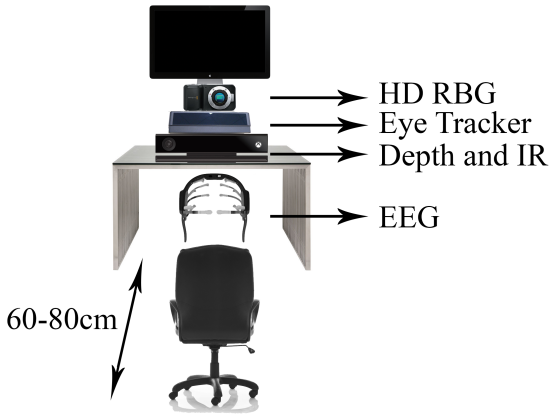

Fig. 2. Location of the devices during the recording of the database.

3.2 Cognitive and emotional behaviour recognition using depth and EEG data based on t-SNE manifolds.

The proposed method aims to classify the spontaneous emotional behaviour of the participants undergoing cognitive tests. Our aim is to detect strong reactions due to the recognition of famous persons or places versus the indifference reaction of the participants, while they are observing unknown subjects and objects. Also, we expect a strong reaction when images of distant past are displayed. A summary of the classes chosen for recognition and the corresponding expected reactions are shown in Table 1.

In our approach features are extracted from two data modalities: 3D facial points from depth data acquired using Kinect and EEG signals. As shown in Fig. 3, two feature descriptors were used in our classification problem based on the corresponding modalities.

Table 1. Classes chosen for recognition and the expected reaction.

Id	Class 1	Class 2	Expected emotion
1	Famous and familiar faces	Unknown faces	Recognition vs Neutral reaction
2	Distant past images of relatives and friends faces	Recent past images of relatives and friends faces	Long term memory recognition vs short
3	Distant past images of group of people including family and friends faces	Recent past images of group of people including family and friends faces	Long term memory recognition vs short term memory recognition
4	Famous places/objects	Unknown places/objects	Recognition vs Neutral reaction

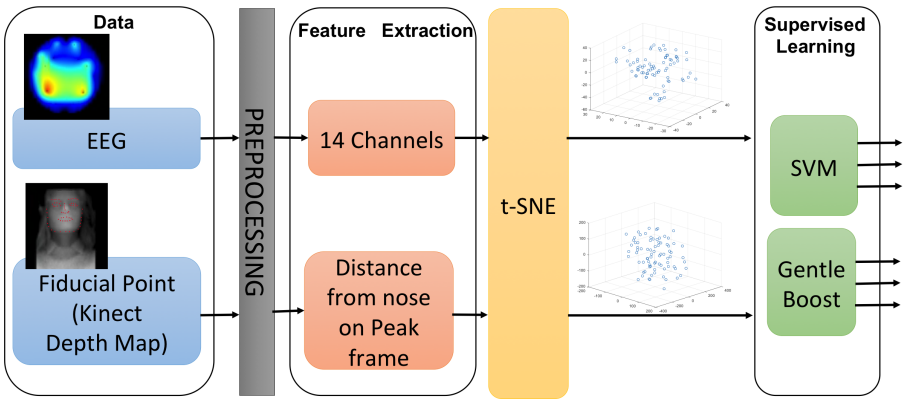

Fig. 3. Diagram representing our approach using the two modalities of data (EEG in the upper part of the diagram and Fiducial Points from depth images in the lower part of the diagram). It is possible to go through the process independently or combining the two modalities concatenating the features extracted from the fiducial points with the EEG features.

The fiducial facial points were obtained from a 30 frame rate IR video using the methods in [1, 19], obtaining 68 facial fiducial points per frame. The third coordinate is obtained from the depth data since their correspondence with the IR frames is provided. In order to obtain rigid head motions invariant features, the coordinated of the fiducial facial points were normalized according to a neutral face points, which correspond to the nose. After preprocessing the data, the feature vectors were extracted.

We expect that at the beginning of each test, before showing the image, the subject is in a neutral pose. Therefore, we select as neutral face the one that correspond to the first frame and we compare each frame to it. The frame that results more distant from the neutral face is selected as the peak frame. The distance of these points to the nose point were measured in the peak frame and all of them are combined in a single feature vector.

Regarding the EEG signals, they were acquired from 14 sensors at 128Hz using an EEG Headset. These 14 channels represent the features that have been obtained from the EEG data. During the preprocessing stage median filter was applied to remove the noise and the mean was subtracted in each channel.

The binary classification is performed separately using the features extracted from the depth frames and the EEG data. Once the features vectors are structured properly, we apply t-Stochastic Neighbour Embedding (t-SNE) method in order to reduce the number of selected descriptors generating a manifold representation.

