Facial Emotion Recognition Using Computer Vision

Conference Paper · September 2018

DOI: 10.1109/INAPR.2018.8626999

5 authors, including:

Jonathan Jonathan
Binus University
1 PUBLICATION **1** CITATION
SEE PROFILE

Andreas Pangestu Lim
Binus University
1 PUBLICATION **1** CITATION
SEE PROFILE

I Gede Putra Kusuma Negara
Binus University
22 PUBLICATIONS **92** CITATIONS
SEE PROFILE

Facial Emotion Recognition Using Computer Vision

Jonathan
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
jonathan016@binus.ac.id

Andreas Pangestu Lim
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
andreas.lim@binus.ac.id

Paoline
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
paoline@binus.ac.id

Gede Putra Kusuma
Computer Science Department, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
inegara@binus.edu

Amalia Zahra
Computer Science Department, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
amalia.zahra@binus.edu

***Abstract*—This paper examines how human emotion, which is often expressed by face expression, could be recognized using computer vision. The study is performed by analyzing journals and researches related to this topic, ranging from psychological to technological journals. A number of algorithms and techniques have been reviewed, and at the end of this paper, a summary of recommendation for performing facial emotion recognition based on the reviews of the techniques/methods is given.**

Keywords— artificial intelligence, computer vision, facial emotion recognition

I. INTRODUCTION

Human face is extremely important in our daily lives. Not only it serves as a placeholder for our senses such as eye, mouth, nose, and skin, it is also a body feature that could be used as the main source of information to identify a person. While humans have no problem at identifying faces, many researchers state that it is hard to understand how human mind could process and interpret faces so easily [1].

This difficulty is mostly caused by the orientation of face [2]. The works of Jourabloo and Liu [3] and Zhu et al. [4] pointed that the change of orientation of face (for example, from completely facing a camera to deviating 30 degrees to the left) could result in several face features to be hidden, and without those features, the face cannot be modelled and cannot be recognized. This problem is solved with face alignment, where the algorithm for face alignment tries to locate face features based on the face, despite the different angle of the face (pose). There are many face alignment algorithms offered, but most are only limited to small to medium poses, with maximum degree of face deviation of 45 degrees. Moreover, there are only few published researches of face alignment for large poses (about 90 degrees of deviation) algorithms. The consequence is there are only a few or, radically speaking, no proper model for developing a reliable system to interpret human face or even facial expressions; furthermore, an intelligent system [5].

An intelligent system that processes seen items such as image and video (including but not limited to human face) is studied under the field of computer vision. One key goal that computer vision researchers try to accomplish is to create an intelligent face recognition system that matches or even surpasses human's capabilities to do the same task [6]. However, despite the collaboration between many disciplines [1], the lack of proper model for understanding human facial expression leads to the difficulty of accomplishing such a key goal.

Even though the difficulty of accomplishing the key goal of computer vision exists, there are some progresses made. These progresses include the recognition that eyebrows play an important part in face recognition [1] [6], the continuous and categorical model in describing how humans understand emotions in facial expression [5], generating face using Viola-Jones algorithm to detect human face [7], and other developments. These developments are fundamental in the attempt of making computer vision advanced even further than mankind, but to accomplish more, further developments must be made.

This paper discusses published journals and articles that are related to facial emotion recognition while defining the best approach to do facial emotion recognition using computer vision. The reviewed literatures include articles, journals, and researches conducted by academicians and researchers that cover many aspects related to this paper's topic. To ease reader's understanding, this paper will first review emotion and how it relates to facial expression. Then, we continue with the discussion on how face and facial expression are recognized and understood by computers. Afterwards, computer's algorithms and tools to perform emotion recognition are elaborated. Lastly, a comparison and recommendation of algorithms and tools used for emotion recognition is presented.

