Seventh International Conference on Information Science and Technology
Da Nang, Vietnam; April 16-19, 2017

An Efficient Real-Time Emotion Detection Using Camera and Facial Landmarks

Binh T. Nguyen
University of Science
Department of Computer Science
Ho Chi Minh city, Vietnam

Minh H. Trinh, Tan V. Phan and Hien D. Nguyen
University of Information Technology
Department of Knowledge Engineering and Machine Learning
Ho Chi Minh city, Vietnam

***Abstract*—Emotion recognition has many useful applications in daily lives. In this paper, we present a potential approach to detect human emotion in real time. For any face detected in camera, we extract the corresponding facial landmarks and examine different kinds of features and models for predicting human emotion. The experiments show that our proposed system can naturally detect human emotion in real time and achieve an average accuracy about 70.65%.**

***Keywords*—emotion detection; facial landmarks; SVM; human emotion.**

I. Introduction

Emotion recognition has been broadly investigated in both theory and applications in many research fields, for example, computer science [7], neuroscience [1][11], biology [12], psychology [13][14][15] and medicine [16]. It is an important step for machine understanding of human behaviors. Typically, there are six types of human emotions i.e. anger, disgust, fear, happiness, sadness and surprise [6]. To detect human emotion, there are several approaches by using various human behavioral signals such as speech signals [2][5], electroencephalogram (EEG) signals [1] and facial images [9].

Horlings and co-workers [1] proposed an emotion recognition system by brain activity, measured by EEG signals. By extracting 114 features from these signals and analyzing classification with three learning models (artificial neural networks (ANNs) [17], support vector machines (SVMs) [18] and naive bayesian classifier (NBC) [18]), they designed a system that could predict five distinct emotions with high accuracy.

Nwe et al. [5] studied the emotion classification problem from speech signals with six kinds of emotions: anger, disgust, fear, joy, sadness and surprise. By utilizing short-time log frequency power coefficients (LFPC) to represent the speech signals and using a discrete hidden Markov model (HMM) [19] as a classifier, their proposed method could achieve an average accuracy about 78% and 96% at the best case. Kun Han and colleagues [2] used deep neural networks for extracting high-level features from raw data and applied an extreme learning machine (ELM) [20] to predict emotion states from speech signals. The experiments on IEMOCAP dataset [3] showed that their work outperformed the state-of-the-art techniques.

Typically, one possible way for building a human emotion recognition system is using facial landmarks, which can be considered as key points in human faces. It is interesting to note that each emotion has unique facial expression, especially on eyes, eyebrows and mouth. Therefore, facial landmarks can be recognized as crucial characteristics to describe human emotions. Pantic and Rothkrantz [4] showed that in order to detect emotions from human images, there were usually three following steps: face detection, feature selection from face images and emotion classification from the extracted features.

In this paper, we present an efficient approach to construct a real-time emotion detection system using one camera. We consider three kinds of emotions: negative, blank and positive emotions. Negative emotion can be considered as any feeling that causes one person to be uncomfortable and sad. Meanwhile, positive emotion can be described as any feeling in which there is a lack of negativity such that no suffering or discomfort is felt. At last, blank emotion means "emotionless". These emotions are illustrated in Figure 1.

To solve the problem, we firstly detect any face appearing in each frame. With each detected face, we extract and normalize the corresponding facial landmarks. Next, we calculate a set of features as inputs for emotion classification later. Finally, we choose an appropriate classifier by analyzing several techniques including SVMs, decision trees and random forests [18]. The experiment shows that our proposed system can achieve a very good performance in speed and accuracy.

This paper is organized as follows. In Section 2, we describe how to design a real-time emotion detection system and show experimental results in section 3. In the last section, we give our conclusions and future work.

II. A proposed approach

To build a real-time emotion detection system, we utilize one camera of resolution 640x480 to capture human faces. There are four modules in our proposed system. For each frame, we firstly detect all human faces and extract facial landmarks from each observed face. Then, we normalize facial landmarks and calculate the corresponding features. Finally, the computed features are used as inputs of a trained classifier to predict emotion for each person. The proposed method can be illustrated in Figure 2. In what follows, we will explain step by step what we do in each module.

