FusionSense: Emotion Classification using Feature Fusion of Multimodal Data and Deep learning in a Brain-inspired Spiking Neural Network

March 2020

1 Abstract

Using multimodal signals to solve the problem of emotion recognition is one of the emerging trends in affective computing. Several studies haved utilized state of the art deep learning methods and combined physiological signals such as electrocardiogram(EEG), electroencephalogram(ECG), skin temperature along with facial expressions, voice, posture to name a few , to classify emotions. Spiking neural networks (SNNs) represent the third generation of neural networks and employ biologically plausible models of neurons. SNNs have been shown to handle in an efficient way spatio-temporal data, which is essentially the nature of data encountered in emotion recognition problems. In this work, for the first time, we propose the application of SNNs to solve the emotion recognition problem with multimodal datasets. Specifically, we use the NeuCube framework, which employs an evolving SNN architecture, to classify emotional valence and evaluate the performace of our approach on the MAHNOB-HCI dataset. The multimodal data used in our work consist of facial expressions along with physiological signals such as ECG, skin temperature, skin conductance, respiration signal, mouth length and pupil size. We perform classification under Leave-One-Subject-Out (LOSO) cross validation mode. Our results show that , the proposed approach achieves an accuracy of 73.15% for classifying binary valence, when applying feature-level fusion, which is comparable to other deep learning methods. We achieve this accuracy even without using EEG data, which other deep learning methods have relied on to achieve this level of accuracy. In conclusion we have demonstrated that the SNN can be successfully used for solving the emotion recognition problem with multimodal data and also provide directions for future research utilizing SNN for affective computing. In addition to the good accuracy, the SNN recognition system is incrementally trainable on new data in an adaptive way and requires only one pass training, which makes it suitable for practical and on-line applications. These features are not manifested in other methods for this problem.

2 Introduction

The central aim of affective computing is to enable seamless communication between humans and computers by developing systems that can detect and

respond to the various affective states of humans [1]. Affective computing is an interdisciplinary field of research involving expertise from computer science, psychology, social and cognitive sciences. Affect recognition has important applications in several fields such as medicine [2], driver fatigue monitoring, human-computer interaction, sociable robotics [3] and security systems, to name a few. Modelling affect can be clasified into three categories: categorical; dimensional; and components. Categorical models classify emotions into a set of discrete classes, which are easy to describe and these include six basic emotions such as happiness, sadness, fear, anger, disgust and surprise. Owing to its simplicity, categorical models have been extensively utilized in affect research. In contrast, dimensional models represent emotion as a point in multidimensional space, where the dimensions include valence, activation and control, allowing the description of more complex and subtle emotions. However, such multidimensional space can pose significant challenge to an automatic emotion recognition system and thus researchers have mostly used the simplified two-dimensional model of arousal and valence proposed in [4], where arousal ranges intensity of emotion from calm to excited, and valence ranges from unpleasant to pleasant [5].

The component model of emotions arrange emotions in a hierarchical fashion, where complex emotions can be derived from the combination of a pair of basic emotions. The most popular component model proposed by Plutchik [6] is based on evolutionary principles and has eight basic bipolar emotions.

Affect can be expressed via facial expression, body movements, voice behavior, gestures and an array of physiological signals such as heart rate, sweat, pupil diameter, brain signals to mention a few. The problem of recognizing emotions by utilizing facial expressions from videos and static images have been addressed by several studies [7, 8, 9]. Advances in deep learning methodologies have created huge interest in application of such methods in facial emotion recognition (FER) [10, 11, 12, 13, 14], most of which are based on supervised learning. The methods do not allow incremental, adaptive learning on new data and are not suitable for on-line applications. For an excellent overview of the application of deep learning and as well as shallow learning approaches to FER , the reader is directed to [15] and the references there in.

Spiking neural networks (SNNs) represent the third-generation of neural networks, modelling neurons and interactions between them in a biologically more realistic manner compared to second-generation neural networks based on ANNs. SNNs are an ideal choice to handle the emotion recognition task from video data, given their ability to handle spatio-temporal data effectively [16](see section 5 for details).

In this work, we propose to build FER system using SNNs. To this end, we use the NeuCube framework (Kasabov2014neucube), which is a type of evolving SNN (eSNN). In this paper we develop an encoding method to map the continuous facial feature values to spikes based on population coding. We use the data from Mahnob-HCI dataset to test the NeuCube framework for the classification of binary valence in response to video stimuli.

The structure of the paper is organized as follows. In section 3 we provide some background literature on various data modalities used in affect detection,

whereas in section 4 we describe strategies for multimodal data fusion. In section 5 we provide some background on SNN and the NeuCube framework. Section 6 details the methodology used in our work and section 7 presents the results. In section 8 we discuss our results and in section 9 direction for future work and conclude the paper.

3 Signals for affect detection

3.1 Facial expression

One of the immediate and natural ways for humans to communicate their emotions is through facial expressions, which constitute about 55 % of the information communicated during face to face human interaction [17]. Thus, affect research has primarily focussed on detecting emotions from the face. Research on facial emotions have shown that the six basic emotions such as fear, anger, sadness, enjoyment and disgust can be detected with facial expressions[18, 19] and dectecting an emotion is equivalent to detetecting the associated prototypic facial expression. Based on the Facial Action Coding System (FACS), which originally described 44 single action units (AU) including head and eye movements, with each action unit linked with an independent motion on the face and the correponding muscles, for example lip suck motion with the muscle orbicularis oris [20]. Several deep learning techniques have been used to build automatic facial emotion recognition (FER) system , including deep boltzmann machine (DBM), deep belief networks (DBNs) [21, 22, 23], convolutional neural networks (CNNs) [24, 25, 26, 27, 11], auto-encoders [28, 29, 30] and recurrent neural networks (RNNs)to mention a few.

3.2 Speech

Affective information from speech can contain lingusistic and paralinguistic features, which refer to what is said and how it is said, respectively. Although speech is a fast and efficient method of communication which can be exploited in affect research, detecting the emotional state of the speaker using speech signal is still a significant challenge. There is no clarity on which features of the speech signal are most powerful is distinguishing different emotions. It has also been shown that, compared to facial expressions, the accuracy of affect detection from speech is lower [31]. For instance, the basic emotions such as sadness, anger and fear can be recognized using speech, where as disgust is hard to detect [1]. Moreover, cultural differences among speakers has not been adressed thoroughly with most of the affect research involving speech focussing on monolingual emotion classification [31]. The features that are typically extracted from speech signal include both global and local features, Local features refer to pitch and energy extracted from small segments into which a speech signal is typically divided to make it stationary, whereas global features refer to statistics of all the local features extracted from a long signal.Studies have

shown that global features have better classification accuracy than local features [32, 33, 34]. However, studies have shown that global features cannot distinguish between emotions that have similar arousal [35] and may prove to be sub-optimal when using classifiers such as Hidden Markov Model (HMM) and Support Vector Machines (SVM) due to insufficient number of training vectors [31]. Since the properties of the different speech sounds can be altered by different emotions , some studies have also explored the benefits of phoneme-level modeling for the classification of emotional states from speech rather than using the prosodic features such as pitch and energy [36]. Their results showed that the using phoneme-class classifiers outperformed HMM classifiers just based on global features. Apart from using HMM or SVM classifiers, several deep learning techniques have been explored for emotion recognition from speech signls including DBM [37, 38], auto-encoders [39, 40], DBNs[41, 42] and CNNs [43, 44, 45], to cite a few. Despite the aforementioned challenges, speech is still an important signal that can be used for affect detection as it is non-intruive and has high temporal resolution.

3.3 Posture and body movements

In comparison to speech and facial expression, perceiving emotions through body movements and postures is a relatively less explored topic in affect research. In fact, 95 % of the literature in research on human emotions focusses on facial expressions and less than 5 % on speech and other physiological signals with the remaining little of body movements. Several studies in the past have shown that body movements and postures can contribute to the recognition of emotional states [46, 47], with perhaps the most influential work in this topic dating back to the second half of 19th century by Charles Darwin [48]. Body postures may offer certain advantages in affect detection given the multiple degrees of freedom human body posseses, which can aid in communication of emotions and susequently affect detection, even at long distances, at which facial emotions are unreliable [49], indicating that postures contain information not present in facial expressions. Another advantage of posture-based affect system could be that in comparison to facial expression which may be intentionally controlled, postures and body movements are unintentional and thus less susceptible to social editing [1]. In a study on deception by Eckman and Friesen [50], it was shown that liars were less successful at deception through body movements compared to more controlled channels of communication such as facial emotions, which they referred to as nonverbal leakage. Gestures, which can be defined as collection of body movements or actions involving head, hands and other parts of the body allow the communication of a range of thoughts and emotions. Some of the basic gestures have been shown to be similar across the cultures. Given the advantages of this non-verbal communication channel, relatively few studies have utilized deep or machine learning framework to recognize emotions using body movements, postures and gestures [51, 52].

3.4 Physiological signals

Physiological signals such as electroencephalography (EEG), electrocardiogram (ECG), electromyogram (EMG), skin conductance, also known as Galvanic skin response (GSR), skin temperature and as well as pupilary diameter can be used for affect detection, apart from the above mentioned non-physiological signals. Physiological signals for affect detection are typically acquired in a non-invasive manner using wearable sensors. Heart rate (HR) and hear rate variability (HRV) can be derived from ECG signals. Skin temperatue has been shown to be a effective indicator of the emotional state as shown in [53] and it primarily reflects the activity of the autonomic nervous system (ANS). Another modality that captures the activity of ANS is the GSR or skin conductance , which can be obtained by measuring the electrical potential on the skin after passing a negligible amont of current. GSR is considered to be a reliable indicator of arousal [54] , as it captures the activity of the sweat glands on the skin.

