Towards emotion recognition in immersive virtual environments: A method for Facial emotion recognition

Kahina Amara
Centre of developpement of advanced technologies
Algiers, Algeria
kahina.amara88@gmail.com

Naeem Ramzan
School of Engineering and Computing, University of the West of Scotland
Paisley, Scotland

Nadia Zenati
Centre of developpement of advanced technologies
Algiers, Algeria

Oualid Djekoune
Centre of developpement of advanced technologies
Algiers, Algeria

Cherif Larbes
ENP Ecole Nationale politechnique
Hassen Badi Avenue, Algiers, Algeria

Mohamed Amine Guerroudji
Centre of developpement of advanced technologies
Algiers, Algeria

Djamel Aouam
Centre of developpement of advanced technologies
Algiers, Algeria

***Abstract*—Virtual Reality (VR) is, thus, proposed as a powerful tool to simulate complex, real situations and environments, offering researchers unprecedented opportunities to investigate human behaviour in closely controlled designs in controlled laboratory conditions. Facial emotion recognition has attracted a great deal of interest for interaction in virtual reality, healthcare system: therapeutic applications, surveillance video application etc. In this paper, we propose a method for facial emotion recognition for immersive virtual environment based on 2D and 3D geometrical features. We used our collected dataset of 17 subjects' performance of six basic facial emotions (anger, fear, happiness, surprise, sadness, and neutral) using three devices: Kinect (v1), Kinect (v2), and RGB HD camera. In addition, we present the performance results of the RGB data for facial emotion recognition using Bagged Trees algorithm. To assess the performance of the proposed system, we used leave-one-out-subject cross-validation. We compared the 2D and 3D data performance for facial expression recognition. The obtained results show the superior performance of the RGB-D features provided by Kinect (v2). Our findings highlight that the 2D images are not robust enough for facial emotion recognition. The built facial emotion models will animate virtual characters that can express emotions via facial expressions. This could be deployed for Chatting, Learning and Therapeutic Intervention.**

***Index Terms*—Virtual Reality, Facial emotion recognition, Immersive Environment, Avatar animation, Interaction, RGB, RGB-D, Machine Learning, Geometrical features.**

I. Introduction

Emotions have a critical impact in our daily lives, so the understanding and recognition of emotional responses is crucial for human behaviour understanding. Emotion recognition research has mostly used non-immersive two-dimensional (2D) images or videos to elicit emotional states. However, immersive virtual reality, which allows researchers to simulate environments in controlled laboratory conditions with high levels of sense of presence, immersion, and interactivity, is becoming more popular in emotion research. Moreover, its synergy with implicit measurements and machine-learning techniques has the potential to impact transversely in many research areas, opening new opportunities for the scientific community.

Healthcare, education and training are examples of application area where VR has been much applied (figure 1. The studies showed that VR can offer great educational advantages. It can solve time-travel problems, for example, students can experience different historical periods. VR can address physical inaccessibility, for example, students can explore the solar system in the first person. It can circumnavigate ethical problems, for example, students can "perform" serious surgery. Surgical training is now one of the most analysed research topics. On the other side, several researchers have also showed the effectiveness of VR in therapeutic applications. To overcome certain inconveniences such us the lack of the dynamism that is inherent to facial expressiveness with some patients, over the last years different authors have proposed the use of virtual avatars. The main goal is to make use of virtual environments and avatars to provide new objective methods for assessing patients' interpersonal behaviour characteristics [1]. VR offers some distinct advantages over standard therapies,

Fig. 1. Virtual reality application: Healthcare, Movie industry

including precise control over the degree of exposure to the therapeutic scenario, the possibility of tailoring scenarios to individual patients' needs and even the capacity to provide therapies that might otherwise be impossible. Taking some examples, studies using VR have analysed the improvement in the training in social skills for persons with mental and behavioural disorders, such as phobias, schizophrenia (SZ) [2], [3] and autism [4]. Moreover, it has been proposed as a key tool for the diagnosis of neuro-developmental disorders [5]. In [2], the authors present a VR-based system that incorporates implicit cues from peripheral physiological signals and eye tracking for the under-standing of facial emotional expression. They compare how a SZ group and a matched group of healthy non-psychiatric adults performed emotion recognition tasks when presented in the form of static slides and when presented in a VR environment with the avatars expressing emotions dynamically. Virtual reality (VR) has been observed to improve both assessment and training of emotional recognition skills of people with Severe mental illness [3]. In [7], a method EEG-based feature extraction technique is presented for emotion recognition using higher order crossings (HOC).

