





DeepVANet: A Deep End-to-End Network for Multi-modal Emotion Recognition

Yuhao Zhang¹, Md Zakir Hossain^{1,2} , and Shafin Rahman³ 

¹ The Australian National University, Canberra, ACT 2601, Australia
{yuhao.zhang,zakir.hossain}@anu.edu.au

² CSIRO Agriculture and Food, Black Mountain, Canberra, ACT 2601, Australia

³ North South University, Dhaka, Bangladesh
shafin.rahman@northsouth.edu

Abstract. Human facial expressions and bio-signals (e.g., electroencephalogram and electrocardiogram) play a vital role in emotion recognition. Recent approaches employ both vision-based and bio-sensing data to design multi-modal recognition systems. However, these approaches require tremendous domain-specific knowledge, complex pre-processing steps and fail to take full advantage of the end-to-end nature of deep learning techniques. This paper proposes a deep end-to-end framework, DeepVANet, for multi-modal valence-arousal-based emotion recognition that applies deep learning methods to extract face appearance features and bio-sensing features. We use convolutional long short-term memory (ConvLSTM) techniques in face appearance feature extraction to capture spatial and temporal information from face image sequences. Unlike conventional time or frequency domain features (e.g., spectral power and average signal intensity), we use a 1D convolutional neural network (Conv1D) to learn bio-sensing features automatically. In experiments, we evaluate our method using DEAP and MAHNOB-HCI datasets. Our proposed multi-modal framework successfully outperforms both single- and multi-modal methods achieving superior performance compared to state-of-the-art approaches and reaches as high as 99.22% correctness.

Keywords: Emotion recognition · Deep learning · Physiological signals

1 Introduction

Emotion is an important mental state of human beings and dominates people's attitudes and behaviors. Recently, emotion recognition has become a popular research topic due to its wide applications in the medical, education, and gaming

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-85613-7_16) contains supplementary material, which is available to authorized users.

fields [5,35]. In particular, existing works predominantly investigate two modalities of data for the emotion recognition task. *Firstly*, vision-based data, i.e., face image/video, widely dominates emotion recognition research. Facial expressions can reflect human's emotions directly, and they are easy to capture by cameras. State-of-the-art studies have successfully explored both hand-crafted and deep learning methods to extract these kinds of facial expression features [16]. Hand-crafted methods include facial specific methods based on facial landmarks (angle and distance of landmark pairs) [8,37] and other effective general visual feature extraction methods such as Local Binary Pattern (LBP) [10] and Histogram of Gradients (HoG) [20,34]. Deep learning methods, particularly convolutional neural network (CNN) and its variations, have shown their powerful visual data processing ability in facial emotion recognition tasks [21]. Especially for facial image sequences or videos, 3D CNN [9], recurrent convolutional neural network (RCNN) [14,23] and ConvLSTM [11] perform well in spatial and temporal facial expression feature extraction. *Secondly*, physiological/bio-sensing signals, i.e., electroencephalogram (EEG), electrocardiogram (ECG), and galvanic skin response (GSR) data are used to extract discriminative information of human emotions [27,31]. Compared to facial expressions, physiological signals are more instinctive and uncontrolled. So, they can reflect the real emotions more accurately. Methods use hand-crafted features like power spectral density (PSD) from different frequency bands, conditional cross-entropy, and statistical features to extract discriminative information [18]. However, due to inconsistency across hardware, the necessity of expert pre-processing, and the expensive nature of bio-sensing data, there is not enough work on deep learning strategies.

Apart from single modality (either bio-sensing or vision) approaches, some multi-modal affective computing efforts have been proposed and shown advantages in emotion recognition accuracy [25,28,38]. Single modal features have their respective strength in some aspects. Thus, their fusion can take advantage of their strengths by generating comprehensive and salient multi-modal features. While dealing with multi-modal data, current literature does not utilize the full benefit of deep learning (e.g. complicated data pre-processing). This paper investigates an end-to-end approach to address the emotion recognition task using both modalities. Human emotional states can be represented qualitatively (discrete model) [6] or quantitatively (dimensional model) [24]. Discrete models generally classifies emotions into six basic types: happiness, sadness, anger, surprise, disgust, and fear. The dimensional model represents emotions in a multidimensional space, such as valence-arousal space, which is more flexible and represents a broader range of emotions. This study focuses on emotion recognition in valence-arousal space using multi-modal data (vision-based and bio-sensing).