In affect research , ECG signals are typically recorded by a pair of electrodes,which are a subset of lead I configuration comprising of 12 electrodes. Features such as HR and HRV can be further derived from ECG that can reflect the activity of the sympathetic and parasympathetic branch of ANS system. Both HR and HRV have been used in several studies to asses the mental states of the subject [55, 56]. An EMG signal is reflective of the strength of muscle movements and is typically recorded by a pair of electrodes placed on the body. Studies have shown that when the subject is under some emotional stress, the changes in the facial expression can be measured using EMG activity [57, 58]. Apart from using electrodes on the face, other studies have also looked into measuring the activity of jaws or shoulders to idendity emotional states [59].

Breathing is another physiological process that is shown to be altered by basic emotions such as happiness, sadness and anxiety [60]. Researchers have observed rapid breathing during arousal state [61] and as well as changes in respiratory pattern of subjects looking at photographs that induce emotions [60]. Respiratory rate is shown to be modulated by emotions, particularly anxiety affecting the expiration rate [60], where as timing and volumetric aspects of breathing are altered by various physical and mental stress [62].

Finally, EEG is probably the most widely used physioloical signal to study emotion. EEG is a low cost technology compared to other neuroimaging modalities and has very good temporal resolution. EEG electrodes record the activity of a large number of synchronous neurons as potential difference on the scalp. Several studies have utilized EEG for emotion recognition [63, 64, 65] and classification of emotional states of arousal, valence and dominance. In addition to EEG, pupilary diameter size is also an indication of emotonal state, with several studies reporting that the size of the pupil discriminates during and after different kinds of emotional stimuli [66, 67].

Several deep learning methodologies have been utilized for emotion recognition using physiological signals [68, 69, 70, 71]. For an exhaustive list of literature, the reader is directed to [1]

4 Multimodal affect recognition

Although majority of the machine learning and deep learning framework for affect recognition uses data from one modality, i.r., video or audio or EEG, recently there has been a considerable interest in fusing data from the above mentioned modalities. Multi-sensor data fusion can be highly advantageous in terms of improving the reliability and accuracy of affect detection and furthermore, multimodal systems have shown to outperform unimodal system as discussed in [72]. Multimodal fusion involves combining data from many different types of sensors and such fusion can be performed primarily at two distinct levels, known as feature-level fusion and decision level fusion.

4.1 Feature-level fusion

In feature-level fusion approach (also known as early fusion), features derived from different modalities are combined into a single feature vector, on which a classifier can then be trained. It is well known that humans use and integrate multiple sensory cues during face-to-face interaction to detect affective states and is the fundamental idea behind feature-level fusion [73]. The main advantage of feature-level fusion is that correlation between multimodal features at an early stage can lead to better performance, requiring only one learning phase on the feature vector. Several studies have utilized this approach for affect research [74, 8, 75]. However, feature-level fusion also has several challenges. Since features obtained from different modalities can have different time-scales, achieving time synchronization to bring the features in same format can be difficult and computationally expensive. Also, given the large feature set one obtains with feature-level fusion, the classification accuracy can be severly affected if the training dataset is limited. Furthermore, learning cross-correlation between the heterogenous features can prove to be difficult [76].

4.2 Decision-level fusion

In decision-level fusion approach (also known as late fusion), first the decisions based on features derived from each modality is obtained separately. A fused decision vector is then obtained using the local decisions, which can be used to obtain the final decision or classification [76]. The fundamental advantage of decision-level fusion over feature-level fusion is that the decisions all have the same format and hence can be fused easily, thus avoiding synchronization issues. Furthermore, using decision-level fusion allows the application of optimal classifier or method suited for each modality, thus providing more flexibility compared to feature-level fusion [77]. Several studies have utilized decision-level fusion for affect research [78, 79, 80] and it has been noted that researchers prefer decision-level fusion over feature-level fusion [77].

5 Spiking neural networks

Human brains encode information via discrete events known as action potentials or spikes, following an all-or-none principle, where a neuron fires an action potential if the stimulus crosses a certain threshold, else it remains silent. Due to this binary nature of information representation , the human brain still outperforms the existing artificial neural networks (ANNs) in terms of both energy and efficiency [81, 82]. Compared to the traditional ANNs, spiking neural networks (SNNs) utilize a more biologically realistic models of neurons [83], thus bridging further the gap between neuroscience and learning algorithms. SNNs have shown the ability to integrate information from different dimensions such as time, phase, frequency and as well as handle large volumes of data in an adaptive and self-organized manner [84, 16],making them particularly suitable to solve online spatio-temporal pattern recognition. SNNs have been shown to be computationally more efficient than ANNs both theoretically [85, 86] and in several real-world applications [87]. SNNs have been used in several real-world learning tasks such as unsupervised classification of non-globular clusters [88], image segmentation and edge detection [89], epileptic seizure detection with EEG [90]. Furthermore, Bohte and colleagues devised a supervised learning rule for the SNNs and denonstrated its application in the XOR classification problem and several other benchmark datasets [87]. The evolving SNN (eSNN), is a class of SNN that utilizes rank order learning [91] and was first proposed in [92]. The eSNN handles spatio-temporal data by increasing the number of spiking neurons in time to learn temporal patterns from data [93]. In addition to the open evolving structure of eSNNs that facilitate addition of new variables and neuronal connections, eSNN have the advantage of fast learning from large amounts of data and can interact with other systems actively . eSNNs also allows the integration of various learning rules such as supervised learning, unsupervised learning, fuzzy rule insertion and extraction, to mention a few and are self-evaluating in terms of system performance. These aforementioned properties consitiute the evolving connectionist systems (ECOS) principles principles on which the eSNN is based [94].

Since in the rank-order learning scheme, the synaptic weights are adjusted only once making it not very efficient for spatio-temporal data, where there may be a need to adjust synaptic weights based on the spikes arriving on the same synapse over time. To overcome this disadvantage, an extension of eSNN known as dynamic eSNN (deSNN) was introduced in [84] that combines rank-order learning with temporal learning rules such as spike-timing dependent plasticity (STDP), which allows dynamic adjustment the synaptic weights . However, both eSNN and deSNN do not encapsulate the structural information of the brain in terms of neuronal locations and their connectivity, which may be crucial for modelling spatio-temporal data . The NeuCube architecture, first proposed in [95], aims at building a eSNN that incorporates structural and as well as functional aspects of the brain along with utilizing STDP learning rules. The following section gives a brief introduction to the NeuCube architechure. For a more detailed introduction, the reader is directed to [95, 96, 97]

5.1 NeuCube

It is well known that the information in human brain is processed at different spatiotemporal levels ranging from molecular information processing to higher order cogitive processes . Data can be acquired at different levels of these spatiotemporal processes and an efficient learning method should be able to handle complex spatio-temporal relation ship from brain data at different levels. Some examples of spatio-temporal brain data (STBD) include EEG, functional magnetic resonance imaging (fMRI), diffusion tensor imaging (DTI), and positron emission tomography (PET) to mention a few. Traditional methods such as support vector machines (SVM) or multilayer perceptron neural networks (MLP) typically deal with the spatial or temporal aspects of brain data and cannot handle the dynamic interaction between these processes [97]. Furthermore, they cannot incorporate any structural prior knowledge of the brain or handle multimodal brain data. NeuCube [96] and also [94, 95,98,99] is a variant of eSNN , initially proposed to handle problems of spatio-temporal pattern recognition in brain data such as EEG, functional magnetic resonance imaging (fMRI) to cite a few, has been further developed to handle various other types of spatio-temporal data such as audio-visual data, climate data, seismic data and ecological data [99]. The typical framework of the NeuCube system comprises of

1. An input encoding module, which converts the STBD into trains of spikes that captures temporal patterns present in the data. Various methods have been proposed to achieve this, including population coding [100], address event representation [101] and Bens Spike algorithm [94].

2. A three-dimensional SNN reservoir (3D-SNNr), which takes the spike trains as input . The 3D-SNNr contains neurons that have pre-defined spatial coordinates and are modelled as leaky integrate and fire neurons. The initial structural connections between the neuroons can be established in several ways including small-world organization [102] or based on the DTI data. Several studies utilizing EEG, fMRI and MEG have demon-strated the presence of small-world connectivity in the brain [103, 104] and thus, this is the preferred initial setup for the spatial structure of 3D-SNNr. Based on the temporal association between the input spikes, connections between the neurons is modified using the spike timing dependent plasticity (STDP) rule. This is a deep unsupervised learning as deep connectionist structures of many neurons are created as a results of the learning in space and time [94].

3. A classification module, which takes the spiking patterns from 3D-SNNr as its input to perform classification.

4. An optional, Gene Regulatory Network (GRN) for controlling the learning parameter and optimization of 3D-SNNr, exploiting the fact that spiking activity is influenced by the gene and protein dynamics.

The details on the implementation of the NeuCube-based SNN models for this study are further described in section 6.6

6 Methods

6.1 Mahnob Database

The MAHNOB-HCI dataset is a multi-modal database for affect recognition and implicit tagging [105]. In this database, 27 subjects (16 females and 11 males) aged between 19 and 40 years old were monitored while watching 20 stimulus clips (34.9 to 117 seconds long) extracted from Hollywood movies and video websites, such as YouTube and blip. The face video, audio and elicited physiological signals (EEG, ECG, respiration amplitude, skin temperature, GSR and gaze data) were acquired while watching the clips. ECG signal was obtained by subtracting a measurement from the upper left corner of chest, under the clavicle bone, from that one on left side of abdomen, below the last rib. Respiration signal was obtained by a belt placed in the subject's abdomen, skin temperature was acquired by a temperature sensor placed at the subject's little finger and GSR was obtained by passing a negligible current between the electrodes on the distal phalanges of the middle and index fingers of the subject. Gaze data was acquired with Tobii X1205 eye gaze tracker providing position of the projected eye gaze on the screen (at 60 Hz), the pupil diameter, the moments when the eyes were closed, and the instantaneous distance of the subject's eyes to the gaze tracker device.