Virtual reality has proven its potential in the movie industry. Reproducing facial and bodily emotions in a completely immersive virtual environment (VE) in a faithful manner is of paramount importance for real virtual rendering (figure 2). Facial emotion expressions are nonverbal way of expressing feeling. Many studies addressed the facial expression [6]–[8].

Recent approaches use 3D facial points, such techniques have gained more attention lately due to the proliferation of affordable commodity depth sensing devices, such as the Kinect. According to the data used in this work, different approaches have been proposed for feature extraction. Positional and temporal features have been investigated in [9] using the facial data collected by Kinect. They defined a feature vector composed of the coordinates of tracked points and Euclidean distance between the tracked points and the angle between those points for the positional features.

Many studies are based on 2D images for facial emotion recognition. In [6], the authors proposed a software for the analysis of facial behaviour. Furthermore, several approaches have demonstrated state-of-the-art performance on RGBD input, or only depth input. We can cite the work presented in [8], [10]. In [10], the authors proposed the skeleton based approach to extract facial features for facial emotion recognition by using a depth camera. Billy et al. [11] used Kinect sensor to recognise emotions under different conditions. The authors used a publicly available database which contains facial images (RGB-D) captured by Kine ct sensor with different poses, expressions, illumination and disguise. Their results demonstrated that using RGB-D information could improve the performance of facial emotion recognition compared with the methods using 2D information. The conventional approaches for facial emotion recognition still suffer from some constraints and limitations which directly affect the system performance [12]. We can cite among these problems, the lack of publicly available database, the environmental changes including illumination changes, the different personnel style for emotion expression.

The existing approaches for facial expression recognition suffer from some constraints and limitations. Firstly, the lack of publicly available database. Furthermore, the selection of non-significant features for depicting different expressions can cause model failure. To deal with these problems, we propose in this work a system for mono-modal facial expression recognition based on facial movements.

In this paper, we present a proposal for facial emotion classification for interaction in immersive virtual environment. The facial emotion recognition is based on 3D angle and 3D Euclidean distance features for the RGB-D data, provided by kinect sensors, and 2D angle and Euclidean 2D distance for the RGB data provided by HD RGB camera. This paper consists of four sections. In the second section, we describe the proposed approach, the feature extraction will be presented in detail. In the third section, we discuss the experimental results. The conclusions and future works will be drawn in the last section.

II. Our proposal

Emotion can be expressed in different ways and plays important role in daily life. The facial expression is a common, nonverbal and effective way of expressing emotion. The presented work is included in this area; the process for facial expression recognition is depicted in figure **??**, we aim to distinguish the expressions as accurate as possible by establishing computational models. We carried out experiments on synthetic RGB-D sequences captured by Kinect (v1) and Kinect (v2) sensors and RGB sequences recorded using RGB HD camera. In this work, we target six basic emotions (anger, fear, happiness, surprise, sadness, and neutral).

We collected our own database including the performance of 17 students (9 male and 8 female) recruited from the School of computing and engineering at the University of West of Scotland. The participants are from different cultures with different skin colour. In order to obtain actual facial expression data, we conducted an emotion priming experiment using different emotional videos. The subjects were first asked to perform emotional states depicted on projected images on screen. In the second part of the experiment, we used 20 emotional videos used in [13] collected from Youtube. We used many types of emotional videos including neutral, happy, surprise, anger, fear and sad video which could induce corresponding emotional state. The participants were asked to perform their feeling according to their personal style. They have to repeat the performance for three times. The face-recorded videos were segmented and stored in a database. The collected dataset contains 1581 RGB videos recorded by RGB HD camera and more than 3000 synthetic RGBD sequences captured by Kinect (v1) and Kinect (v2). Figure displays the participants' facial expression.