Limitation of Prior Works: We identify several issues in existing works of the valence-arousal model of emotion classification: *(a)* Although some models used deep learning in vision pipeline, they failed to incorporate the same on bio-sensing data [12]. *(b)* Some models pre-processed raw bio-sensing signals to 2D-frames (e.g. EEG heatmap of PSD bands) to extract CNN features [13,28,36].

(c) Other models usually extracted individual modal features separately and then applied a decision-level fusion on them [12, 25]. All the mentioned issues hamper the end-to-end nature of deep learning.

This paper proposes a deep end-to-end network, DeepVANet, to recognize humans emotions using both face videos and physiological signals by extracting vision-based and bio-sensing features, respectively. At the vision pipeline, we use a Convolutional Recurrent Neural Network (CRNN) based on convolutional long short-term memory (ConvLSTM) as face appearance feature extractor to obtain spatio-temporal information from face videos [26]. Similarly, we use Conv1D neural network at the bio-sensing pipeline instead of traditional hand-crafted time/frequency-based extraction methods. Later, we employ a fusion pipeline to fuse both modality features enriching the feature quality for better classification. Our fusion pipeline supports both feature- and decision-level fusion strategies. Different modal features are concatenated together then pass through some fully-connected (FC) classifier layers to predict the valence/arousal label for feature-level fusion. In contrast, for decision-level fusion, FC sub-classifiers accept single modal features and output predict scores separately. Then, adaptive boosting techniques are employed to yield the final emotion label. Our method combines all steps while maintaining the end-to-end nature of the work. We present experimental results on DEAP [17] and MAHNOB-HCI [29] datasets demonstrating new state-of-the-art results on multi-modal emotion recognition. In summary, the main contributions of our work are: (1) To the best of our knowledge, we propose the first end-to-end deep network for multi-modal (face and bio-sensing) valence-arousal-based emotion recognition. (2) Instead of the traditional use of hand-crafted features, we fuse Conv1D based bio-sensing and CRNN/ConvLSTM based vision feature to learn a rich feature representation jointly for emotion classification. (3) We perform extensive experiments on two well-known datasets achieving state-of-the-art performance: 98.56/98.18% and 99.19/99.22% (valence/arousal) on DEAP and MAHNOB-HCI, respectively.

2 Multi-modal Emotion Recognition

Problem Formulation: Suppose, for i th facial emotion video frames, $\mathbf{F}_i = \langle \mathbf{I}_t | 1 \dots n_i \rangle$ where, \mathbf{I}_t represents t th image frame, we collect bio-sensing signals, $\mathbf{B}_i = \langle e_t | 1 \dots m_i \rangle$, where, e_t represent t th data point. n_i and m_i are the length of the video and physiological signal, respectively. For i th instance the ground-truth annotation is \mathbf{y}_i which can be valence or arousal value. We train an end-to-end parameterized model, \mathcal{F}_θ using a set of tuples $\{ \langle (\mathbf{F}_i, \mathbf{B}_i), \mathbf{y}_i \rangle : i \in [0, T] \}$, $\mathbf{y}_i \in [0, 1]$, where T is the total number of instances in the dataset, $\mathbf{y}_i = 0$ represents low valence/arousal, high valence/arousal otherwise. During inference, given a test video and physiological signal pair $(\mathbf{F}_j, \mathbf{B}_j)$ as input, \mathcal{F}_θ predicts $\hat{\mathbf{y}}_j$ which approximates ground-truth annotation \mathbf{y}_j as follows: $\hat{\mathbf{y}}_j = \mathcal{F}_\theta((\mathbf{F}_j, \mathbf{B}_j); \Theta)$.

2.1 Architecture

Our proposed DeepVANet for VA based emotion recognition has three blocks: Face Appearance, Bio-sensing Data, and Classification Pipelines (Fig. 1).