Physiological signals, except the gaze data, were acquired at a sampling rate of 1024 Hz (down sampled to 256 Hz for further analysis) while six different views of subject's facial expressions were recorded simultaneously by six video cameras at 60 fps. In this work, the video taken only by the color camera above the screen were used. After watching each stimulus, the participants used a keyboard interface for answering five questions related to emotional label, arousal, valence, dominance and predictability. Participants answered each question using nine numerical keys, selecting nine emotional labels for the first question and nine possible levels for the last question. In this work, only the binary valence scale was used where levels 1 to 5 were considered as low valence (unpleasant) and levels 6 to 9 as high valence (pleasant). The database is available online here.

The multimodal emotion recognition (valence) pipeline starts with face detection in video, followed by face landmark detection, features extraction from face and peripheral signals, and ends with training and signals classification using NeuCube.

6.2 Face detection and tracking

The first step for analysing face emotion recognition in video is face detection and tracking in frames. Computer Vision (CV) Matlab Toolbox was used for this task. The output of this step is the corner coordinates for the polygon enclosing the face for each frame in the video.

The face detection carried out in this work included the following steps,

1. The face in the first frame was detected using the *vision.CascadeObjectDetector* object in the CV toolbox. This function uses the Viola-Jones algorithm

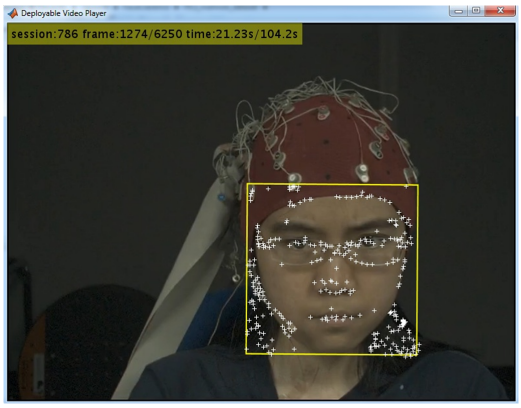


Figure 1: Example of face detection in Mahnob- HCI dataset showing the feature points tracked along the video

[106] to detect people's faces, noses, eyes, mouth, or upper body. It outputs the region of interest (ROI) for the face as a polygon, enclosing the face. Specifically, the algorithm uses the histogram-of-oriented gradients (HOG), Local Binary Patterns (LBP), Haar-like features and a cascade of classifiers trained using boosting.

2. The corner features in the first frame ROI were detected using the *detectMinEigenFeatures* function in CV toolbox, which uses the minimum eigenvalue algorithm [107].

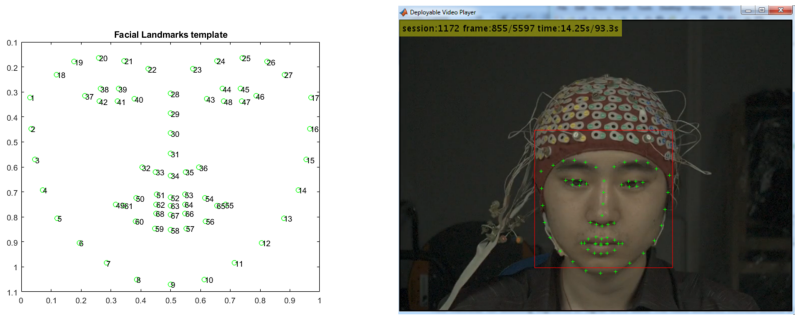
3. For tracking of feature points in the remaining frames, we used Kanade-Lucas Tomasi (KLT) algorithm [108, 107].

4. Finally, in order to estimate the motion of the face, we used *estimateGeometricTransform* function in the CV toolbox to apply the same transformation to the ROI detected in the previous frame to obtain the ROI in the next frame.

Figure 1 shows the output of the face detection step. We found that point tracking in frames to detect face is computationally more efficient than face detection in each frame. Furthermore, point tracking can manage problems that can emerge in face detection such as making gestures with hand that may occlude parts of the face.

6.3 Face landmarks detection

Using the detected ROIs (See section 6.2), a trained model (DLIB) for 68 facial landmarks detection was used for each frame in the video [109]. DLIB library can be obtained from this link. The processing time for this task was around 100 seconds per video (i.e., approximately 30 minutes per subject). Figure 2



(a) Model template

(b) Detected facial landmarks in one example video frame

Figure 2: Facial landmarks detection.

shows the model template (a) and one example video frame with detected facial landmarks (b) adjusted to relevant facial structures (mouth, eyebrows, eyes, nose and face borders).

6.4 Face features extraction

We extracted the following featured from facial landmarks (see figure 3),

1. Vertical distance between the horizontal line connecting the inner corners of the eyes and outer eyebrow (f1, f2).
2. Vertical distances between the upper eyelids and the lower eyelids (f3, f4).
3. Distances between the upper lip and mouth corners (f5, f6).
4. Distances between the lower lip and mouth corners (f7, f8).
5. Vertical distance between the upper and the lower lip (f9) and distance between the mouth corners (f10)

We assume that participants hold a neutral face while first 2 seconds after starting the stimulus. As we want to detect changes in facial features, therefore the mean features in first 2 s are subtracted from facial features for each response video.

6.5 Physiological Features

Heart rate variability (HRV), respiration variability, respiration depth, skin temperature, GSR and pupil diameter are used as physiological features in this study. ECG signal is pre-processed by mean subtraction and band pass filtered with a low pass and high pass filter in cascade (Least Square FIR, 70 dB, $0.05 - 40$ Hz, 1dB ripple) for reducing high frequency noise as muscular

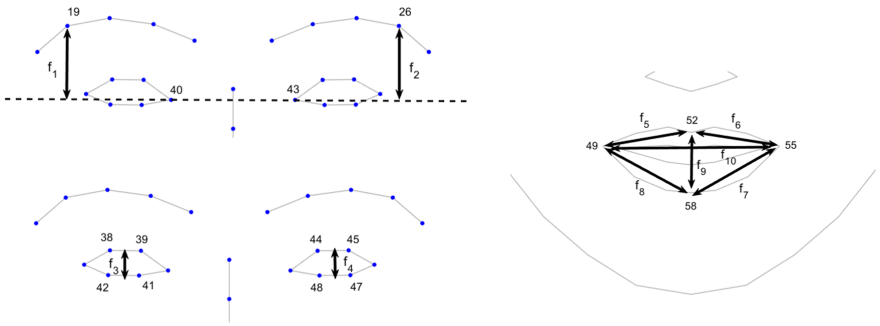


Figure 3: Facial features

activation and reducing shifting due to respiration. R waves are detected using Pan and Tompkins algorithm [110] for calculating the RR interval (for HRV) as a feature. The *findpeaks.m* function in Matlab (Signal Processing Toolbox) was applied to the respiration signal to detect valleys and peaks in signal and further obtain the respiration variability (time between cycles) and respiration depth (cycle amplitude). The raw Temperature (Celsius) and GSR measurements were also considered as feature signals. And from the gaze data the mean pupil diameter (from both eyes) was computed as an additional feature signal. Examples of physiological features are shown in Figure 4.

All facial and peripheral physiological features obtained in the analyzed window (last 30 seconds of video) were resampled to 64 samples. In order to capture changes in physiological feature, all features are calculated in whole video response too, and resampled to 64 points. The first sample is substracted from features in windows for further analysis. We suppose that this first measurement in whole video mean for resting or neutral state for physiological signals. Figure 5 shows the distribution of normalised features. It can be noted that mouth-related features and pupil size have better discriminative power between low and high valence. Outliers are omitted for visualisation purposes.

6.6 NeuCube SNN for facial emotion recognition

We used NeuCube proposed in [96] to build a system for emotion valence classification. A general scheme of our approach based on NeuCube is presented in Figure 6. As described in section 5.1, NeuCube structure includes Encoding, 3D-SNNr, output neuron layer and KNN classifier. Training and classifying spatio-temporal data using NeuCube have the following stages:

- **Encoding:** Encode the spatio-temporal data (features) into trains of spikes.
- **SNNr:** Construct a recurrent 3D SNNr and initialize the connection weights among neurons.

