2005－AVM－48（6）
2005／3／11

Automatic Smile Detection System

Uwe KOWALIK† Terumasa AOKI‡ Hiroshi YASUDA‡

† Research Center of Advanced Science and Technology

The University of Tokyo 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8904 Japan

E-mail: ‡{uwe,aoki, yasuda}@mpeg.rcast.u-tokyo.ac.jp

Abstract Facial Expressions are playing an important role in human communication. Intentional or not, they betray one's emotional state. It is challenging to make a computer understand a person's emotional condition, since it could serve the goal to provide natural ways of human-machine interaction. We propose a system that can detect the facial expression of 'smile' automatically from video by applying a neural network classifier to a set of feature points provided by a face tracker. We believe, that detecting a smile automatically by a computer system opens a door for various kinds of applications and nevertheless it will be a step towards a natural way of human-machine interaction, since the machine may understand whether we enjoy the communication or not.

Keywords Human Machine Interface, Face Recognition, Facial Expression, Affective Computing

1. Introduction

In the last decades much research has been done to discover the secrets of facial expressions in human communication. Spontaneous or not, facial expressions emphasize the point of speech or betray the emotional state of an individual non-intentionally. In [1] Ekman has proved that facial expressions are universal amongst different cultures. No matter to which countries people we will talk, we will always recognize a smile as a smile and an angry face always emphasizes non-agreement (this does not imply that we also understand the reason). Based on there previous work P.Ekman and W. Friesen developed the 'Facial Action Coding Scheme' (FACS) that provides a unambiguous way of categorizing facial expressions based on 'Action Units' (AUs). The AUs are related to the face muscles. The FACS system distinguishes 44 different basic action units. All possible (visually perceptible) facial expressions can be encoded by combinations of those basic AUs in connection with a related activity measure. By using FACS the analysis of facial expressions will be separated into the two steps i.e. description and interpretation. In [4] a data base for interpreting FACS-encoded facial expressions is provided. Today's technology and recent advance in image processing and pattern recognition provide means for automatic face detection and facial expression classification.

Many different approaches and combinations of algorithms have been proposed for the purpose of face detection and face feature detection. Color based approaches try to detect the face by skin color [6]. The approach of [7] directly locates eyes, mouth, and face boundary based on measurements derived from the color-space components of an image. [8] uses an invariant feature based approach exploiting the geometrical structure of a face to detect facial feature points in gray level images by graph matching. A combination of feature-based and image-vector-based approach is presented in [9] combining Higher-order Local Auto-Correlation features and fisher weight maps to detect facial expressions of smile.

In [10] we introduced our system approach for a TV-Program Rating System using facial expressions. The goal of this system is to parameterize emotions derived from facial expressions of the TV-audience in order to achieve a direct feedback of satisfaction level compared to today's indirect statistical methods. In this paper we propose a part of this system that will provide a measure for deriving the emotion 'happiness' from the facial expression of a 'smile'. Smile and laughter carry important information for measuring the satisfaction level of TV-audience, since they point out the grade of enjoyment a person feels. The current system recognizes the presence and strength of a smile by classifying a set of face features provided by a feature tracker module exploiting a feed-forward neural network. An important constraint for developing the smile-detection system was the processing time. The module will be a part of a complex analysis system for facial expressions and should perform well while consuming not too much computational power. In this paper we will present the structure of our low-complexity smile-detection classifier. We will show its capability of providing sufficient classification results with little computational effort.

2. System Overview

In Figure 1 the – compared to [10]– slightly modified system structure of the outline of the facial expression analyzer system can be seen.

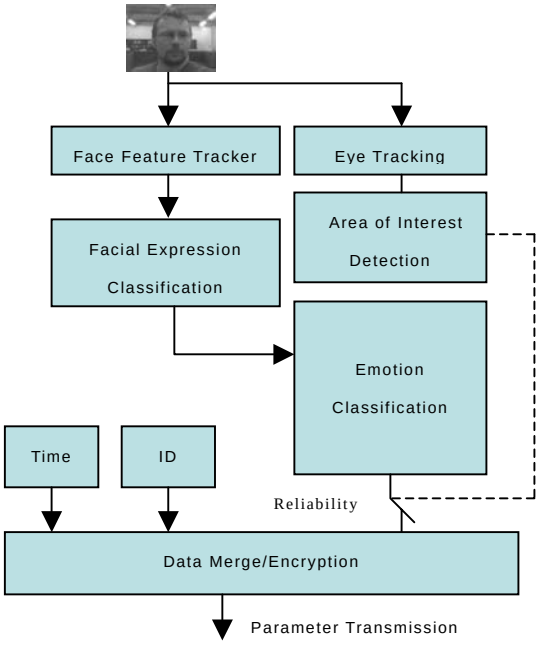

Figure 1 Facial Expression Analyzer Outline

Video images taken from a video camera are fed in to a 'Face Detection/Face Feature Tracker' module (FD/FFT). The 'FD/FFT' module detects the face by searching deformation invariant features in a gray level image derived from the current video frame. The output is a set of 22 feature points. Once the face and the features have been detected, they will be tracked over subsequent frames. In the case where the reliability measure of the tracked feature points falls below a certain threshold, the face/facial feature detection step will be automatically repeated. Figure 2 visualizes the output of the 'FD/FFT' module. The 'FD/FFT' module output builds a vector of 44 elements consisting of the x- and y-components of the features. In addition to the facial features, we obtain the global position of the face' projection in the image

plane determined by P(x, y) with 0<=x, y<=1.0 independent of the image size.