In more details, t-SNE [33] is a non-linear dimensionality reduction technique used to embed high-dimensional data into a low-dimensional space (e.g., two or three dimensions for human-intuitive visualization). Given a set of N high-dimensional faces of people under different illumination conditions (i.e. data-points) $x_1, ..., x_N$, t-SNE starts by converting the high-dimensional Euclidean distances between data-points ($\|x_i - x_j\|$)into pairwise similarities given by symmetrized conditional probabilities. In particular, the similarity between data-points x_i and x_j is calculated from (1) as:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \tag{1}$$

where $p_{i|j}$ is the conditional probability that x_i will choose x_j as its neighbour if neigbours were picked in proportion to their probability density under a Gaussian centred at x_i with variance σ_i^2, given by (2):

$$p_{i|j} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_k - x_i\|^2 / 2\sigma_i^2\right)} \tag{2}$$

In the low-dimensional space the Student-t distribution (with a single degree of freedom: $f(x) = 1/(\pi(1 + x^2))$) that has much heavier tails than a Gaussian (in order to allow dissimilar objects to be modelled far apart in the map) is used to convert distances into joint probabilities. Therefore, the joint probabilities q_{ij} for the low-dimensional counterparts y_i and y_j of the high-dimensional points x_i and x_j are given by

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}. \tag{3}$$

The objective of the embedding is to match these two distributions (i.e., (1) and (2)), as well as possible. This can be achieved by minimizing a cost function which is the Kullback-Leibler divergence between the original (p_{ij}) and the induced (q_{ij}) distributions over neighbours for each object

$$D_{KL}(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{4}$$

The minimization of the cost function is performed using a gradient decent method which have the following simple form:

$$\frac{\delta D_{KL}}{\delta y_i} = 4 \sum_j \frac{(p_{ij} - q_{ij})(y_i - y_j)}{(1 + \|y_i - y_j\|^2)} \tag{5}$$

The reduced feature vectors (manifolds) from the two data modalities represent the input to two supervised learning algorithms: SVM and GentleBoost. Features from eight participants have been used for training, and data from one participant have been used for testing, according to the leave one out approach. The final results are the average of all the iterations, since k-folding cross-validation has been applied.

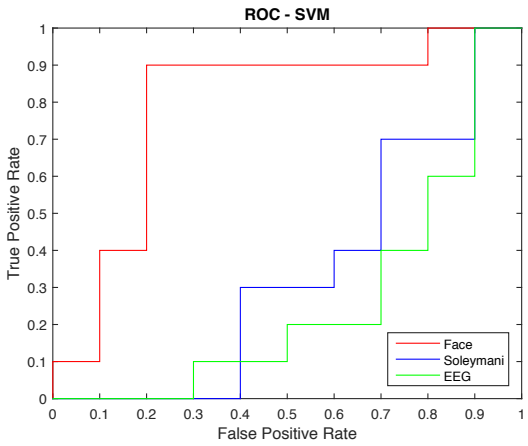

Fig. 4. ROC curve of the proposed method based on facial features in comparison to the ones proposed by Soleymani.

4 Results

This section shows and analyses the recognition results obtained using the facial and EEG features using SVM and gentleboost classifiers. The results are represented by the F1 score, which is a measure of accuracy that takes into account the precision and recall.

Tables 2 and 3 show the F1 scores for all the modalities and both classifiers, SVM and gentleboost, respectively. Furthermore, the ROC curves of the proposed method based on facial and EEG features in comparison to the ones proposed by Soleymani is shown in figure 4. The results of both individual modalities (EEG and facial) are coherent and adequate for the detection of emotions with overall F1 values above 70%. Comparing both data modalities, face depth data provide slightly better results than EEG for both classifiers. The classifiers have provided similar results, with SVM to result more accurate estimates in the case of face depth data. The proposed facial features also provide better F1 scores than the state of the art in most of the classifications. On the other hand, the EEG features are not so consistent and the best results are provided using gentleboost.

The classification of famous faces vs unknown faces, this is, the recognition of expected recognition vs neutral emotions, have been the recognised with lower accuracy in both cases of facial and EEG data. On the other hand, the classification of distant past faces vs recent past faces and distant past groups vs recent past groups, or the classification long vs short term memory reactions resulted higher recognition rates when facial features were utilised.

Table 2. F1 scores obtained using SVM classifier. See Table 1 for id information.

SVM	id 1	id 2	id 3	id 4	Overall
EEG	0.6069	0.6854	0.7106	0.7416	0.6861
Face	0.7035	0.7776	0.7276	0.6001	**0.7022**
Soleymani [30]	0.6235	0.6699	0.6722	0.6942	0.6650

Table 3. F1 scores obtained using Gentleboost classifier. See Table 1 for id information.

Boost	id 1	id 2	id 3	id 4	Overall
EEG	0.6307	0.6789	0.7183	0.7338	0.6904
Face	0.6646	0.7826	0.6871	0.7143	**0.7121**
Soleymani [30]	0.7068	0.7362	0.7295	0.6579	0.7076

5 Conclusions

In this work a novel database (SEM-db) has been introduced focused on natural reactions to specific autobiographical and non-autobiographical stimulus that intend to elicit different emotions. This database provides facial videos captured using depth sensors and EEG signals, amongst other information, that can be used for cognitive behaviour analysis. Using this database this work presents an approach for expected emotional behaviour recognition comparing state of the art and novel descriptors. Furthermore, advanced features were suggested based on t-SNE manifolds providing accurate representation of the depth information and the EEG data. Overall the face depth representation provides more accurate classification rates in comparison to the other descriptors and data modalities utilised in our comparative study.