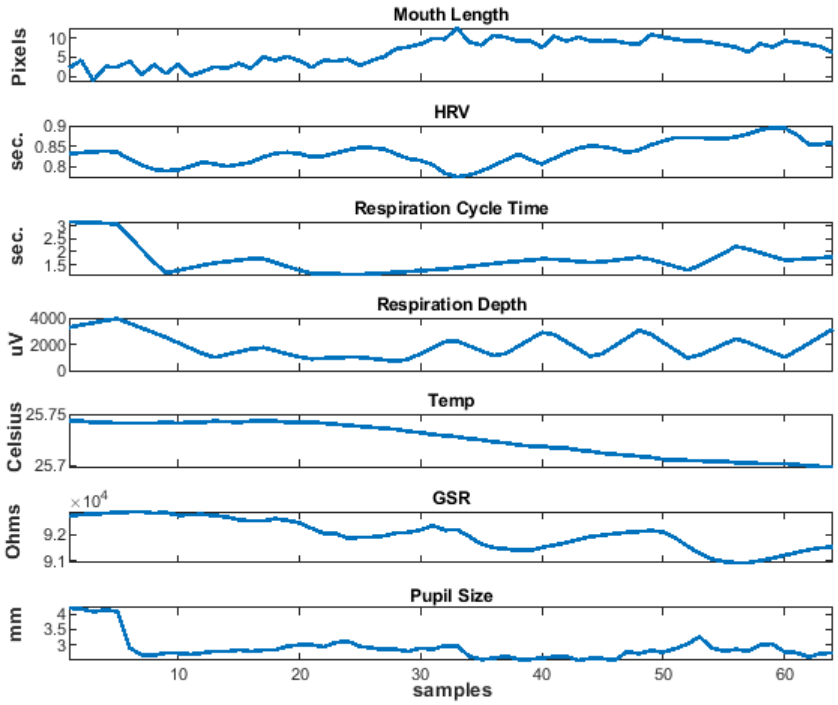


Figure 4: Elicited signal features in the last 30 seconds of video

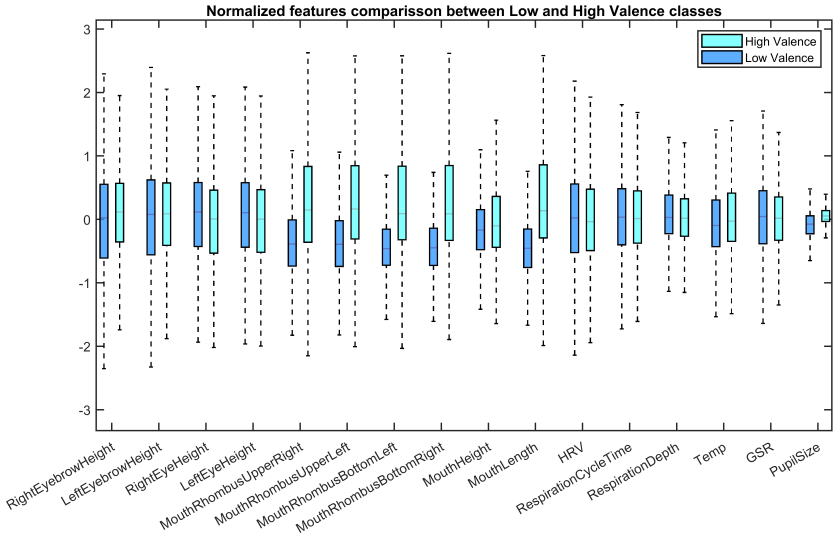


Figure 5: Boxplot for features in Mahnob-HCI dataset for valence emotional dimension

- **Input neurons location:** Locate the input neurons in the SNNr keeping related inputs near in space.
- Deep, **unsupervised learning:** Feed the SNNr with training data to learn in an unsupervised mode the spatio-temporal patterns in the data.
- **Supervised learning:** Construct an eSNN classifier to learn to classify different dynamic pattern in SNNr activities.
- **Classification:** Feed the SNNr with testing data for classification purposes.

We briefly explain each stage in the following sections,

6.6.1 Encoding

The coding method we used was inspired by Gaussian Receptive Field population-based sparse coding proposed in [88, 87]. This method codes each continuous value from a time-based feature to spikes emitted at different times by a neuron population [94]. The whole feature range is covered for the neurons and the time for generating the spikes depends on the distance from the current value to the centre of a Gaussian receptive field covering each value interval. We used a population of five neurons per feature, in which only a neuron from the group spikes at the current time step. Figure 7 shows an example of coding the mouth length feature. Note that the dimension of feature is 64 and the temporal dimension of each spikes train is 129 because zeros are inserted between the spikes.

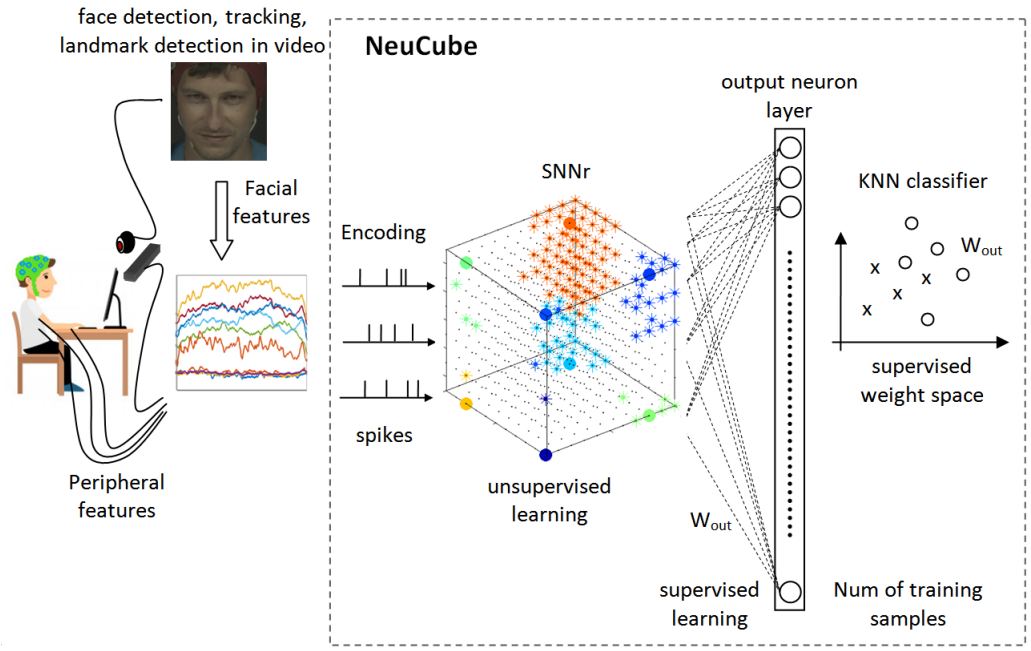


Figure 6: Proposed SNN method for emotion valence classification using the NeuCube framework

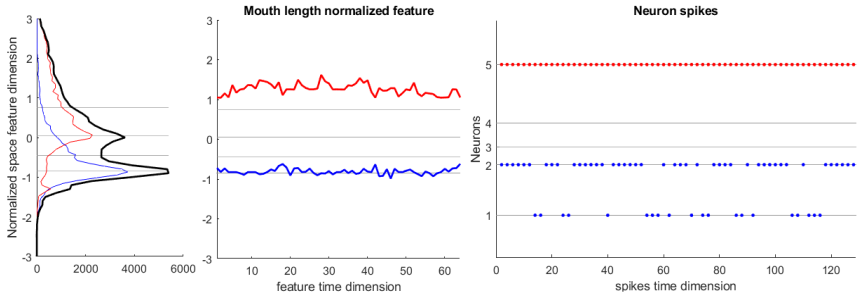


Figure 7: Encoding Continuous feature values to five spiking neurons

It can be noted from the distribution of mouth length feature (Figure 7, left plot; blue: low valence, red: high valence, black: low and high valence), that there are two peaks in the distribution indicating the separation between the two class. In the middle plot (Figure 7), the time course of mouth length feature in a low valence event (blue) and one high valence event (red) for the subject 1 are shown. In the right plot (Figure 7) spikes generated for these two events are shown (low valence in blue, high valence in red). Levels that define the receptive fields or range for exciting each neuron are defined using the feature distribution in the data from all detected events for all analyzed subjects. Levels for each five neuron population are obtained automatically by analyzing the histogram in such a way that the five ranges have the same count of value occurrences. Levels are shown as gray lines (Left and middle plots in Figure 7). Note that each feature value in time produces a spike in only one neuron from the population. As we have ten facial features, and six peripheral signal then eighty input neuron are allocated in NeuCube network.

6.6.2 Construction of SNNr

When brain imaging data such as EEG is used, the SNNr can be built with a shape resembling the human brain [96] and the input neurons can be located based on the anatomical location of the EEG electrodes. However, in this study, as we are building a general classifier of facial features, we chose to build an $11 \times 11 \times 7$ array of neurons (equally spaced in x and y axes) as shown in Figure 8. Each five neuron population are spatially arranged in NeuCube structure in lines as illustrated in Figure 8, this way neighbor neurons code similar feature values favoring spatial neuron specialization.

The SNNr was made with leaky integrate and fire model (LIFM) spiking neurons with recurrent connections. In this neuron model, the post-synaptic potential (PSP) increases or decreases with every input spike from pre-synaptic neurons. The effect of each spike is modulated by the corresponding synaptic connection weight. If PSP reaches a specific threshold (0.5 in this work), the neuron emits an output spike toward its connected neighbours and the PSP resets to a reference value. The PSP can leak between spikes with a predefined time constant τ, if we are using an exponential model or a constant leak time. The latter is used in this work and is set to 0.002. After a neuron spike, the absolute refractory time (equal to 1 in this work) is simulated by disabling it to increase the PSP until a certain unit time has passed. Figure 9 shows an example of LIFM neuron simulation with a refractory time equal to 3 seconds, potential leak rate equal to 0.02, a threshold of firing that is equal to 0.5 and synapses weights of 0.1, 0.1 and 0.35. It can be noted in figure 9 that the accumulation of spikes in time lead to an increase of PSP until a spike is generated and the effect of disregarding input spikes immediately after a spike is generated.

We set the initial connections (synapses strength) between neurons in SNNr using small-world connectivity [102, 111]. The connection probability was set such that neurons were more likely to be connected to neighboring neurons than to the distant ones. It has been shown that such an approach brings some ad-

vantages with regard to learning speed, parallel processing, and also favours the linking of specialized processing cluster units [112]. Additionally, we defined a radius r to be the maximum distance of connections of one neuron to another in the reservoir ($r = 25$ in this study). The initial weights were assigned as the product of random values $[-1, +1]$ divided by Euclidean distance between pre-synaptic and post-synaptic neuron so that 80% of them were positive values (excitatory connection) while 20% of them were negative values (inhibitory connections). Neuron connections are unidirectional, and the direction of communication was selected randomly. Connections between input neurons and other neuron are always positive and with doubled weight in comparison with other random connections. These connections were modified in the unsupervised learning stage to adapt to spatio-temporal patterns in input data.