978-1-5090-5401-5/17/$31.00

(a) negative (b) blank (c) positive

Fig. 1. Three types of human emotions: negative, blank and positive emotions.

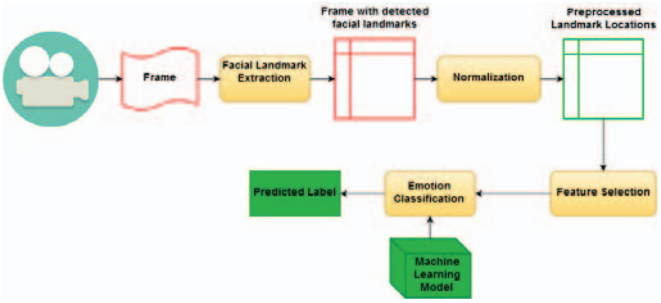


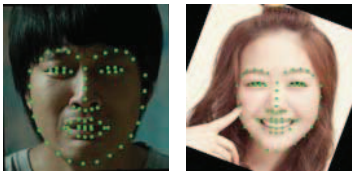
Fig. 2. There are four major steps (yellow) in the proposed system: facial-landmark extraction, normalization, feature selection and emotion classification.

Fig. 3. 68 chosen facial key-points in a given face.

A. Facial-landmark extraction

For detecting facial landmarks in face images, one can apply two traditional approaches, namely regression-based approach and template fitting approach. For the regression-based method, one iteratively refines the first initialization of landmark locations and then use image features to predict them explicitly by regression [8]. Meanwhile, for the template fitting approach, one does not need any initial guess of facial landmarks but establishes face templates to fit input images [21].

It seems to be true that the more landmark locations one can find in face images, the better emotion prediction one can achieve. In addition, the performance of facial landmark extraction is one of the most important requirements for building a real-time emotion detection system. In this work, we choose an approach of Kazemi and Sullivan that can extract 68 key points in human faces effectively [8].



(a) Sadness (b) Happiness

Fig. 4. Geometrical structure of facial landmarks

Figure 3 depicts the exact locations of facial landmarks. They can be grouped into six parts of the human face: eyes, eye-brows, eyelids, nose, lips and jaw. Reasonably, each emotion has a unique geometrical structure of landmark points. For instance, when people smile, all their muscles at the sides of the mouth are flexed and the mouth seems to be opened at that time. While when people are shameful, their mouths are mostly neutral or frowning. Consequently, we can learn feature representation from the geometrical structure for emotion prediction.

B. Normalization

Since one person may have different head poses in camera, for enhancing the performance of the proposed system, we apply a normalization step before selecting features. First, we rotate each detected face such that the angle formed by the head and the horizontal axis of the image is always a right angle. That is, from 68 facial landmark locations, we calculate the angle between the vertical axis of the image and the line connected the 32^{th} and the 36^{th} landmark locations (as shown in Figure 3) and then do a proper rotation. Due to each detected face has various size, after rotation step, we rescale the face image into 100×100 pixels. At the end, we exploit the updated facial landmarks to compute geometrical features in the next section.

C. Feature selection

In this section, we aim at describing how to choose a learning model for emotion prediction. From detected facial landmarks in a given face, we search for fitting features and a

(a) Before (b) After

Fig. 5. Normalization step

suitable machine learning model to obtain the predicted label. In this paper, we only consider three labels, which are -1 for negative emotion, 0 for blank emotion and 1 for positive emotion.

There are several works related to the feature selection based on facial landmarks. Loconsole and co-workers [9] determined 19 key points in human face to compute 26 geometrical features for emotion prediction, including 10 eccentricity features, 3 linear features and 13 differential features. Palestra et al. [10] presented a potential approach for automatic facial expression recognition. From a detected face in human image, they extracted 20 facial landmarks and then used these successful detected landmarks to construct a set of 32 geometric facial features. These geometrical features are combined from linear, eccentricity, polygonal and slope features.

In this work, we extract 68 facial landmarks from a single face image. Not only taking the advantage of 26 geometrical features from [9], we introduce a set of new features for the classification problem.