II. LITERATURE REVIEW

A. Emotion and Facial Expression

Emotion is often referred to as a conscious experience that one undergoes, where in that experience exists a corresponding specific affective state [8]. Generally, emotions could be classified as either joy, surprise, anger,

978-1-5386-9422-0/18/$31.00 ©2018 IEEE

sadness, disgust, or fear [5]. These emotions are produced by many neural links and organs in our brain, including the amygdala, a part of our brain that is located near our hippocampus, in the temporal lobe's frontal portion (see Fig. 1). Amygdala, collaborating with other brain parts, produces emotions based on stimulus or stimuli. For instance, a threatening stimulus would produce neural response in the form of fearful expressions [9]. However, it is appropriate to note that amygdala is not as critical as it seems in producing emotions [8].

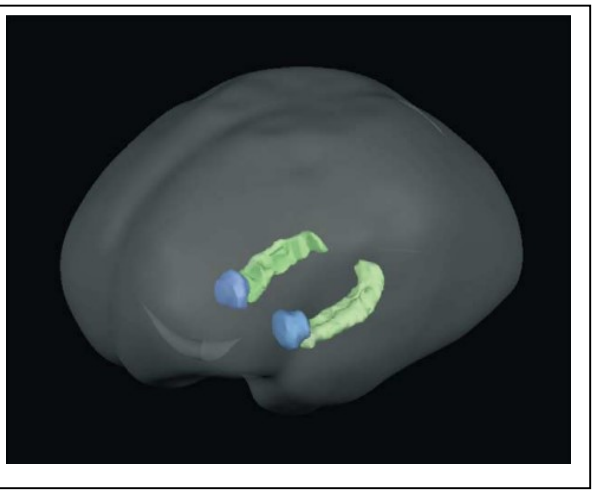

Fig. 1 - Amygdala (Blue) Location [10]

Emotions do not exist only to be felt by humans, but humans would express them, either consciously or unconsciously. Most of the time, emotion is expressed by face expression [11], which Martinez and Du stated as an engineering marvel as the face muscles allow us to produce many face expression configurations [5]. These configurations could be set by proper training or by predetermined configurations, but only several emotion expressions could be well-recognized by humans. In addition, Bartlett et al. elaborated that Charles Darwin recognizes facial expression as a mean for communicating emotions [12]. Therefore, it can be inferred that facial expression is closely related to human emotion since it is often generated based on emotion.

B. Computer's Perception of Face

After understanding that facial expression is often dependent on human emotion, to make computer capable of performing facial emotion recognition, we need to understand how face could be detected and processed by computers.

Machine (computer) could only process things that it has been introduced to, and to do so, computer requires a pattern as its main reference and comparison resource. The same principle applies to researchers' attempts to make computer capable of detecting and recognizing human face. Computer needs a model that could represent the human face, and that model is later trained and then tested using positive and negative values.

Moura and Ferreira-Lopes cited that one development made by Paul Viola and Michael Jones [13] enabled computer vision researchers to make a system that detects object(s), including face, in real-time [7]. The Viola-Jones technique or algorithm was later extended to create OpenCV, a free open-source library that uses a cascade function trained by positive and negative images, which is then used to detect objects. Other than OpenCV and the Viola-Jones technique, many techniques were developed that explore the face through different approaches.

One approach that seems to be quite adequate and efficient is the one conducted by Wong et al. [14]. The paper pointed occurring problems when performing face detection and facial feature extraction, where the difficulty lied in locating face's position, variables that may affect detection performance, and run time when extracting facial features. The algorithm devised in the paper uses genetic algorithm to detect face regions and eigenface technique to verify the fitness of those regions, which could be applied to detect one or more faces. It first evaluates all valley regions in a gray-level image to find the possible eye region using genetic algorithm. Afterwards, the shirring effect is normalized using a transformation matrix that is dependent on the degree of shirr. The process is then continued by normalizing the possible face's histogram region to erase the possibility of asymmetrical face in the image because of lighting effect. After these processes, the normalized possible face region is projected to eigenface space to see if that face region is a face or not, and the output is stored as fitness value where only high fitness values indicating possible face regions are selected to be verified further and undergo feature extraction. The base for verification is the symmetry of a face region and existence of different face features, which are limited to eyebrows, eyes, nose, and mouth that are projected. If both y-projection and x-projection of the facial feature extraction return a valid result (all facial features exist), then the face region is declared as face and its features are located. The results of this algorithm are reliable, since the algorithm produces hit rates above 81% for the test data from MIT and some other images, despite several failures due to external conditions such as moustache and the use of glasses.