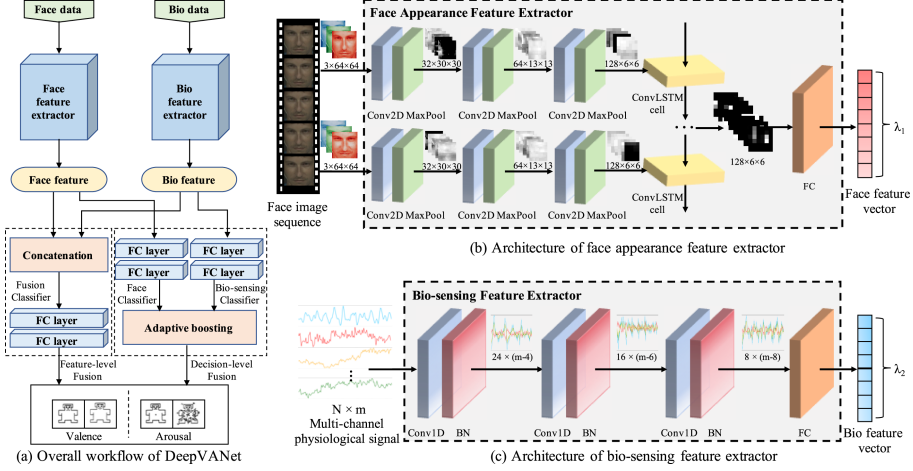


Fig. 1. Our proposed DeepVANet for Valence and Arousal based emotion recognition; FC and BN stand for fully connected layer and batch normalisation respectively.

Face Appearance Pipeline: We deploy 3-layer CNN (pretrained on AFEW-VA dataset [19]) at the top of our face appearance feature extractor to capture high-level spatial information. The inputs of the face appearance feature extractor are face image sequences (with size of 64×64). Each sample F_i has the shape of $n \times 3 \times 64 \times 64$, where n denotes the sequence length and 3 represents the number of RGB channels. The kernel number of three convolutional layers are 32, 64, and 128, respectively, where all the kernels have the same size of 3×3 . Each layer is followed by a ReLu activator and a 3×3 max pooling. So, through a series of 2D convolutional and max-pooling operations, the CNN layers output the feature maps F' with the shape of $n \times 128 \times 6 \times 6$. Facial expressions are dynamic and evolve during the video containing temporal patterns. We extract face expression information in both spatial and temporal dimensions using a ConvLSTM [26]. For each timestep t , input gate i_t , forget gate f_t , cell gate g_t and output gate o_t are defined as $i_t = \sigma(W_{xi} * F'_t + W_{hi} * h_{t-1} + b_i)$, $f_t = \sigma(W_{xf} * F'_t + W_{hf} * h_{t-1} + b_f)$, $g_t = \tanh(W_{xg} * F'_t + W_{hg} * h_{t-1} + b_g)$ and $o_t = \sigma(W_{xo} * F'_t + W_{ho} * h_{t-1} + b_o)$; cell state c_t and hidden state h_t are defined as $c_t = f_t \odot c_{t-1} + i_t \odot g_t$ and $h_t = o_t \odot \tanh(c_t)$, where \odot and $*$ demote the Hadamard product and 2D convolutional operator. A followed FC layer accepts the final hidden state and outputs the face appearance features which is a $1 \times \lambda_1$ vector $f_{face} = (a_1, a_2, \dots, a_{\lambda_1})$.

Bio-sensing Data Pipeline: Instead of traditional time or frequency domain methods, we use deep learning methods to extract bio-sensing features automatically. To extract high-level bio-sensing features, we apply a multi-layer CNN architecture consisting of three 1D convolutional layers and a fully connected layer [32]. Each convolutional layer is followed by a batch normalization layer and a ReLu activator, and the last fully connected layer is deployed for further feature extraction and flattening. The inputs to the bio-sensing data pipeline are multi-channel physiological signals (concatenate EEG and peripheral physiological signals over channel dimension). For each bio-sensing input instance \mathbf{B}_i with the shape of $c_{bio} \times m$, where c_{bio} and m represent the channel number and signal length respectively, the forward propagation in our bio feature extractor is shown as follows. Firstly, \mathbf{B}_i flows into three-layer 1D CNN. The kernel sizes of three 1D CNN layers are 5, 3, 3, respectively. We choose to decrease output channel numbers for CNN layers 24, 16, and 8, to reduce the bio-sensing feature size and computation complexity. So, the output of convolutional layers is in the shape of $8 \times (m - 8)$, where m denotes the signal length. Then it is put into a FC layer and flattened into a $1 \times \lambda_2$ bio feature vector $f_{bio} = (b_1, b_2, \dots, b_{\lambda_2})$.