In this work, the facial expressions were tracked using the face and skeleton tracking API available in the Microsoft Kinect Software Development Toolkit. The synthetic RGB-D sequences captured by Kinect sensors provided 3D facial

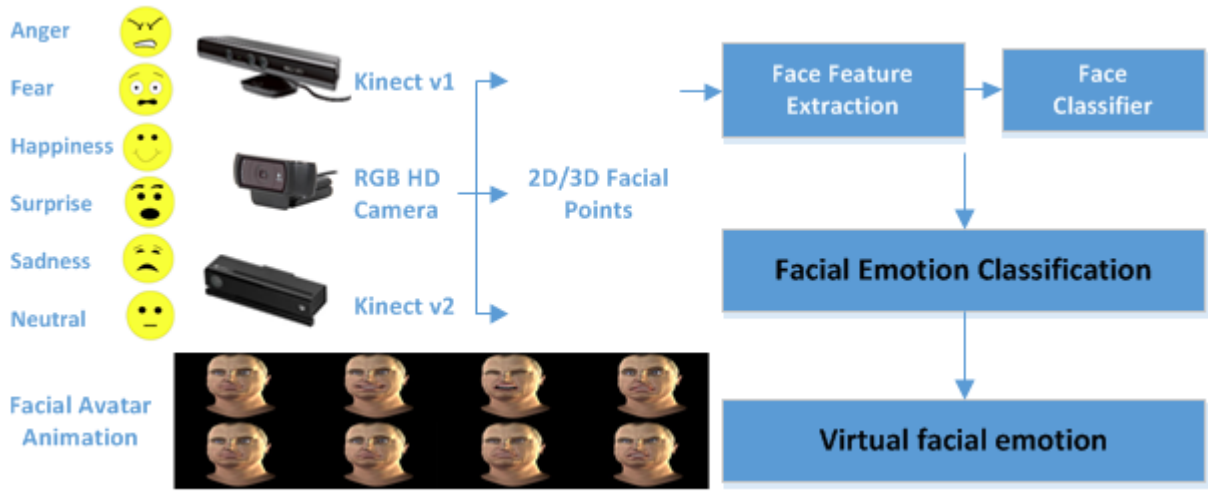

Fig. 2. The proposed framework for virtual facial expression recognition.

points. We choose representative points, which represent significant movement in order to describe the subtle changes of facial expression. The face tracking Kinect toolkit provides 121 facial points for Kinect 1 and 1347 facial points for Kinect 2. However, not all of these points are significant to facial expressions. In [17], the authors proposed that the main areas englobing the eyes, eyebrows, and the mouth are involved in facial expression displays. Out of the available points, facial points around eyebrows, eyes, mouth, nose, chin and cheeks were tracked and some other key positions were finally selected to improve the recognition accuracy. For the RGB video recorded by RGB HD camera, we used the open source tool OpenFace [6]. It provides facial landmark using the Conditional Local Neural Fields (CLNF) [15] (see figure 3). The CLNF performs the detection of 68 facial landmark. The CLNF is an instance of a Constrained Local Model (CLM) [16]. The CLM is composed of two main components: Point Distribution Model (PDM) which captures landmark shape variations; patch experts, which capture local appearance variations of each facial landmark [6]. Finally, we choose 26 facial points. The new face feature vector (FV_{Face} is defined using geometric features: distance and angle with the horizontal axis and the coordinates of tracked points from one frame.

Given two facial points $P_n^{face}(t)$ and $P_{n-1}^{face}(t)$ with coordinates $(x_n(t), y_n(t), z_n(t))$ and $(x_{n-1}(t), y_{n-1}(t), z_{n-1}(t))$ respectively at frame t,

$$D_n^{face}(P_{n-1}^{face}(t), P_n^{face}(t) = \begin{cases} (x_{n-1}(t) - x_n(t)) \\ (y_{n-1}(t) - y_n(t)) \\ (z_{n-1}(t) - z_n(t)) \end{cases} \quad (1)$$

$$\theta_n^{face}(P_{n-1}^{face}(t), P_n^{face}(t)) = \begin{cases} \theta(x_{n-1}(t), x_n(t)) \\ \theta(y_{n-1}(t), y_n(t)) \\ \theta(z_{n-1}(t), z_n(t)) \end{cases} \quad (2)$$

$$FV_{Face} = \begin{cases} D_1^{face}(P_0^{face}(t), P_1^{face}(t)), ..., D_n^{face}(P_{n-1}^{face}(t), \\ P_n^{face}(t)), \theta_1^{face}(P_0^{face}(t), P_1^{face}(t)), \\ ..., \theta_n^{face}(P_{n-1}^{face}(t), P_n^{face}(t)) \end{cases} \quad (3)$$

Fig. 3. The facial key points generated bu OpenFace [23]

The feature vector is based on position of the tracked points from one frame. The face feature vector is defined as follows (equation 3). It is a set of distance difference $D(t)$ and $\theta(t)$ which is the angle between each selected facial points which are depicted in equation 1 and equation 2. We calculated 36 distance and 36 angle between each tracked points. We selected key facial points representing significant changes based on psychological studies [17].