C. Algorithms and Tools for Facial Emotion Recognition

The explanation presented in the previous subsection is just one among many modifications made to detect faces. However, in order to perform facial emotion recognition, we need to define algorithm(s) that process images more advanced than just a detection to be implemented by computers using tools provided or designed by ourselves. Several algorithms and tools have been developed or used to accomplish this task, and the results will be reviewed in the following paragraphs.

The first study that discusses an algorithm to accomplish the task was conducted by Bartlett et al. [12]. The paper attempted to make a system capable of automatically detecting frontal face from a video stream and classified what emotion that face was showing based on its facial expressions. In the paper, emotions expressed by facial expression were classified as either happiness, sadness, surprise, disgust, fear, anger, or neutral. The first action taken by the system was classifying images to either face or non-face based on a development of Viola-Jones' work [13]. Afterwards, filters that contained Haar Basis functions from cascade of classifiers were chosen using feature selection

procedure that was based on Adaboost. This resulted in much faster process than comparing all possible filters. The process was then continued with iteratively adjusting weights over the examples according to each performance using Adaboost rule until a classifier reached the minimum desired performance rate. The image was later rescaled and converted into a Gabor magnitude representation. To proceed with the facial emotion recognition, the paper classified facial expressions using SVM (Support Vector Machine) with linear and RBF (Radial Basis Function) kernels. The result was later compared to the results obtained from Adaboost, but SVM was found to be faster in training rather than Adaboost. However, the general performance of Adaboost was significantly faster than SVM. The paper also discusses the combination of Adaboost and SVM, called AdaSVM's, where Adaboost chose Gabor Features which were then used for SVM training. This combination outperformed both SVM and Adaboost significantly (see Table 1).

TABLE 1 BARTLETT ET AL. PERFORMANCE COMPARISON OF ADABOOST, SVM, AND ADASVM FOR 48X48 IMAGES [12]

	Leave-group-out		Leave-subject-out	
	Adaboost	***SVM***	***SVM***	***AdaSVM***
Linear	85.0	84.8	86.2	88.8
RBF		86.9	88.0	90.7

The second study reviewed is the use of deep-learning in recognizing emotions by Ko [15]. Deep-learning methods in performing face emotion recognition include CNN-based (Convolutional Neural Network) and a hybrid CNN-LSTM-based (CNN-Long Short-Term Memory) method. CNN can reduce dependency of physics-based models and/or other pre-processing techniques. A CNN does this by enabling "end-to-end" learning directly from input images. To do so, CNN is equipped with three layers: convolution layer, max pooling layer, and fully connected layers. Convolution layer takes on image input or feature map and convolves the input with a filter banks set to output feature maps that represent the facial image's spatial arrangement. The max pooling layer averages or max-pools the given input to reduce dimensions and ignore noises. The fully connected layers use the original image to compute the class scores. However, CNN is unable to reflect temporal variations in facial components. This drawback is overcome by the hybrid method, CNN-LSTM. LSTM is a special type of RNN (Recurrent Neural Network) that possesses a chain-like structure despite the repeating modules sharing a different structure. The LSTM model itself is straightforward in tuning with other models, and it supports both fixed- and variable-length inputs or outputs. It is important to note that many types have been derived from CNN-based and CNN-LSTM-based methods in face emotion recognition [15]. In all, deep-learning applied in recognizing face expression has many benefits, including easy-to-use structure and high performance. However, large dataset and massive computational device are necessary for the method to function well.