We represent the first 17 key-points in the jaw by $F_1, F_2, \ldots, F_{17}$, the 49^{th} key-point by A_M, the 55^{th} key-point by B_M, the 52^{nd} and 63^{rd} key-points by U_{m1} and U_{m2}, and the 58^{th} and 67^{th} key-points by D_{m1} and D_{m2}.

For all key-points with indexes in the set:

$$I = \{50, 51, 53, 54, 56, 57, 59, 60, 61, 62, 64, 65, 66, 68\},$$

we denote them by M_i, where i is the corresponding index. From these notations, we propose a new set of geometrical features as following:

(a) Nine linear features which are the normalized distances from each landmark F_i $(i = 1, \ldots, 9)$ to the point A_M:

$$d_i = \frac{|F_i A_M|}{|A_M B_M|} \tag{1}$$

(b) Nine trigonometric features that are the angles between each vector $\overrightarrow{F_i A_M}$ $(i = 1, \ldots, 9)$ and the vector $\overrightarrow{A_M B_M}$ where:

$$\cos(\overrightarrow{F_i A_M}, \overrightarrow{A_M B_M}) = \frac{\left|\overrightarrow{F_i A_M}.\overrightarrow{A_M B_M}\right|}{\left|\overrightarrow{F_i A_M}\right| \left|\overrightarrow{A_M B_M}\right|} \tag{2}$$

(c) Nine linear features which are the normalized distances from each landmark F_i $(i = 9, \ldots, 17)$ to the point B_M:

$$e_i = \frac{|F_i B_M|}{|A_M B_M|} \tag{3}$$

(d) Nine trigonometric features that are the angles between each vector $\overrightarrow{F_i B_M}$ $(i = 9, \ldots, 17)$ and the vector $\overrightarrow{B_M A_M}$ such that

$$\cos(\overrightarrow{F_i B_M}, \overrightarrow{B_M A_M}) = \frac{\left|\overrightarrow{F_i B_M}.\overrightarrow{B_M A_M}\right|}{\left|\overrightarrow{F_i B_M}\right| \left|\overrightarrow{B_M A_M}\right|} \tag{4}$$

(e) Let C be the intersection between two lines $A_M B_M$ and $U_{M1} D_{M_1}$. We define 20 following linear features which are the normalized distances from each landmark around the mouth M_i $(i \in \{49, 50, \ldots, 67, 68\})$ to the point C by:

$$f_i = \frac{|M_i C|}{|A_M B_M|}, \tag{5}$$

where $M_{49} \equiv A_M$, $M_{55} \equiv B_M$, $M_{52} \equiv U_{m1}$, $M_{63} \equiv U_{m2}$, $M_{58} \equiv D_{m1}$ and $M_{67} \equiv D_{m2}$.

(f) Next, we use 20 trigonometric features that are the angles between each vector $\overrightarrow{M_i C}$ $(i = 49, \ldots, 68)$ and the vector $\overrightarrow{B_M A_M}$ satisfying that:

$$\cos(\overrightarrow{M_i C}, \overrightarrow{B_M A_M}) = \frac{\left|\overrightarrow{M_i C}.\overrightarrow{B_M A_M}\right|}{\left|\overrightarrow{M_i C}\right| \left|\overrightarrow{B_M A_M}\right|} \tag{6}$$

(g) Finally, we choose three features which are related to the openness of mouth and two eyes:

$$m_1 = \frac{|U_{m1} D_{m1}|}{|A_M B_M|}, \tag{7}$$

$$o_L = \frac{|EL_{38} EL_{42}|}{|EL_{37} EL_{40}|} \tag{8}$$

$$o_R = \frac{|ER_{44} ER_{48}|}{|ER_{43} ER_{46}|} \tag{9}$$

in which $EL_{38}, EL_{42}, EL_{37}, EL_{40}$ are four key-points in the left eye (the 38^{th}, 42^{nd}, 37^{th} and 40^{th} ones) and $ER_{44}, ER_{48}, ER_{43}, ER_{46}$ are four key-points in the right eye (the 44^{th}, 48^{th}, 43^{rd} and 46^{th} ones).