6.6.3 Deep, unsupervised SNN training

We adjusted the connections between the neurons using the training data and a learning rule-based on Hebbian plasticity called spike-time-dependent plasticity (STDP) [113]. STDP learning modifies the neuronal connection weights taking into account the time difference between post- and pre-synaptic firing. A connection is strengthened, if postsynaptic firing occurs after presynaptic firing; otherwise, it is decreased. After STDP learning, the spatio-temporal pattern was saved in the value of connection weights in the SNNr. STDP learning rule is given as,

$$\Delta w = \text{sgn}(\Delta t) \frac{LR}{|\Delta t| + 1} \tag{1}$$

where LR is the STDP Learning Rate (0.001 in this work), $\text{sgn}(\cdot)$ is the function sign (-1 for negative values and 1 for positive), Δt is the difference between post- and pre-synaptic times ($\Delta t = t_{post} - t_{pre}$) and Δw is the change in the connection weight. The Hebbian relation Δw vs Δt is depicted in Figure 10. The learning results in the creation of deep structures of connections between neurons in the SNNr.

6.6.4 Supervised output neurons training

The deSNN is applied for supervised learning [84]. For every single training sample, an output neuron was created and connected to all the neurons in the trained SNNr (see Figure 6). Each output neuron was trained using the corresponding training sample by propagating the signal through the network once more. The neuron's connections weights $w_{i,j}$ between neurons i (in the reservoir) and j (output neuron) were initially established using rank order (RO) rule [84]. The RO method ranks the order in which the first spike arrives in the j neuron and the weights are given as,

$$w_{i,j}(0) = \alpha \mod^{order(i,j)} \tag{2}$$

where α is a learning parameter (in a partial case, equal to 1), $\mod$ is a modulation factor that defines how important the order of the spike is (0.8

in this study), $order(i,j)$ represents the order (the rank) of the first spike at synapse (i,j) ranked among all spikes arriving from all synapses to the neuron j. Furthermore, $order(i,j) = 0$ for the first spike to neuron j and increases according to the input spike order at other synapses.

Once a synaptic weight $w_{i,j}$ is initialized, based on the order of the first spike from i to j, the synapse becomes dynamic. It increases its value with a small positive value ($drift = 0.005$) at any time t a new spike arrives at this synapse and decreases its value if there is no spike at this time, as described in the following formula,

$$w_{i,j}(t) = \begin{cases} w_{i,j}(t-1) + drift, & if \quad S_{i,j}(t) = 1 \\ w_{i,j}(t-1) - drift, & if \quad S_{i,j}(t) = 0 \end{cases}$$

where $S_{i,j}(t)$ describes the existence of spike from neuron i entering to neuron j at time t. Every generated output neuron was trained to recognise and classify spatio-temporal patterns of weights adjusted by a corresponding labelled input training sample.

6.6.5 Classification

At classification stage, the NeuCube is fed with validation data. For each sample data synaptic weights for output neurons are calculated using the same supervised rules used in supervised training procedure. The connection weights that are learned in this process are then classified using a K-nearest neighbor (KNN, with $K = 3$ neighbors) algorithm and the labels that are known for all the samples.

We ran the whole NeuCube framework in a leave one subject out mode (LOSO) to test its capacity for learn spatio–temporal features from subjects and classify an unseen new subject.

6.6.6 Fusion of multimodal signals

Two schemas for the fusion of multimodal signals were explored - 1) features-level and 2) decision-level fusion. For features-level fusion, we coded all features (facial and peripheral) and included as input in NeuCube. Regarding decision-level fusion, for each subject, we calculated the accuracy of NeuCube classification in training data (rest of subjects) for separated modalities (facial and peripheral), and we chose the method with higher accuracy as the method for doing validation classification for the specific subject.

6.6.7 NeuCube parameters

NeuCube performance in analysing spatio-temporal data depends on several parameters. We chose a set of default parameter values equal to that used in the NeuCube development system publicly available online here, with the exception in refractory time. We used 1 time unit for this parameter in order to

Small world radius (r)	25
STDP learning rate (LR)	0.001
Threshold of firing	0.5
Potential leak rate	0.002
Refractory time	1 second
mod	0.84
drift	0.005
K	3

Table 1: NeuCube parameters

increase neuron activity. The NeuCube parameters used in this work are given in Table 1.

7 Results

NeuCube framework was fed with coded data under a LOSO cross validation scheme, i.e. all the data from a specific subject were excluded from the training set. All the parameters were fixed with values mentioned in Method section. Table 2 shows classification accuracy results in Mahnob-HCI dataset.

In total 390 videos (207 : 53.07% low valence, 183 : 46.92% high valence) were analyzed. Paired sample t-test comparing mean accuracy from Peripheral and Facial features result in no difference between them ($p < 0.05$). Mean classification accuracy using decision-level features results minor than using facial and greater than peripheral ($p < 0.03$). And feature-level fusion accuracy of 73.15% results better than decision-level fusion accuracy of 65.11% ($p < 0.004$).

For doing decision-level fusion of we obtained a mean accuracy of 83.7% for classifying the training data using facial features and 80.94% using peripheral training data.

7.1 Clustering Spike Communication

NeuCube framework has an option to analyze clusters of neuron-surrounding input neurons using the spike amount communicated between a pair of neurons. Figure 11 shows an example using this tool when the neuron reservoir is trained separately with one class (low valence) and the other one (high valence). For visualization purposes and taking account that mouth length and pupil size have more discriminative power regarding the rest of features, only input neurons coding higher and lower values in mouth length and pupil size are shown. Fig. 11 shows that neurons coding high values of mouth length and pupil size are more active for high valence and for this reason the cluster of spiking communication surrounding these neurons are bigger. Note that neuron coding low values of pupil size is more active for low valence. This results agree with features distribution in Figure 4.

Subject ID	Facial features accuracy (%)	Physiological features accuracy (%)	Fusion detection accuracy (%)	Fusion features (%)
1	73.33	66.67	66.67	73.33
2	62.5	50	62.5	56.25
3	75	72.73	75	81.82
4	78.57	66.67	66.67	83.33
5	75	50	75	68.75
6	58.82	70.59	58.82	70.59
7	75	60	75	93.33
8	64.29	50	64.29	64.29
9	60	100	60	90
10	61.54	69.23	61.54	69.23
11	78.57	61.54	61.54	76.92
13	64.29	71.43	64.29	85.71
14	50	57.14	50	71.43
16	72.73	63.64	63.64	63.64
17	62.5	80	62.5	40
18	41.67	50	41.67	62.5
19	61.54	75	61.54	58.33
20	53.33	73.33	53.33	80
21	66.67	71.43	66.67	71.43
22	66.67	66.67	66.67	80
23	75	50	75	75
24	69.23	53.85	69.23	76.92
25	66.67	77.78	66.67	55.56
27	68.75	60	68.75	80
28	66.67	66.67	66.67	86.67
29	68.75	50	68.75	64.29
30	78.57	57.14	78.57	78.57
Total	66.67	63.84	65.11	73.15

Table 2: Video valence classification accuracy in Mahnob-HCI dataset using NeuCube on individual and fusion features in a LOSO mode of cross-valiadation. Using fusion of features results in a better accuracy of classification at average.

8 Discussion

In this work we developed an approach based on NeuCube [96], which is an eSNN framework, to classify emotional valence using multimodal dataset that included video and physiological signals. We used a population coding scheme, based on ROC to encode input data into spikes, that SNNs can handle. When tested on the benchmark dataset, the MAHNOB-HCI, our approach resulted in a accuracy about 73.15 % for emotion classification. To the best of our knowledge, there has not been any other study to utilize SNN for affect recognition with multimodal data. In addition to the good accuracy of classification, the SNN system can be incrementally trained on new data and new features in an adaptive way, allowing the system to be used in an on-line applications [94].

8.1 Related work

Owing to its difficulty for the classification of spontaneous emotional responses from subjects, the MAHNOB-HCI dataset has been used in several studies. Since the MAHNOB-HCI dataset also contains multimodal data in the form of physiological and audio signals, several studies have resorted to a multimodal approach.