Classification Pipeline: The classification pipeline predicts a score for valence or arousal. For each instance, we have ground-truth $\hat{\mathbf{y}} \in \{0, 1\}$ to represent the emotion label (valence or arousal), where 0 and 1 represent low and high, respectively. Here, we investigate two kinds of fusion methods (feature-level and decision-level fusion) to make full use of different modalities. For *feature-level fusion*, face appearance feature f_{face} and bio feature f_{bio} are concatenated to generate a multi-modal feature vector $f_{multi} = f_{face} \oplus f_{bio}$, which is fed into the fusion classifier containing a two-layer fully-connected neural network. Using different value for λ_1 and λ_2 , we can set different feature size combinations and explore the contribution of each single modality. For *decision-level fusion*, we use Adaptive Boosting (AdaBoost) [7, 12]. The face and bio-sensing features are used to train two sub-classifiers separately. The sub-classifier network is the same as the fusion classifier. Each sub-classifier yields a predicted score S_i , which are used together to calculate a fusion score S_{fusion} using the Adaboost to get the proper weights of each sub-classifier. Generally, the sub-classifier having the less predict error holds the higher weight in calculating the fusion score. Training samples are also given the weights which influence their contribution to the error. The sample weights are initialized uniformly. The weight decreases if the sample is predicted correctly and increases otherwise. After finishing all training iterations, we can get the sub-classifier weights and calculate the fusion score.

2.2 Training and Inference

We use binary cross-entropy as a loss function for both feature- and decision-level fusion cases. Considering \mathcal{N}, \mathbf{y} and S_i as the training batch size, target

emotion label and predicted score, respectively, the loss is calculated as:

$$Loss = -\frac{1}{N} \sum_{i=1}^N (\mathbf{y}_i \cdot \ln S_i + (1 - \mathbf{y}_i) \cdot \ln (1 - S_i))$$

During inference, we forward test video, \mathbf{F}_j and physiological signal, \mathbf{B}_j to our proposed network, obtain score S_{fusion} and use the following formula for final prediction, $\hat{\mathbf{y}}$.

$$\hat{\mathbf{y}} = \begin{cases} High, & S_{fusion} > 0.5 \\ Low, & S_{fusion} \leq 0.5 \end{cases}$$

3 Experiment

Dataset: Two public multi-modal emotion datasets, namely DEAP, MAHNOB-HCI are used to evaluate our network separately. A brief statistics about the dataset is shown in Table 1.

Table 1. Statistics of the datasets used in this work

Criterion	DEAP	MAHNOB-HCI
Dataset size	22 subjects \times 40 trials	27 subjects \times 20 trials
Stimulus duration	60 s	49.7–117 s
Modality	Vision, EEG, Peripheral	Vision, EEG, Peripheral
Used bio signals	EEG, EOG, EMG, GSR, Respiration belt, Plethy-smograph, Temperature	EEG, ECG, GSR, Respiration belt, Temperature
Labels	Continuous valence and arousal from 1 to 9	Discrete valence and arousal of integers from 1 to 9

Experimental Evaluation: In line with the previous works [12, 28], we divide valence and arousal into ‘High’/‘Low’ using value = 5 as the threshold. We notice that physiological signals vary significantly across subjects [2, 17, 36], and this inter-subject difference affects the emotion recognition accuracy while predicting valence and arousal across people. Therefore, similar to the past works [2, 36], we train and evaluate our model for every single subject (called per-subject experiment) to improve the recognition performance. In addition to the per-subject experiment, we also conduct inter-subject experiments. Here, all subjects’ data is used to train and test one model aiming to evaluate our proposed network’s generalization ability. For both per- and inter-subject experiments, we run 10-fold cross-validation and take the average testing accuracy as the model performance metric. Besides our multi-modal network, we also evaluate two single modalities,

face and bio-sensing, separately. Furthermore, EEG and peripheral physiological signals are used separately in bio-sensing modality experiments.