Using both 79 new features and 26 geometrical features from [9], we have a list of 105 features as inputs for the emotion classification module, which is described in the next section.

D. Emotion classification

From calculated features, we need to find an appropriate model for emotion classification. we compare three different supervised classification methods: multi-class Support Vector Machine (multi-SVM), decision tree and random forest. To evaluate these methods, we use 5-fold cross-validation and the results are given in Section III.

III. Experiments

In this section, we describe our experiments and the performance of the proposed system.

TABLE I
THE PERFORMANCE OF THREE LEARNING MODELS: MULTI-SVM, DECISION TREE AND RANDOM FOREST ON TWO SETS $S1$ AND S_2.

Model	Decision Tree	Random Forest	Multi-class SVM (Linear Kernel)	Multi-class SVM (RBF Kernel)
Accuracy (S_1)	0.65	0.715	0.7513	0.7714
Accuracy (S_2)	0.70	0.7681	0.7983	0.8185

TABLE II
THE NUMBER OF SAMPLES FOR EACH EMOTION IN THE DATASET

Total	Negative	Blank	Positive
1079	309	359	411

TABLE III
THE ACCURACY FOR FIVE RANDOM TESTS ON THE REAL-TIME EMOTION DETECTION SYSTEM

Test	#1	#2	#3	#4	#5	Average
Accuracy	0.6726	0.9080	0.9526	0.306	0.6933	0.7065

Fig. 6. The emotion prediction system.

A. Training model

We use a dataset about 1079 images with three kinds of emotions: negative, blank, and positive emotions. These images are collected by our volunteers. Each volunteer is requested to stand in front of a camera and shows his/her personal emotion in a period of time. The number of images for each type of emotion can be given in Table II. In the experiments, we evaluate the performance of each approach by 5-fold cross-validation.

We then compare five cases: decision tree, random forest, multi-class SVM with two kernels (linear and RBF kernels). We analyze each approach by two sets of features: S_1 (the set of 26 geometrical features used in [9]) and S_2 (the set of 105 complete features). The accuracy of our experiments is shown in Table I.

In each dataset, one can see that the multi-class SVM with RBF kernel outperforms three other approaches where the corresponding accuracy is 0.7714 in S_1 or 0.8185 in S_2. In addition, each method using the set S_2 gains higher accuracy than another. It implies that the set of new features can achieve much better than the set of features proposed by Loconsole et al. For this reason, we opt for the multi-class SVM with RBF kernel as the final model used in the real-time emotion detection system.

B. Real-time emotion detection

We intend to design a system that is able to detect human emotion in real time. Using one camera, we detect all possible faces and then calculate the corresponding feature for each detected face. These computed features are later used as inputs for a trained model to obtain the predicted labels. Finally, the predicted emotion for each detected face can be displayed by emotional icons in the output (as displayed in Figure 6).

In real-time applications, it it important to understand how many frames per second they can be executed. In our experiment, we use one camera with the resolution 640×480 for testing and the number of executed frame rates in our proposed system is measured around 10.4 frames per second (fps) on average.

One can see that there is no much difference among three consecutive frames for human emotion. That means, in practice, for a given person, we do not need to detect his/her emotion frame by frame. For this reason, the proposed system only needs to handle about 10 frames (not 30 frames) per second to ensure that human emotion can be tracked naturally as well.

To evaluate the accuracy of our proposed system in real time, we ask five volunteers to show all possible emotions (negative, blank an positive emotions) in front of the camera during 20 seconds. After that, we compare the predicted emotions in each frame and the ground truth data. The accuracy among these tests can be illustrated in Table III. The average accuracy is about 70.65%. This is a very encourage result for a real-time emotion detection using camera.

IV. CONCLUSIONS AND FUTURE WORK

We have presented a potential approach for human emotion detection using facial landmarks and a camera in real time. Using facial landmarks extracted from detected faces, we have proposed a set of new features as representation data

for human emotion and analyze different learning techniques for emotion classification. The experimental results show that the multi-class SVM with RBF kernel can outperform other methods and is chosen as the final model of the final system. The proposed system can detect human emotion naturally with the average accuracy about 70.65%.