In a study by Koelstra and Patras [114] EEG and facial expressions were fused to perform affect recognition and implicit tagging . In case of EEG, power spectral density (PSD) features were used and for facial expression, an AU detection method was used, which was originally proposed in [115]. Basically, the AU detection was perormed using Free-form Deformations (FFDs) and Motion History Images to. For facial recognition, they trained the system using the MMI dataset [116] and obtained 64.5% of binary valence classification using only facial features and 74% by combining facial and EEG features. They performed a per-subject leave-one-trial-out cross-validation, where the classifier is trained on 19 trials from the same subject and tested on the 20th. As can be seen from their study, only using facial features result in low accuracies and fusion with EEG signal improved the classification accuracy. Boxuan and colleagues developed a temporal information preserving framework by splitting signals into multiple stages in each video. They achieved a valence (unpleasant, neutral, pleasant) classification accuracy of 54% using only facial expression and 69% when fusing with physiological signals [117]. They used Affdex SDK software [118], trained in 10,000 manually labelled facial images, which classify emotion-based on HOG features and support vector machine (SVM) classifier. Huang and colleagues obtained 50.57% for valence classification using appearance descriptors based facial features (Local binary pattern from three orthogonal planes, LBP-TOP) and 66.28% using fusion it with global EEG features [119]. They used the LOSO cross-validation scheme in nine emotion categories. A convolution deep belief network (CDBN) was proposed in [120] to learn emotional features from multimodal datasets and the authors reported a classification accuracy of 58.5%with the MAHNOB-HCI dataset. Torres et al. [121] performed feature selec-tion using discriminant-based algorithms, using EEG and peripheral signals. Their resuts showed that EEG-related features show the highest discrimina-tion ability. Furthermore, it was shown that EEG features along with GSR

achieved the highest discrimination for arousal index, whereas for the valence index, EEG features are accompanied by the heart rate features in achieving the highest discrimination power. For the MAHNOB-HCI dataset, they obtain a classification accuracy of 66.09% and 69.59% in the valence and arousal dimension respectively. Liu et al. [122] tested a deep learning approach based on multi-layer Long short-term memory recurrent neural network (LSTM-RNN) for emotion recognition, which combined temporal attention and band attention. They achieved an accuracy of 74.5% in valence classification (9 class) fusing video and EEG analysis. They used 20 participants for training, 4 participants for validation and 3 participants for testing. Huang et al. [123] used transfer learning technique (pre-trained convolutional neural network, CNN) to obtain an 73.33% in binary valence accuracy in MAHNOB-HCI dataset using facial features and 75.21 fusing with EEG features. Overall, the results we have obtained from the MAHNOB-HCI dataset are comparable with the state of the art work learning methods applied on this database. We observe that in some cases , the classificationa ccuracy obtaned using our SNN approach is better than the ANN approach, that have also used EEG sinals, which we have excluded. It is also to be noted that it is difficult to establish a fair comparison with most of the previous works as we did not include EEG features and use pretrained models as in [123]. We also disregarded all the data related with the subject in a Leave-subject out validation scheme.

8.2 Limitations

Our work has several limitations. First, we did not include any EEG features, because changes in EEG features associated with emotion are lumped features and we wanted to test NeuCube with temporal spatial patterns. As discussed previously, several studies have shown that including EEG features increases considerably the classification accuracy. However, there are several challenges in using EEG for emotion recognition [124], including selection of robust features, continuous decoding of affective states, reliable decoding of long-term reliability of EEG recordings for such studies, long preparation time and most importantly adopting a proper model of emotion with regard to EEG and understanding the EEG representation of affective states. For an excellent overview of these challenges the reader is directed to [124]. Nonetheless, the possibility of using EEG with the NeuCube framework will be explored in our future studies. Second, other important features that could be utilized from the mulltimodal data could be speech and postures. Several studies have considered the implications of including speech in affect recognition , with pitch considered to be an index into arousal [1], although the classification accuracy is shown to be lower than facial expression. Nonetheless, given the noninvasive and easy procedure to acquire voice, this feature should definitely be considered in the future studies with SNN. With regard to posture tracking, again it is a non-intrusive acquisition to the user's experience, but the equipment, but requires more expensive equip-ment compared to speech. Also there are some constraints with regard to the user's position, for example the user should be sitting [1].

We have also assumed that the face captured during the first two seconds after the stimulus is presented is neutral and consider it as the baseline. This could be problematic, especially if the participant is tired. Since we used Otsu's algorithm [125] for event detection, we do not take into account the long-lasting facial expression. It could be interesting if long term facial variation inside the video could be considered as detected events. Also, it could be interesting to incorporate detecting facial micro-expressions in our framework but this is in general challenging due to limited availability of such data and as well as difficulties in analyzing minute changes in expression [126]. Few methods have been proposed to address the problem of detecting micro-expressions using spatio-temporal local texture descriptor[127], Gabor filter with SVM classifier [128] and LBP-TOP with nearest neighbor classifier [129], which can be incorporated to add more information for the SNN framework. Another improvement could be to normalise expression between subjects by using pose estimation [130], correction of a 3D model [131, 132]. Further improvements could be made along the lines of detecting non-frontal head poses, identity bias and as well as illumination variation.

Although we studied the effect of varying certain NeuCube parameters, the performance of the proposed system may be affected by the choice of several other parameters. For instance, effect of varying other NeuCube parameters such as small world radius, firing threshold, refractory time, time resolution should be carefully investigated. The NeuCube framework also provides parameter optimization tool, which could be utilized instead of setting the parameters in an ad hoc manner.

9 Conclusion

Utilizing multimodal data to solve the problem of affect recognition with state of the art deep learning methods has gained a lot of popularity. SNNs offer an alternative to ANNs, where in the former is biologically more realistic model of neurons. In this work, we proposed a novel SNN method and system based on the known NeuCube framework, which is an eSNN, to sove affect recognition problem using multimodal data obtained from MAHANOB-HCI dataset. The eSNN is based on the ECOS principles which includes, efficient processing of spatio-temporal data and open evolving structure. Despite not including EEG, our approach gave results comparable to deep learning methods that utilize multimodal data, including EEG. In addition to the good accuracy of classification, the SNN system can be incrementally trained on new data and new features in an adaptive way, allowing the system to be used in an on-line applications.

References

[1] Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37, 2010.

[2] Jane Edwards, Henry J Jackson, and Philippa E Pattison. Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review. *Clinical psychology review*, 22(6):789–832, 2002.

[3] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003.

[4] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[5] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Face and Gesture 2011*, pages 827–834. IEEE, 2011.

[6] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[7] Antonios Danelakis, Theoharis Theoharis, and Ioannis Pratikakis. A survey on facial expression recognition in 3d video sequences. *Multimedia Tools and Applications*, 74(15):5577–5615, 2015.

[8] Soujanya Poria, Erik Cambria, Amir Hussain, and Guang-Bin Huang. Towards an intelligent framework for multimodal affective data analysis. *Neural Networks*, 63:104–116, 2015.

[9] Mohammed Yeasin, Baptiste Bullot, and Rajeev Sharma. Recognition of facial expressions and measurement of levels of interest from video. *IEEE Transactions on Multimedia*, 8(3):500–508, 2006.

[10] Yichuan Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

[11] Amogh Gudi, H Emrah Tasli, Tim M Den Uyl, and Andreas Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–5. IEEE, 2015.

[12] Radu Tudor Ionescu, Marius Popescu, and Cristian Grozea. Local learning to improve bag of visual words model for facial expression recognition. In *Workshop on challenges in representation learning, ICML*, 2013.

[13] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM, 2013.

[14] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450. ACM, 2016.

[15] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*, 2018.

[16] Kshitij Dhoble, Nuttapod Nuntalid, Giacomo Indiveri, and Nikola Kasabov. Online spatio-temporal pattern recognition with evolving spiking neural networks utilising address event representation, rank order, and temporal spike learning. In *The 2012 international joint conference on Neural networks (IJCNN)*, pages 1–7. IEEE, 2012.

[17] Albert Mehrabian. Communication without words. *Psychology today*, 2(4), 1968.

[18] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[19] Paul Ekman. Are there basic emotions? 1992.

[20] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[21] Rana El Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*, pages 181–200. Springer, 2005.

[22] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1805–1812, 2014.

[23] Md Zia Uddin, Mohammad Mehedi Hassan, Ahmad Almogren, Atif Alamri, Majed Alrubaian, and Giancarlo Fortino. Facial expression recognition utilizing local direction-based robust features and deep belief network. *IEEE Access*, 5:4525–4536, 2017.

[24] Ran Breuer and Ron Kimmel. A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv:1705.01842*, 2017.

[25] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991, 2015.

[26] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.

[27] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449, 2015.

[28] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*, pages 808–822. Springer, 2012.

[29] Nianyin Zeng, Hong Zhang, Baoye Song, Weibo Liu, Yurong Li, and Abdullah M Dobaie. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, 273:643–649, 2018.

[30] Wenyun Sun, Haitao Zhao, and Zhong Jin. An efficient unconstrained facial expression recognition algorithm based on stack binarized autoencoders and binarized neural networks. *Neurocomputing*, 267:385–395, 2017.

[31] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[32] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

[33] Mohammad T Shami and Mohamed S Kamel. Segment-based approach to the recognition of emotions in speech. In *2005 IEEE International Conference on Multimedia and Expo*, pages 4–pp. IEEE, 2005.

[34] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm. In *2005 IEEE International Conference on Multimedia and Expo*, pages 1500–1503. IEEE, 2005.

[35] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.

[36] Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition based on phoneme classes. In *Eighth International Conference on Spoken Language Processing*, 2004.

[37] Enrique M Albornoz, Diego H Milone, and Hugo L Rufiner. Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language*, 25(3):556–570, 2011.

[38] Zheng-wei Huang, Wen-tao Xue, and Qi-rong Mao. Speech emotion recognition with unsupervised feature learning. *Frontiers of Information Technology & Electronic Engineering*, 16(5):358–366, 2015.

[39] Neri E Cibau, Enrique M Albornoz, and Hugo L Rufiner. Speech emotion recognition using a deep autoencoder. *Anales de la XV Reunion de Procesamiento de la Informacion y Control*, 16:934–939, 2013.

[40] Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction*, pages 511–516. IEEE, 2013.

[41] Chenchen Huang, Wei Gong, Wenlong Fu, and Dongyu Feng. A research of speech emotion recognition based on deep belief network and svm. *Mathematical Problems in Engineering*, 2014, 2014.

[42] Guihua Wen, Huihui Li, Jubing Huang, Danyang Li, and Eryang Xun. Random deep belief networks for recognizing emotions from speech signals. *Computational intelligence and neuroscience*, 2017, 2017.

[43] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE, 2016.

[44] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804, 2014.

[45] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)*, pages 1–5. IEEE, 2017.

[46] Richard D Walk and Carolyn P Homan. Emotion and dance in dynamic light displays. *Bulletin of the Psychonomic Society*, 22(5):437–440, 1984.

[47] Marco De Meijer. The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal behavior*, 13(4):247–268, 1989.