Validation Strategy: We have two hyper-parameters in our model: the face feature dimension λ_1 and the bio-sensing feature dimension λ_2 . For both hyper-parameter cases, we choose a set of candidates $\{16, 32, 64, 128, 256\}$, and perform a grid search to determine the best hyper-parameter. We use the mean recognition accuracy of valence and arousal as validation metric, and 16 and 64 are chosen as the face appearance feature and bio-sensing feature size.

Implementation Details¹: To overcome the limited data size for deep learning, we divide each trial into 1-second length segments as proposed in [36]. We extract frames from face videos 5 Hz for vision-based data and perform face detection [3], cropping, and alignment on each frame. These face images are resized to 64×64 resolution and normalized within the interval $[0, 1]$ before feeding into our model. For bio-sensing data, we perform average reference, 4–45 Hz bandpass filter, and artifact removal using EEGLab [4] on 32-channel EEG signals. Both EEG and peripheral signals are downsampled 128 Hz and concatenated together along with channel dimension. For each segment, we apply the baseline removal on bio-sensing signals by subtracting the mean of 3-second pre-trial data [36]. We use the Adam algorithm [15] as the optimizer during training. The batch size, numbers of training epochs, and learning rate for all face-, bio-, and multi-modal are 64, 50, and 10^{-3} , respectively. We implement our work with the *Pytorch* framework using a single Nvidia V100 GPU.

3.1 Quantitative Results

Overall Result: Table 2 reports the inter- and per-subject experiment results of proposed DeepVANet on DEAP and MAHNOB-HCI datasets. We choose [17, 29] as baseline methods and consider some studies applying deep learning for comparison. These studies used the same binary VA labels as this paper. Some notable observations are: (1) Multi-modal methods [12, 18, 28] achieved better accuracy than single model methods [17, 29] which justifies the significance of fusion strategy on face images and physiological signals. (2) Some deep learning-based approaches [1, 2, 36] get an accuracy above 90% by only using EEG signals. This is because the deep learning-based (LSTM) EEG features are more discriminative than handcrafted features. (3) We notice that the per-subject experiment result is better than the inter-subject experiment because the bio-sensing data variance across subjects impacts emotion recognition performance. However, our DeepVANet also performs well in inter-subject experiments and shows a good generalization ability to predict emotion for different people. (4) Overall, our feature-level fusion approach achieves the highest average recognition accuracy of 98.56%/98.18% and 99.19%/99.22% on DEAP and MAHNOB-HCI datasets, respectively, and surpasses the previous methods by a considerable margin. This

¹ Code and evaluation are available at: <https://github.com/geekdanielz/DeepVANet>.

Table 2. Overall results of VA based emotion recognition. Both inter-subject and per-subject experiment results are reported. For accuracy (%), valence at left and arousal at right.

Data	Study	Modalities	Accuracy
DEAP	Inter-subject experiments		
	Siddharth et al. [28]	Face, EEG, Peripheral	79.52/78.34
	Tang et al. [30]	EEG, Peripheral	83.82/83.23
	This work	Face, EEG, Peripheral	95.56/95.33
	Per-subject experiments		
	Keoltra et al. [17]	EEG, Peripheral	62.70/62.00
	Huang et al. [12]	Face, EEG	80.30/74.23
	Liu et al. [22]	EEG, Peripheral	85.20/80.50
	Alhagry et al. [1]	EEG	85.65/85.45
	Yang et al. [36]	EEG	90.80/91.03
MAHNOB	Inter-subject experiments		
	Soleymani et al. [29]	EEG, Peripheral	57.00/52.40
	Wiem and Lachiri [33]	EEG, Peripheral	68.75/64.23
	Siddharth et al. [28]	Face, EEG, Peripheral	85.49/82.93
	This work	Face, EEG, Peripheral	96.82/97.79
	Per-subject experiments		
	Keolstra et al. [18]	Face, EEG	74.00/70.00
	Huang et al. [12]	Face, EEG	75.21/75.63
	This work	Face, EEG, Peripheral	99.19/99.22

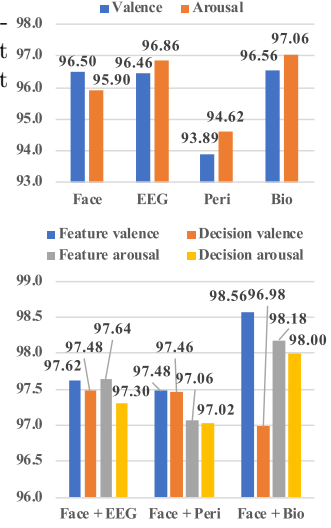


Fig. 2. Ablation study results (recognition accuracy %) for DEAP dataset

success is a result of utilizing advanced deep learning-based feature extraction for bio-sensing data.