It is important to note that we only consider three types of emotions including negative, blank and positive emotions. In future works, we will generalize our approach for other kinds of emotions.

ACKNOWLEDGMENT

We would like to thank University of Science and University of Information Technology for the support during this project.

REFERENCES

[1] Robert Horlings, Dragos Datcu and Leon J. M. Rothkrantz, "Emotion recognition using brain activity", *CompSysTech*, 2008.

[2] Han Kun, Dong Yu and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine", *Interspeech*, 2014.

[3] Busso Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database", *Language resources and evaluation*, Vol. 42, 4, pp. 335-359, 2008.

[4] Pantic Maja and Leon J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art", *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 22, 12, pp. 1424-1445, 2000.

[5] Tin Lay Nwe, Say Wei Foo and Liyanage C. De Silva, "Speech emotion recognition using hidden Markov models", *Speech communication*, Vol. 41, 4, 603- 623, 2003.

[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*, Vol. 89, 4, pp. 344-350, 2001.

[7] Michel Philipp and Rana El Kaliouby, "Real time facial expression recognition in video using support vector machines", *Proceedings of the 5th ACM international conference on Multimodal interfaces*, pp. 258-264, 2003.

[8] Vahid Kazemi and Josephine Sullivan, "One millisecond face alignment with an ensemble of regression trees", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867-1874, 2014.

[9] C. Loconsole, C. R. Miranda, G. Augusto, A. Frisoli and V. Orvalho, "Real-time emotion recognition novel method for geometrical facial features extraction", *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP)*, pp. 378-385, 2014.

[10] Giuseppe Palestra , Adriana Pettinicchio, Marco Del Coco, Pierluigi Carcagn, Marco Leo and Cosimo Distante, "Improved Performance in Facial Expression Recognition Using 32 Geometric Features", *Proceedings of the 18th International Conference on Image Analysis and Processing (ICIAP)*, pp 518-528, 2015.

[11] Liu Yisi, Sourina Olga and Nguyen Minh Khoa, "Real-Time EEG-Based Emotion Recognition and Its Applications", *Transactions on Computational Science XII: Special Issue on Cyberworlds*, pp. 256–277, 2011.

[12] Stickel Christian, Ebner Martin, Steinbach-Nordmann Silke, Searle Gig and Holzinger Andreas, "Emotion Detection: Application of the Valence Arousal Space for Rapid Biological Usability Testing to Enhance Universal Access", *Lecture Notes in Computer Science*, Vol. 5614, pp. 615-624, 2009.

[13] P. Ekman and H. Oster, "Facial expressions of emotion", *Annual Review of Psychology*, 30, pp. 527-554, 1979.

[14] D. Galati , R. Miceli and B. Sini, "Judging and coding facial expression of emotions in congenitally blind children", *International Journal of Behavioral Development*, Vol. 25, 3, pp. 268-278, 2001.

[15] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: A meta-analysis", *Psychological Bulletin*, Vol. 128, 2, pp .205-235, 2002 .

[16] A. R. Daros, K. K. Zakzanis and A. C. Ruocco, "Facial emotion recognition in borderline personality disorder", *Psychological Medicine*, 43, 9, pp. 1953-1963, 2013.

[17] Xiang-Sun Zhang, "Introduction to Artificial Neural Network", *Neural Networks in Optimization*, pp. 83-93, 2000.

[18] Trevor Hastie, Robert Tibshirani and Jerome Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", *Springer Series in Statistics*, 2009.

[19] L. Rabiner and B. Juang, "An introduction to hidden Markov models", *IEEE ASSP Magazine*, Vol.3, 1, pp. 4-16, 1986.

[20] Guang-Bin Huang, Qin-Yu Zhu and Chee-Kheong Siew, "Extreme learning machine: theory and applications", *Neurocomputing*, Vol. 70, 1, pp. 489-501, 2006.

[21] Zhanpeng Zhang, Ping Luo, Chen Change Loy and Xiaoou Tang, "Facial Landmark Detection by Deep Multi-task Learning", *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 94-108, 2014.