[48] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[49] Mark Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior*, 28(2):117–139, 2004.

[50] Paul Ekman and Wallace V Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969.

[51] Sriparna Saha, Shreyasi Datta, Amit Konar, and Ramadoss Janarthanan. A study on emotion recognition from body gestures using kinect sensor. In *2014 International Conference on Communication and Signal Processing*, pages 056–060. IEEE, 2014.

[52] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1667–1675, 2017.

[53] O Barnea and V Shusterman. Analysis of skin-temperature variability compared to variability of blood pressure and heart rate. In *Proceedings of 17th International Conference of the Engineering in Medicine and Biology Society*, volume 2, pages 1027–1028. IEEE, 1995.

[54] Arturo Nakasone, Helmut Prendinger, and Mitsuru Ishizuka. Emotion recognition from electromyography and skin conductance. In *Proc. of the 5th international workshop on biosignal interpretation*, pages 219–222. Citeseer, 2005.

[55] Jennifer A Healey and Rosalind W Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2):156–166, 2005.

[56] Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Nils Fallentin, Ulf Lundberg, and Karen Søgaard. The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology*, 92(1-2):84–89, 2004.

[57] Jocelyn Scheirer, Raul Fernandez, and Rosalind W Picard. Expression glasses: a wearable device for facial expression recognition. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems*, pages 262–263, 1999.

[58] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*, volume 11. Elsevier, 2013.

[59] Jennifer A Healey. Affect detection in the real world: Recording and processing physiological signals. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6. IEEE, 2009.

[60] Ikuo Homma and Yuri Masaoka. Breathing rhythms and emotions. *Experimental physiology*, 93(9):1011–1021, 2008.

[61] Ivan Nykliček, Julian F Thayer, and Lorenz JP Van Doornen. Cardiorespiratory differentiation of musically-induced emotions. *Journal of Psychophysiology*, 1997.

[62] Paul Grossman and Cees J Wientjes. How breathing adjusts to mental and physical demands. In *Respiration and emotion*, pages 43–54. Springer, 2001.

[63] Wei-Long Zheng, Jia-Yi Zhu, Yong Peng, and Bao-Liang Lu. Eeg-based emotion classification using deep belief networks. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014.

[64] Guillaume Chanel, Julien Kronegg, Didier Grandjean, and Thierry Pun. Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals. In *International workshop on multimedia content representation, classification and security*, pages 530–537. Springer, 2006.

[65] Robert Horlings, Dragos Datcu, and Leon JM Rothkrantz. Emotion recognition using brain activity. In *Proceedings of the 9th international conference on computer systems and technologies and workshop for PhD students in computing*, pages II–1, 2008.

[66] Eric Ed Granholm and Stuart R Steinhauer. Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*, 2004.

[67] Timo Partala, Maria Jokiniemi, and Veikko Surakka. Pupillary responses to emotionally provocative stimuli. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 123–129, 2000.

[68] Xiaowei Jia, Kang Li, Xiaoyi Li, and Aidong Zhang. A novel semi-supervised deep learning framework for affective state recognition on eeg signals. In *2014 IEEE international conference on bioinformatics and bioengineering*, pages 30–37. IEEE, 2014.

[69] Tzyy-Ping Jung, Terrence J Sejnowski, et al. Multi-modal approach for affective computing. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 291–294. IEEE, 2018.

[70] Youngjun Cho, Nadia Bianchi-Berthouze, and Simon J Julier. Deepbreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 456–463. IEEE, 2017.

[71] Qiang Zhang, Xianxiang Chen, Qingyuan Zhan, Ting Yang, and Shanhong Xia. Respiration-based emotion recognition with deep learning. *Computers in Industry*, 92:84–90, 2017.

[72] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47(3):1–36, 2015.

[73] Maja Pantic and Leon JM Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[74] Chandrima Sarkar, Sumit Bhatia, Arvind Agarwal, and Juan Li. Feature analysis for computational personality recognition using youtube personality data set. In *Proceedings of the 2014 ACM multi media on workshop on computational personality recognition*, pages 11–14, 2014.

[75] Shangfei Wang, Yachen Zhu, Guobing Wu, and Qiang Ji. Hybrid video emotional tagging using users' eeg and video content. *Multimedia tools and applications*, 72(2):1257–1283, 2014.

[76] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[77] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[78] Firoj Alam and Giuseppe Riccardi. Predicting personality traits using multimodal information. In *Proceedings of the 2014 ACM multi media on workshop on computational personality recognition*, pages 15–18, 2014.

[79] Guoyong Cai and Binbin Xia. Convolutional neural networks for multimedia sentiment analysis. In *Natural Language Processing and Chinese Computing*, pages 159–167. Springer, 2015.

[80] Toshihiko Yamasaki, Yusuke Fukushima, Ryosuke Furuta, Litian Sun, Kiyoharu Aizawa, and Danushka Bollegala. Prediction of user ratings of oral presentations using label relations. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pages 33–38, 2015.

[81] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[82] Wei Wang, Giacomo Pedretti, Valerio Milo, Roberto Carboni, Alessandro Calderoni, Nirmal Ramaswamy, Alessandro S Spinelli, and Daniele Ielmini. Learning of spatiotemporal patterns in a spiking neural network with resistive switching synapses. *Science advances*, 4(9):eaat4752, 2018.

[83] Aboozar Taherkhani, Ammar Belatreche, Yuhua Li, Georgina Cosma, Liam P Maguire, and TM McGinnity. A review of learning in biologically plausible spiking neural networks. *Neural Networks*, 2019.

[84] Nikola Kasabov, Kshitij Dhoble, Nuttapod Nuntalid, and Giacomo Indiveri. Dynamic evolving spiking neural networks for on-line spatio-and spectro-temporal pattern recognition. *Neural Networks*, 41:188–201, 2013.

[85] Wolfgang Maass and Henry Markram. On the computational power of circuits of spiking neurons. *Journal of computer and system sciences*, 69(4):593–616, 2004.

[86] Wolfgang Maass. Fast sigmoidal networks via spiking neurons. *Neural Computation*, 9(2):279–304, 1997.

[87] Sander M Bohte, Joost N Kok, and Han La Poutre. Error-backpropagation in temporally encoded networks of spiking neurons. *Neurocomputing*, 48(1-4):17–37, 2002.

[88] Sander M Bohte, Han La Poutré, and Joost N Kok. Unsupervised clustering with spiking neurons by sparse temporal coding and multilayer rbf networks. *IEEE Transactions on neural networks*, 13(2):426–435, 2002.

[89] Boudjelal Meftah, Olivier Lezoray, and Abdelkader Benyettou. Segmentation and edge detection based on spiking neural network model. *Neural Processing Letters*, 32(2):131–146, 2010.

[90] Samanwoy Ghosh-Dastidar and Hojjat Adeli. Improved spiking neural networks for eeg classification and epilepsy and seizure detection. *Integrated Computer-Aided Engineering*, 14(3):187–212, 2007.

[91] Simon Thorpe and Jacques Gautrais. Rank order coding. In *Computational neuroscience*, pages 113–118. Springer, 1998.

[92] Nikola K Kasabov. *Evolving connectionist systems: the knowledge engineering approach*. Springer Science & Business Media, 2007.

[93] Simei Gomes Wysoski, Lubica Benuskova, and Nikola Kasabov. Evolving spiking neural networks for audiovisual information processing. *Neural Networks*, 23(7):819–835, 2010.

[94] Nikola K Kasabov. *Time-space, spiking neural networks and brain-inspired artificial intelligence*, volume 7. Springer, 2018.

[95] Nikola Kasabov. Neucube evospike architecture for spatio-temporal modelling and pattern recognition of brain signals. In *Iapr workshop on artificial neural networks in pattern recognition*, pages 225–243. Springer, 2012.

[96] Nikola K Kasabov. Neucube: A spiking neural network architecture for mapping, learning and understanding of spatio-temporal brain data. *Neural Networks*, 52:62–76, 2014.

[97] Nikola Kasabov, Jin Hu, Yixiong Chen, Nathan Scott, and Yulia Turkova. Spatio-temporal eeg data classification in the neucube 3d snn environment: methodology and examples. In *International Conference on Neural Information Processing*, pages 63–69. Springer, 2013.

[98] Nikola Kasabov and Elisa Capecci. Spiking neural network methodology for modelling, classification and understanding of eeg spatio-temporal data measuring cognitive processes. *Information Sciences*, 294:565–575, 2015.

[99] Nikola Kasabov, Nathan Matthew Scott, Enmei Tu, Stefan Marks, Neelava Sengupta, Elisa Capecci, Muhaini Othman, Maryam Gholami Doborjeh, Norhanifah Murli, Reggio Hartono, et al. Evolving spatio-temporal data machines based on the neucube neuromorphic framework: design methodology and selected applications. *Neural Networks*, 78:1–14, 2016.

[100] Henk A Mastebroek, Johan E Vos, and JE Vos. *Plausible neural networks for biological modelling*, volume 13. Springer Science & Business Media, 2001.

[101] Shih-Chii Liu and Tobi Delbruck. Neuromorphic sensory systems. *Current opinion in neurobiology*, 20(3):288–295, 2010.

[102] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186, 2009.

[103] Cornelius J Stam. Functional connectivity patterns of human magnetoencephalographic recordings: a 'small-world'network? *Neuroscience letters*, 355(1-2):25–28, 2004.

[104] Zhang J Chen, Yong He, Pedro Rosa-Neto, Jurgen Germann, and Alan C Evans. Revealing modular architecture of human brain structural networks by using cortical thickness from mri. *Cerebral cortex*, 18(10):2374–2381, 2008.

[105] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2011.