Ablation Study: To explore the emotion recognition performance of different modalities, we test our approach using both single- and multi-modal input on DEAP dataset. From the results in Fig. 2, we find that: *(1)* EEG features are more salient than peripheral features, and bio-sensing modality (EEG+Peripheral) performs better than individual EEG and peripheral modality. *(2)* multi-modality outperforms all the single modalities, which demonstrates the effectiveness of our feature- and decision-level fusion strategies. *(3)* feature-level fusion method performs better than decision-level fusion in most cases. We believe it is because we only have two sub-classifiers in our approach, limiting the effectiveness of the decision-level fusion based on AdaBoost.

3.2 Qualitative Visualization

We use t-SNE to reduce the feature dimension into 2D space and visualize different modal features in Fig. 3. We find that the HIGH and the LOW class features successfully form two rough clusters in 2D space. Notably, for the fusion modal, two clusters are separated more clearly. It also demonstrates that our multi-modal features are salient and discriminative.

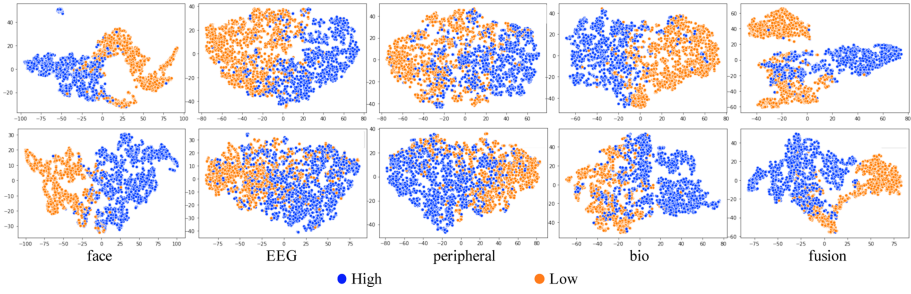


Fig. 3. 2D t-SNE visualization of features extracted from subject 1 of DEAP dataset; top and bottom rows represent valence and arousal, respectively.

4 Conclusion

In this paper, we propose a deep end-to-end network, DeepVANet, for multi-modal emotion recognition. The network accepts face image sequences and bio-sensing signals (e.g., EEG, EOG, ECG, GSR etc.) as input and yields valence-arousal labels for emotion recognition. We adopt ConvLSTM and Conv1D to extract face appearance and bio-sensing computing features, respectively, demonstrating a significant advantage in the valence-arousal emotion recognition model. Furthermore, we evaluate our method with per-subject and inter-subjects experiments. For both strategies, the experiments' results show that our end-to-end network outperforms previous studies and establishes new state-of-the-art performance on DEAP and MAHNOB-HCI datasets.

References

1. Alhagry, S., Fahmy, A.A., El-Khoribi, R.A.: Emotion recognition based on EEG using LSTM recurrent neural network. *Emotion* **8**(10), 355–358 (2017)
2. Anubhav, Nath, D., Singh, M., Sethia, D., Kalra, D., Indu, S.: An efficient approach to EEG-based emotion recognition using LSTM network. In: *IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*, pp. 88–92 (2020)
3. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: *International Conference on Computer Vision* (2017)
4. Delorme, A., Makeig, S.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**(1), 9–21 (2004)
5. Dzedzickis, A., Kaklauskas, A., Bucinskas, V.: Human emotion recognition: review of sensors and methods. *Sensors* **20**(3), 592 (2020)
6. Ekman, P.: Are there basic emotions? *Psychol. Rev.* **99**, 550–553 (1992)
7. Freund, Y., Schapire, R.E., et al.: Experiments with a new boosting algorithm. In: *ICML*, vol. 96, pp. 148–156 (1996)
8. Ghimire, D., Lee, J.: Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors* **13**(6), 7714–7734 (2013)