Another study that implemented CNN was conducted by Duncan et al. [16] who used VGG_S model, created based on CNN, that was retrained using dataset from CK+ (Cohn-Kanade) dataset and JAFFE (Japanese Female Facial Expression) database to predict the emotion shown on a live video stream. To accomplish this, from an input video, Duncan et al. used Haar-Cascade filter from OpenCV to find face(s) on the screen. Then, the detected faces were analyzed through five convolutional layers, three fully-connected layers which were retrained, and a softmax classifier (see the architecture in Fig. 2). From the softmax classifier, a prediction was produced by labelling the face using one of the six emotion labels that were used in the experiment, which were angry, fear, happy, neutral, sad, and surprise. After the faces' emotions had been predicted, the faces were replaced with an emoji symbolizing the emotion. The result produced reached over 90% accuracy level for the inputs that had perfect lighting, camera position at eye level, and subject(s) facing the camera showing exaggerated expression. However, for the inputs that did not have the specified attributes, the accuracy level might plummet significantly, which was one key issue that still needed to be corrected.

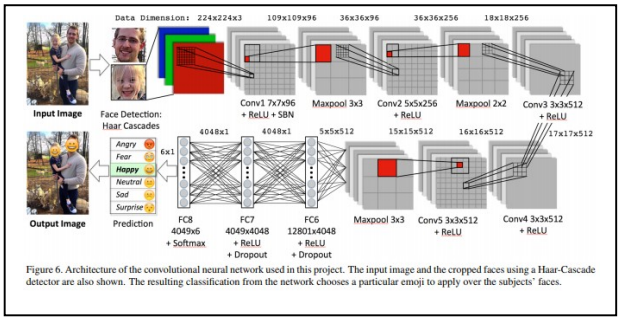

Figure 6. Architecture of the convolutional neural network used in this project. The input image and the cropped faces using a Haar-Cascade detector are also shown. The resulting classification from the network chooses a particular emoji to apply over the subjects' faces.

Fig. 2 - Duncan et al. CNN Architecture [16]

The use of AU (Action Units) with the help of k-NN (k-Nearest Neighbor) and MLP (Multilayer Perceptron) is useful in recognizing facial emotion as well. As Tarnowski et al. [17] researched, the use of AU (six AU units) formed with the help of Microsoft Kinect 3D for face modelling resulted in distinguishable emotions based on the AUs' distributions. The AUs formed a face model, and the 3D model was used to train the classifier. The selected results were the ones during the last four seconds of recording to minimize noise caused by respondent's preparation for expressing certain emotion, and they were used to train the k-NN and MLP. The k-NN (3-NN) is three-nearest neighbors classifier, and the MLP is a two-layer neural network classifier with seven neurons in the hidden layer (see Fig. 3). Back propagation algorithm with conjugate gradient method was used to train the neural network, and the experiment was performed by comparing the results produced by 3-NN and MLP. There were two ways performed by Tarnowski et al. to recognize emotions [17]: subject-dependent -for each user separately- and subject-independent -for all users together-. The dataset used were divided into two distributions: the first was a random distribution and the second was a "natural" distribution. Both 3-NN and MLP produced good results for both subject-dependent and subject-independent ways of recognizing emotions, and for both dataset distributions, with an average of 90% accuracy for random dataset and 70% accuracy for "natural" dataset. However, changing head orientation affected the AU coefficients value greatly, and resulted in 20% decrease in accuracy level for MLP.