Experiments in this study were conducted on a computer with an Intel® Xeon ® CPU E3-1245 v3 3.40 Ghz and 8 GB RAM. All the experiments have been run in Matlab 2016b environment, using Matlab's own implementation of classification algorithms (Bagged Trees, k-NN, Linear SVM). Support vector machine [18] proposed by Vapnik and Chervonenk is a powerful statistical learning method, it models the situation by creating a feature space. The goal is to train

a model that assigns new unseen objects into a particular category. Linear SVM is one method used in statistics and machine learning to find a linear combination of features which characterize or separate two or more classes or events. Since emotion recognition may not be linearly separable, we also considered non-linear classification algorithms. Bagging is a method for improving results of machine learning classification algorithms. This method was formulated by Leo Breiman and its name was deduced from the phrase "bootstrap aggregating" [14]. Bagged Trees can be used to reduce the variance associated with prediction and improve the prediction process. Many bagging samples are drawn from the available data, some prediction method is applied to each bagging sample, and then the results are combined, by simple voting process for classification, to obtain the overall prediction, with the variance being reduced due to the averaging. The bagging method generates additional data for training from the original dataset using combinations with repetitions to produce multi-sets of the same cardinality/size as the original data. By increasing the size of the training set, it can not improve the model predictive force, but just decrease the variance, narrowly tuning the prediction to expected outcome.

The k-NN algorithm as non-parametric lazy learning algorithm is one of the simplest classification algorithm [19]. This is pretty useful , as in the real world , most of the practical data does not obey the typical theoretical assumptions made (gaussian mixtures, linearly separable etc). It is also a lazy algorithm which means is that k-NN does not use the training data points to do any generalization. There is no explicit training phase or it is very minimal and fast . All the training data is needed during the testing phase, the k-NN algorithm keeps all the training data. This is in contrast to other techniques like SVM where it is possible to discard all non support vectors without any problem. Most of the lazy algorithms – especially k-NN – makes decision based on the entire training data set. Predictions are made for a new instance by searching through the entire training set for the k most similar instances (the neighbors) and summarizing the output variable for those k instances. To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance [20] which is calculated as the square root of the sum of the squared differences between a new point and an existing point.

III. Results and discussion

To solve our multi-classification problem, we defined six models to distinguish the six target emotional states (anger, fear, happiness, sadness, surprise, and neutral). As different classifiers may yield to different classification performance for the same dataset, we used for the training linear and non-linear classifiers including Bagged Trees, Fine k-NN and Linear SVM.

one-subject-out validation. We left the performance of one participant for testing, and the data of 16 participants were used for training and. Comparing the results of different

TABLE I
THE OBTAINED FACIAL EMOTION RECOGNITION RESULTS USING k-NN.

Devices	Accuracy%	Recall%	F1-score%	Precision%
Kinect 1	97.09	91.94	91.74	91.67
Kinect 2	97.40	92.86	92.65	92.49

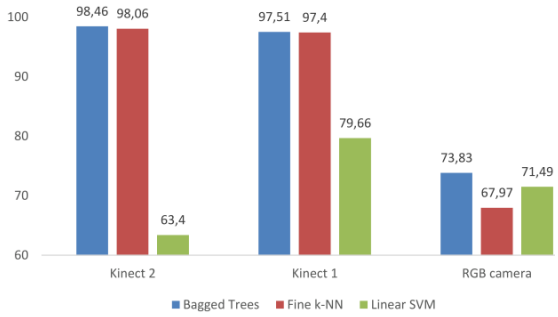

Fig. 4. The classifiers accuracy performance.

training algorithms, we notice that Bagged Trees algorithm outperforms the remaining classifiers with accuracy rate of 98.46%, 97.51%, and 73.83% for Kinect (v2), Kinect (v1) and HD RGB camera respectively. Comparing the devices used to collect the data, the Kinect captors achieved better results than OpenFace, which gives the lowest performance with 73.83%, 67.97%, and 71.49% of accuracy for Bagged Trees, Fine k-NN and Linear SVM respectively. The obtained results showed that Kinect (v2) performs better than Kinect (v1). The figure 4 showcases the obtained results using the three devices. The histogram presents the accuracy rates obtained by each classifier. Based on our experiment, the data collected by the HD RGB camera showed the lowest performance, which can be explained by the sensitivity to the surroundings, especially to illumination conditions [21]. The RGB-D images can capture essential geometrical features, and enable higher precision and preservation of facial details insensitive to different conditions. Table III shows the performance comparison between the proposed work in this paper and state-of-the-art works.