References

1. T. Baltru, P. Robinson, and L. P. Morency. Openface: an open source facial behavior analysis toolkit. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016.
2. V. Bettadapura. Face expression recognition and analysis: the state of the art. *Tech Report arXiv:1203.6722*, pages 1–27, 2012.
3. Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. M. Siskind, and S. Wang. Recognize human activities from partially observed videos. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2658–2665, June 2013.
4. Alexandros Andre Chaaraoui and Francisco Florez-Revuelta. Optimizing human action recognition based on a cooperative coevolutionary algorithm. *Engineering Applications of Artificial Intelligence*, 31:116–125, 2014.
5. M. A. D. Chowdhuri and S. Bojewar. Emotion detection analysis through tone of user: A survey. *Emotion*, 5(5):859–861, 2016.
6. T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.
7. J.W. Davis and A. Tyagi. Minimal-latency human action recognition using reliable-inference. *Image Vis. Comput.*, 24(5):455–472, May 2006.
8. P. Ekman and W.V. Friesen. The facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press, San Francisco*, 1978.
9. K. C. Huang, S. Y. Huang, and Y. H. Kuo. Emotion recognition based on a novel triangular facial feature extraction method. *2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, 2010.
10. C. E. Izard. Human emotions. *Springer Science & Business Media*, 2013.
11. T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.
12. S. Koelstra, C. Muehl, M. Soleymani, J.S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

13. Y. Kong, D. Kit, and Y. Fu. A discriminative model with multiple temporal scales for action prediction. *in ECCV 2014 - European Conference on Computer Vision*, pages 596–611, 2014.
14. T. Lan, T. Chen, and S. Savarese. A hierarchical representation for future action prediction. *Comput. VisionECCV*, pages 689–704, 2014.
15. K. Li and Y. Fu. Arma-hmm: A new approach for early recognition of human activity. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1779–1782, Nov 2012.
16. G. C. Littlewort, M. S. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, 27(12):1797–1803, 2009.
17. S. Lokannavar, P. Lahane, A. Gangurde, and P. Chidre. Emotion recognition using eeg signals. *Emotion*, 4(5):54–56, 2015.
18. G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.
19. P. Michel and R. El Kaliouby. Real time facial expression recognition in video using support vector machines. *In Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264, 2003.
20. G. R. Mller-Putz, R. Riedl, and S.C. Wriessnegger. Electroencephalography (eeg) as a research tool in the information systems discipline: Foundations, measurement, and applications. *Communications of the Association for Information Systems*, 37(46):911–948, 2015.
21. M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.
22. J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. *14th ACM international conference on Multimodal interaction*, pages 501–508, 2012.
23. M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. *IEEE international conference on multimedia and Expo*, pages 317–321, 2005.
24. I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 97–102, 2004.
25. P. C. Petrantonakis and L. J. Hadjileontiadis. Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis. *IEEE Transactions on Affective Computing*, 1:81–97, 2010.
26. M.S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. *Int. Conf. Comput. Vis., no. Iccv*, page 10361043, Nov 2011.
27. E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2015.
28. E. Sariyanidi, H. Gunes, M. Gkmen, and A. Cavallaro. Local zernike moment representation for facial affect recognition. *British Machine Vision Conf*, 2013.
29. A. T. Sohaib, S. Qureshi, J. Hagelbck, O. Hilborn, and P. Jeri. Evaluating classifiers for emotion recognition using eeg. *International Conference on Augmented Cognition. Springer Berlin Heidelberg*, pages 492–501, 2013.

30. M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic. Analysis of eeg signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, 7(1):17–28, 2016.
31. M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.
32. M. Szwoch and P. Pieniek. Facial emotion recognition using depth data. In *2015 8th International Conference on Human System Interaction (HSI)*, pages 271–277, June 2015.
33. G van Der Maaten, L ; Hinton. Visualizing data using t-sne. *Journal Of Machine Learning Research*, 9:2579–2605, November 2008.
34. R.L. Vieriu, S. Tulyakov, S. Semeniuta, E. Sangineto, and N. Sebe. Facial expression recognition under a wide range of head poses. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–7, May 2015.
35. A. E. Vijayan, D. Sen, and A. P. Sudheer. Eeg-based emotion recognition using statistical measures and auto-regressive modeling. *IEEE International Conference on Computational Intelligence & Communication Technology (CICT)*, 14(1):587–591, 2015.
36. F. Weninger, M. Wllmer, and B. Schuller. Emotion recognition in naturalistic speech and languagea survey. *Emotion Recognition: A Pattern Analysis Approach*, pages 237–267, 2015.
37. M. Wllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. *Interspeech*, pages 597–600, 2008.
38. W. J. Yan, X. Li, S. J. Wang, G. Zhao, Y. J. Liu, Y. H. Chen, and X. Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014.
39. G. Zhao and M. Pietikinen. Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. *Pattern recognition letters*, 30(12):1117–1127, 2009.