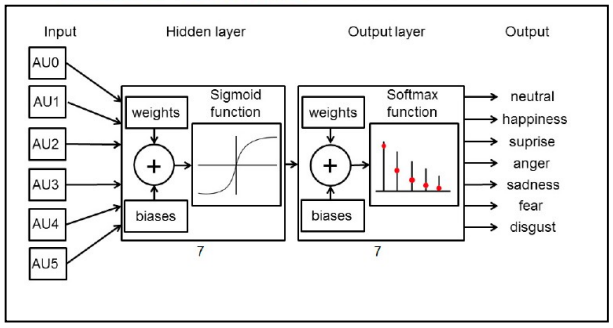

Fig. 3 - Tarnowski et al. Neural Network Structure [17]

Another research reviewed was the attempt of recognizing facial expression from video with the use of Bayesian networks as a 'static' approach in recognizing facial expressions [11]. Cohen et al. performed this by comparing two models: Naive-Bayes classifier with the change of distribution from Gaussian to Cauchy, and the Gaussian-TAN (Tree-Augmented Naive Bayes) classifiers [11]. To track human face, the paper used a system called PBVD (Piecewise Bézier Volume Deformation) tracker created by Tao and Huang [18]. The 'static' approach aimed to classify each frame in a video to one category of facial expressions. By using the Naive-Bayes classifier, the Cauchy distribution indicated higher classification results than those of Gaussian. The Naive-Bayes classifier assumes that the features are independent, given the class. However, this becomes the problem of Naive-Bayes classifier because such independence assumption may be too strong when being applied. To overcome this, TAN classifier was used using the assumption that the features were Gaussian. After learning the data's structure, Gaussian-TAN classifier yielded more optimal result with small increase in complexity in the Naive-Bayes classifier. However, Gaussian-TAN should only be used when sufficient training data exists and when the learned structure is reliable. In addition, it is difficult to use.

While previous methods seemed to be complex, the work of Huang and Tai [19] used SURF (Speeded-Up Robust Features) as SURF is deemed more powerful than SIFT (Scale Invariant Feature Transform). Besides, SURF has faster performance and is comparable in terms of repeatability, distinctiveness, and robustness to other existing schemes. To perform facial emotion recognition, the first step was to capture keypoint descriptors. Using Hessian matrix's determinant, the keypoint detector could be located. Keypoint detector was assigned with finding keypoints in the image for the keypoint descriptor to describe the feature and constructed the feature vectors of those keypoints. The keypoint descriptor itself was created by using Haar wavelet responses and splitting each region into smaller sub-regions. Then, the results (feature vectors) were normalized to unit length, thus creating PDF (Probability Density Function) descriptors and the selected ones were used for classification. To select PDF descriptors, KL (Kullback Leibler) divergence was used. The KL divergence was useful in calculating the distance between two PDF descriptors, and a recognition tally was created using the PDF descriptors. The recognition tally is a utilized equation, and it calculates and extracts important features. The denominator of the equation is the representation of the minimum PDF descriptor of a class, while the numerator is the representation of the minimum PDF descriptor of classes other than the denominator's represented class. From the recognition tally, a 4x4 uniform grid was created, where in each grid there existed selection of four largest PDF descriptors to be used for classification. The grid was then masked with Gaussian mask, with more weight in the center of the mask as Huang and Tai deemed that the central part of the face was the container of most important information. The masked result was then classified by WMV (Weighted Majority Voting) which classified an image based on the sum of matching scores of each grid. An example of WMV classifier can be seen in Fig. 4. The result of this method produced an average of 93% recognition rate for different subjects, whereas SIFT only had an average of 71.67% of recognition rate for the same subject category. Note that the accuracy of recognition rate using this method relies more on the number of PDF descriptors, and that excessive number of PDF descriptors would result in lower performance.