For performance comparison, we choose the accuracy, recall, precision, F1-score metrics that can be estimated by describing random errors (TP: True positive; TN: True negative; FP: False positive; FN: False negative), a measure of statistical changes (equation 4, equation 5,equation 7, equation 6). The obtained results are depicted on table II and table III.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{7}$$

Fig. 5. Future work

For the RGB data collected using RG HD camera, we use the open source OpenFace [23]. The results are depicted in table II.

TABLE II
RGB DATA PERFORMANCE: ACCURACY RATES FOR EACH CLASS USING BAGGED TREES.

Device	Accuracy (%)	anger	fear	happiness	sadness	surprise	neutral
RGB HD camera	72.39	67.62	77.31	64.73	86.92	75	71.39

TABLE III
ACCURACY PERFORMANCE COMPARISON WITH STATE OF THE ART WORKS. $^{*}Kinect1$, ** KINECT 2, ^{1} K-NN, ^{2} BAGGED-TREES.

Works	**Number of classes**	**Accuracy %**
[12] ^{**1}	6	89.44
	8	90.33
[22] *	6	80.75
	7	80.57
[8] ^{**1}	6	96.74
	8	96.92
Proposed approach ^{**2}	6	**98.46**
Proposed approach ^{*2}	6	**97.51**

To the best of our knowledge and based on the comparison depicted on table III, we believe that the proposed approach for emotion recognition outperform the state-of-the-art works.

At this point, we want to reproduce the emotions detected from the facial expressions in a virtual environment. This is by exploiting the models of the six basic emotions already built during the classification phase using machine learning (figure 4). We want the avatar in the virtual environment faithfully reproduce the facial emotions expressed by a user present in a real scene (figure 5). This proposal could be used in healthcare application (Patients with Schizophrenia therapies), serious game, human-computer interaction, etc.

IV. CONCLUSION

The paper proposes a concept for virtual avatar animation. This could be used for therapeutic intervention, chatting: social media interaction, and maybe for learning activities. We presented results of machine learning classification of facial emotion expression. The recognition was mainly based on a combination of 2D, 3D angle and Euclidean distance between facial key points as features carefully selected. We provided a comparison between RGB and RGB-D data performance for facial emotion recognition. Our findings deduced that the non-linear algorithms presented the best performance due the data nature. The Bagged Trees and k-NN consistently outperformed all the tested classification algorithms on our collected dataset. We observed that the 2D data are not robust enough for facial emotion recognition, which are 3D changes of facial expression. The RGB-D data can capture essential geometrical features, and enable higher precision and preservation of facial emotion critical details. The RGB-D data are more insensitive to different conditions compared to RGB data. Future works will concentrate on building real time system for facial virtual avatar animation in an immersive virtual environment.

REFERENCES

[1] Marcos-Pablos, Samuel González Pablos, Emilio Martín-Lorenzo, Carlos Flores Pérez, Luis Gómez-García-Bermejo, Jaime Zalama, Eduardo. 2016. Virtual Avatar for Emotion Recognition in Patients with Schizophrenia: A Pilot Study. Frontiers in Human Neuroscience. 10. 10.3389/fnhum.2016.00421.

[2] Bekele, Esube Bian, Dayi Zheng, Zhi Peterman, Joel Park, Sohee Sarkar, Nilanjan. 2014. Responses during Facial Emotional Expression Recognition Tasks Using Virtual Reality and Static IAPS Pictures for Adults with Schizophrenia. Human-Computer Interaction. 8526. 10.1007/978-3-319-07464-1-21.

[3] Souto, Teresa Silva, Hugo Leite, Ângela Baptista, Alexandre Queirós, Cristina Marques, António. 2019. Facial Emotion Recognition: Virtual Reality Program for Facial Emotion Recognition—A Trial Program Targeted at Individuals With Schizophrenia. Rehabilitation Counseling Bulletin. 63. 003435521984728. 10.1177/0034355219847284.

[4] Nyaz Didehbani, Tandra Allen, Michelle Kandalaft, Daniel Krawczyk, Sandra Chapman, Virtual Reality Social Cognition Training for children with high functioning autism, Computers in Human Behavior, Volume 62, 2016, Pages 703-711, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2016.04.033.