[106] Paul Viola, Michael Jones, et al. Rapid object detection using a boosted cascade of simple features. *CVPR (1)*, 1(511-518):3, 2001.

[107] Jianbo Shi et al. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, pages 593–600. IEEE, 1994.

[108] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[109] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[110] Jiapu Pan and Willis J Tompkins. A real-time qrs detection algorithm. *IEEE transactions on biomedical engineering*, (3):230–236, 1985.

[111] Valentino Braitenberg and Almut Schüz. *Cortex: statistics and geometry of neuronal connectivity*. Springer Science & Business Media, 2013.

[112] D Simard, L Nadeau, and H Kröger. Fastest learning in small-world neural networks. *Physics Letters A*, 336(1):8–15, 2005.

[113] Sen Song, Kenneth D Miller, and Larry F Abbott. Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature neuroscience*, 3(9):919, 2000.

[114] Sander Koelstra and Ioannis Patras. Fusion of facial expressions and eeg for implicit affective tagging. *Image and Vision Computing*, 31(2):164–174, 2013.

[115] Sander Koelstra, Maja Pantic, and Ioannis Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1940–1954, 2010.

[116] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65. Paris, France, 2010.

[117] Boxuan Zhong, Zikun Qin, Shuo Yang, Junyu Chen, Nicholas Mudrick, Michelle Taub, Roger Azevedo, and Edgar Lobaton. Emotion recognition with facial expressions and physiological signals. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE, 2017.

[118] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 3723–3726. ACM, 2016.

[119] Yongrui Huang, Jianhao Yang, Siyu Liu, and Jiahui Pan. Combining facial expressions and electroencephalography to enhance emotion recognition. *Future Internet*, 11(5):105, 2019.

[120] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[121] Cristian Torres-Valencia, Mauricio Álvarez-López, and Álvaro Orozco-Gutiérrez. Svm-based feature selection methods for emotion recognition from multimodal data. *Journal on Multimodal User Interfaces*, 11(1):9–23, 2017.

[122] Jiamin Liu, Yuanqi Su, and Yuehu Liu. Multi-modal emotion recognition with temporal-band attention based on lstm-rnn. In *Pacific Rim Conference on Multimedia*, pages 194–204. Springer, 2017.

[123] Xiaohua Huang, Jukka Kortelainen, Guoying Zhao, Xiaobai Li, Antti Moilanen, Tapio Seppänen, and Matti Pietikäinen. Multi-modal emotion analysis from facial expressions and electroencephalogram. *Computer Vision and Image Understanding*, 147:114–124, 2016.

[124] Xin Hu, Jingjing Chen, Fei Wang, and Dan Zhang. Ten challenges for eeg-based affective computing. *Brain Science Advances*, 5(1):1–20, 2019.

[125] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[126] Yandan Wang, John See, Raphael C-W Phan, and Yee-Hui Oh. Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. *PloS one*, 10(5):e0124674, 2015.

[127] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*, pages 1–6. IEEE, 2013.

[128] Qi Wu, Xunbing Shen, and Xiaolan Fu. The machine knows what you are hiding: an automatic micro-expression recognition system. In *international conference on affective computing and intelligent Interaction*, pages 152–162. Springer, 2011.

[129] Yanjun Guo, Yantao Tian, Xu Gao, and Xuange Zhang. Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method. In *2014 international joint conference on neural networks (IJCNN)*, pages 3473–3479. IEEE, 2014.

[130] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626, 2008.

[131] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

[132] Amin Jourabloo and Xiaoming Liu. Pose-invariant 3d face alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3694–3702, 2015.

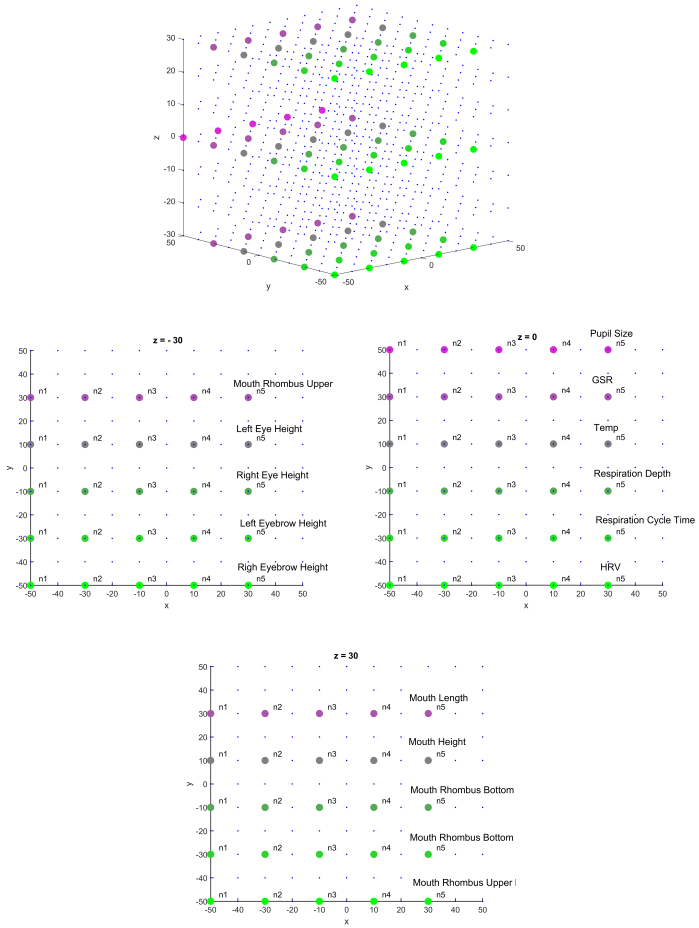


Figure 8: Input neurons location for facial and peripheral features classification. n1 means for the neuron coding the lowest values and n5 the highest ones for each feature. Note there are 3 layers of input neuron in the cube, located at $z = -30$ (facial), $z = 0$ (peripheral) and $z = 30$ (facial).

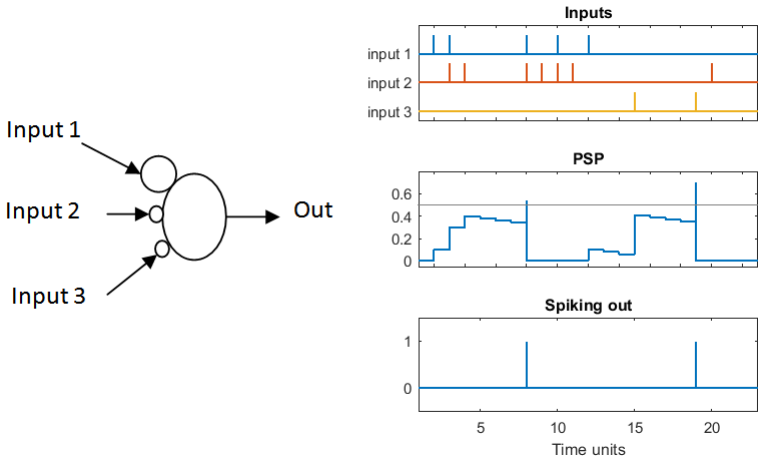


Figure 9: LIFM neuron model. Small circles at neuron inputs represent connection weights. Note that input 1 has a bigger weight and it produces a larger effect in PSP

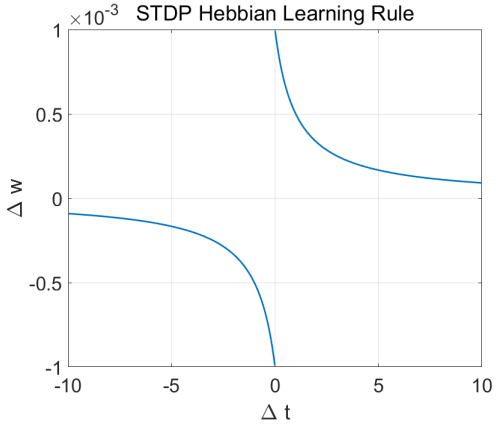


Figure 10: Hebbian Learning rule, connection (synaptic modification) vs difference between post- and pre-synaptic times

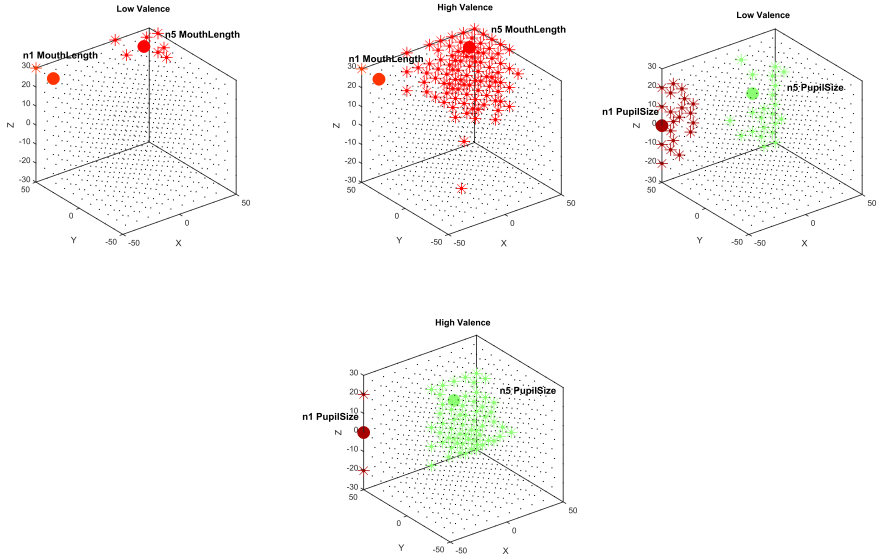


Figure 11: Neuron activity pattern example when NeuCube is trained using each separately data (low and high valence).