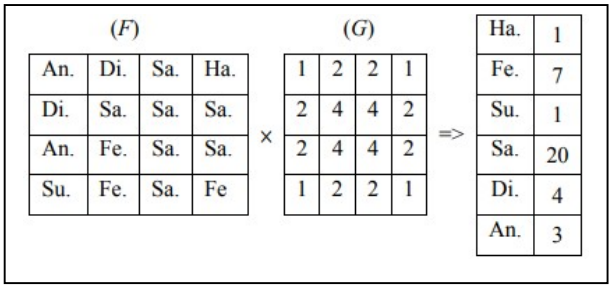

Fig. 4 - Huang and Tai WMV Classifier [19]

III. Conclusion

A. Summary

This paper has discussed a number of aspects ranging from emotion and how it is related with facial expression, computer's understanding of face, to the approaches for accomplishing facial emotion recognition. The discussions are based on international journals and researches published from varying periods, ranging from 2003 to 2018. The researches reviewed are works of Cohen et al. about the use of Naïve-Bayes and Gaussian-TAN [11], Bartlett et al. about the use of Adaboost, SVM, and AdaSVM [12], Ko about the use of CNN and CNN-LSTM [15], Duncan et al. about the use of CNN with VGG_S model [16], Tarnowski et al. about the use of k-NN and MLP [17], and lastly by Huang and Tai about the use of SURF and WMV classifier [19].

Facial expression is dependent on emotion. To recognize a facial expression, an intelligent system must detect whether a face exists or not in an input, then perform feature extraction, and do recognition process. This kind of intelligent system is developed under the field of computer vision.

Six papers have been discussed, and we have seen how several algorithms and tools in those papers have been defined and/or designed to do facial emotion recognition, including SVM, Adaboost, neural networks (CNN, CNN-LSTM, k-NN, MLP), Bayesian networks (Naive-Bayes and Gaussian-TAN), VGG_S, Microsoft Kinect 3D, and SURF.

978-1-5386-9422-0/18/$31.00 ©2018 IEEE

Each algorithm and/or tool has its own merits and drawbacks.

B. Merits and Drawbacks of Each Algorithm and/or Tool

Using AdaSVM offers a better performance than Adaboost and SVM, and is quite simple to be implemented. The use of CNN-LSTM, which is based on CNN, offers a structure that is easy-to-use and high performance, but it needs some expensive resources, which include large dataset and massive computational devices. Using VGG_S, a CNN-based model that is retrained produces high level of accuracy but it has many constraints for the best result. AU with k-NN and MLP offers a unique perspective as it models the face at first, then classifies the emotion shown. However, this sequential process might be time costly, and minor deviance could decrease the accuracy level. The last one is the use of Bayesian networks that compare two classifiers: Gaussian-TAN and Naive-Bayes. Both classifiers show good performance, but Gaussian-TAN requires expensive resources and is difficult to use, and Naive-Bayes produces less optimal results. The use of SURF produces good results, but to use SURF in facial emotion recognition, we need to use the appropriate number of PDF descriptors since the result of WMV classifier is dependent on the number of PDF descriptors.

All of them offer a solution from many perspectives for recognizing emotion from facial expression, but most of them cannot cope with deviance such as orientation, lighting, and the number of dataset used for training. These issues are challenging, but they are necessary and must be overcome for facial emotion recognition to become better. Also, these issues open possibilities for future research.

C. Recommendation

We believe that researchers may have different goals and considerations in choosing which algorithm to use in conducting their research, and because of that, we would like to give our recommendations based on the reviews of the discussed papers.

If we aim for results with high level of accuracy without many datasets, it is better to use SURF and/or neural network using additional face model structure. Results produced from researches using any of these methods are accurate with proper dataset for training. However, we should take care of every constraint that our CNN-based model has, or the number of PDF descriptors for SURF method.

If we have many datasets, the use of CNN-LSTM and/or Bayesian networks is appropriate for classifying emotions as they would produce high level of accuracy results based on the large training datasets. However, they are both quite difficult to be implemented. An effort to combine deep neural network, which is CNN, and Bayesian network has been undertaken in [20], and the result seems promising. Thus, it might be interesting to also apply a hybrid method by combining CNN-LSTM and Bayesian network and observe the outcomes.