[5] Alcañiz Raya, Mariano Olmos, Elena Abad, Luis. 2019. Use of virtual reality for neurodevelopmental disorders. A review of the state of the art and future agenda. Medicina. 79. 77-81.

[6] T. Baltrušaitis, P. Robinson and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit", 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, 2016, pp. 1-10. doi: 10.1109/WACV.2016.7477553

[7] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion Recognition From EEG Using Higher Order Crossings", in IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 2, pp. 186-197, March 2010. doi: 10.1109/TITB.2009.2034649

[8] N. Chanthaphan, K. Uchimura, T. Satonaka and T. Makioka, "Facial Emotion Recognition Based on Facial Motion Stream Generated by Kinect," 2015 11th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS), Bangkok, 2015, pp. 117-124. doi: 10.1109/SITIS.2015.31

[9] Z. Zhang, L. Cui, X. Liu and T. Zhu, "Emotion Detection Using Kinect 3D Facial Points," 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Omaha, NE, 2016, pp. 407-410.doi: 10.1109/WI.2016.0063

[10] X. Zhao, J. Zou, H. Li, E. Dellandréa, I. A. Kakadiaris and L. Chen, "Automatic 2.5-D Facial Landmarking and Emotion Annotation for Social Interaction Assistance," in IEEE Transactions on Cybernetics, vol. 46, no. 9, pp. 2042-2055, Sept. 2016. doi: 10.1109/TCYB.2015.2461131

[11] B. Y. L. Li, A. S. Mian, W. Liu and A. Krishna, "Using Kinect for face recognition under varying poses, expressions, illumination and disguise," 2013 IEEE Workshop on Applications of Computer Vision (WACV), Tampa, FL, 2013, pp. 186-192. doi: 10.1109/WACV.2013.6475017

[12] N. Chanthaphan, K. Uchimura, T. Satonaka, T. Makioka, "Novel facial feature extraction technique for facial emotion recognition system by using depth sensor", 2016, International Journal of Innovative Computing, Information and Control ,12 2067–2087.

[13] C . A. Gabert-Quillen, E. E. Bartolini, B. T. Abravanel, C. A. Sanislow, Ratings for emotion film clips, Behavior Research Methods 47, 2015, 773–787.

[14] Breiman, L. Machine Learning (1996) 24: 123. https://doi.org/10.1023/A:1018054314350.

[15] T. Baltrusaitis, P. Robinson and L. P. Morency, "Constrained Local Neural Fields for Robust Facial Landmark Detection in the Wild," 2013 IEEE International Conference on Computer Vision Workshops, Sydney, NSW, 2013, pp. 354-361. doi: 10.1109/ICCVW.2013.54

[16] Cristinacce D, Cootes TF. Feature detection and tracking with constrained local models. British Machine Vision Conference. 2006:929–938.

[17] B. Fasel, J. Luettin, Automatic Facial Expression Analysis: A Survey, 2003,Idiap-RR Idiap-RR-19-1999, IDIAP, 1999. Published in Pattern Recognition, 36(1):259-275.

[18] Cortes, C. and Vapnik, V. Machine Learning ,1995, 20: 273. https://doi.org/10.1023/A:1022627411411.

[19] S. Zhang, X. Li, M. Zong, X. Zhu and R. Wang, "Efficient kNN Classification With Different Numbers of Nearest Neighbors," in IEEE Transactions on Neural Networks and Learning Systems, vol. PP, no. 99, pp. 1-12. doi: 10.1109/TNNLS.2017.2673241.

[20] Hui Wang, "Nearest neighbors by neighborhood counting," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 6, pp. 942-953, June 2006.doi: 10.1109/TPAMI.2006.126

[21] S. M. Lajevardi and H. R. Wu, "Facial Expression Recognition in Perceptual Color Space," in IEEE Transactions on Image Processing, vol. 21, no. 8, pp. 3721-3733, Aug. 2012. doi: 10.1109/TIP.2012.2197628

[22] Mao, Qi-rong,Pan, Xin-yu, Zhan, Yong-zhao, and Shen, Xiang-jun,"Using Kinect for real-time emotion recognition via facial expressions", Frontiers of Information Technology & Electronic Engineering,2015,vol.16, no.4, pp.272-282,issn="2095-9230".

[23] B. Amos, B. Ludwiczuk, M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.