If we would like to implement a simple yet functional model, we could use AdaSVM or a retrained model such as VGG_S. They offer high quality results, with the drawback of not being able to restructure the model, unlike other methods.

REFERENCES

[1] R. Chellapa, P. Sinha and P. J. Phillips, "Face recognition by computers and humans," Computer, pp. 46–55, February 2010.

[2] N. Rathore, D. Chaubey and N. Rajput, "A survey on face detection and recognition," *International Journal of Computer Architecture and Mobility,* vol. 1, no. 5, March 2013.

[3] A. Jourabloo and X. Liu, "Large-pose face alignment via CNN-based dense 3D model fitting," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 4188–4196, June 2016.

[4] X. Zhu, Z. Lei, X. Liu, H. Shi and S. Z. Li, "Face alignment across large poses: A 3D solution," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 146–155, June 2016.

[5] A. Martinez and S. Du, "A model of the perception of facial expressions of emotion by humans: research overview and perspectives," Journal of Machine Learning Research, vol. 13, pp. 1589–1608, May 2012.

[6] P. Sinha, B. Balas, Y. Ostrovsky and R. Russell, "Face recognition by humans: 20 results all computer vision researchers should know about".

[7] J. M. Moura and P. Ferreira-Lopes, "Generative face from random data, on 'How computers imagine humans'," Proceedings of ARTECH2017 8th International Conference on Digital Arts, 6-8 September 2017.

[8] A. K. Anderson and E. A. Phelps, "Is the human amygdala critical for the subjective experience of emotion? evidence of intact dispositional affect in patients with amygdala lesions," Journal of Cognitive Neuroscience, vol. 14, no. 5, pp. 709–720, 2002.

[9] J. S. Morris, K. J. Friston, C. Büchel, C. D. Frith, A. W. Young, A. J. Calder and R. J. Dolan, "A neuromodulatory role for the human amygdala in processing emotional facial expressions," Brain, vol. 121, pp. 47–57, 1998.

[10] A. E. Phelps, "Human emotion and memory: interactions of the amygdala and hippocampal complex," Current Opinion in Neurobiology, vol. 14, p. 198–202, 2004.

[11] I. Cohen, N. Sebe, A. Garg, L. S. Chen and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," Computer Vision and Image Understanding, vol. 91, pp. 160–187, 2003.

[12] M. S. Bartlett, G. Littlewort, I. Fasel and J. R. Movellan, "real time face detection and facial expression recognition: development and applications to human computer interaction," Computer Vision and Pattern Recognition Workshop, vol. 5, July 2003.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Accepted Conference on Computer Vision and Pattern Recognition, 2001.

[14] K.-W. Wong, K.-M. Lam and W.-C. Siu, "An efficient algorithm for human face detection and facial feature extraction under different conditions," Pattern Recognition, vol. 34, pp. 1993–2004, 2001.

[15] B. C. Ko, "A brief review of facial emotion recognition based on visual information," Sensors, vol. 18, no. 401, 30 January 2018.

[16] D. Duncan, G. Shine and C. English, "facial emotion recognition in real time," 2016.

[17] P. Tarnowski, M. Kołodziej, A. Majkowski and R. J. Rak, "Emotion recognition using facial expressions," Procedia Computer Science, vol. 108C, pp. 1175–1184, 12-14 June 2017.

[18] H. Tao and T. S. Huang, "Connected vibrations: A modal analysis approach to non-rigid motion tracking," Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998.

[19] H.-F. Huang and S.-C. Tai, "Facial expression recognition using new feature extraction algorithm," Electronic Letters on Computer Vision and Image Analysis, vol. 11, no. 1, pp. 41–54, 2012.

[20] L. Surace, M. Patacchiola, E. Battini Sönmez, W. Spataro, and A. Cangelosi, "Emotion recognition in the wild using deep neural networks and Bayesian classifiers," Proceedings of the 19th ACM International Conference on Multimodal Interaction, ACM, pp. 593–597, November 2017.