Figure 2 The system tracks 22 features

A scale factor z with z ~ d, where d is the distance of the face to the camera plane, as well as the Euler-angles for the rotation of the face around the x-, y- and z-axis are provided too. The feature vector represents low-level information of local points in a human face. Taking into account the physical correlation between the feature points (connected through skin, driven by muscle contraction), always a set of feature points is involved to form a specific facial expression. The 'Facial Expression Classification' module (FEC) exploits this correlation to perform facial expression classification. The last step in the processing chain of the in [10] proposed system is the classification of emotions. The idea is to use the fact, that a certain emotional state will be physically represented by an appropriate facial expression. Therefore the facial expression analyzer builds a hierarchical chain of from low-level to high-level classifiers to achieve emotion classification based on facial expressions.

3. Facial Expression Classification

In Figure 3 a more detailed block diagram of the facial expression classification module (FE) can be seen. The feature vector (output of 'FD/FFT' module) has to be preprocessed appropriately to the requirements of the following classifier. This may imply reduction of dimension, transform, etc. A FE-classifier has always one specialization (e.g. smile) and its output is an intensity measure of the occurrence of this facial expression. One or more feature points might be involved in building a certain expression.

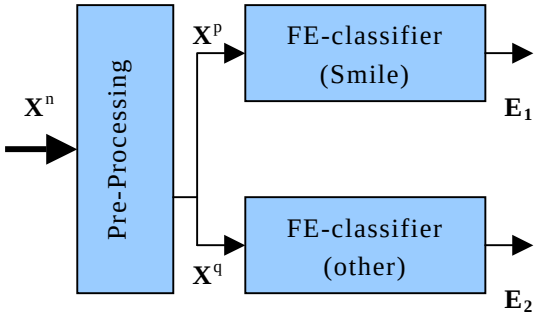

Figure 3 Block diagram of FEC module

3.1. 'Smile'-Classifier

A subset of eight feature points is involved in our approach of smile-detection. The pre-processing step performs a selection of the eight features defining the mouth shape (Figure 4).

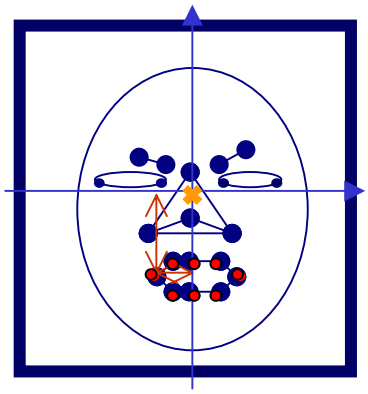

Figure 4 Mouth features (red)

A compensation of head translation, rotation and scale (z-component) is performed before applying a neural network classifier to the mouth feature set. The resulting feature vector is 16-dimensional.

3.2. Neural Network Classifier

For classification of the mouth shape represented by the 16-dimensional input vector a feed-forward network with three layers has been designed. The input layer consists of sixteen neurons to fit the feature vector's dimension. A hidden layer with two neurons and an output layer with one neuron complete the network. A training set of 268

samples has been used for training. 50% of the input pattern have been labeled as ‘smile’ and 50 % have been labeled as ‘no smile. The feature patterns for the mouth shape description have been extracted semi-automatically from video sequences taken from three persons posing ‘smile’ and ‘no smile’ expressions. The whole video database consists of sequences of six people. Each sequence contains about one minute of video. A set of three sequences has been used to extract the data for ‘untrained’ patterns to verify the performance. A helper application has been implemented to support the generation of pattern from video sequences (Figure 5).

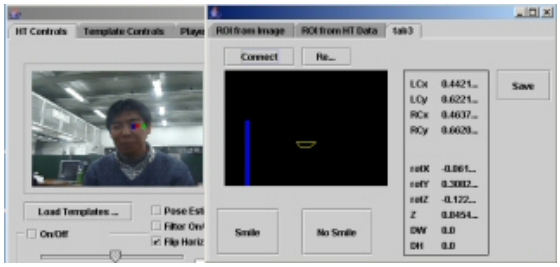

Figure 5 Pattern generation application

During the pattern generation step the video of facial expressions is presented to the test person. In the case the person recognizes a smile in the video sequence, he/she clicks a button for smile. The same procedure has been applied to ‘no smile’ expressions. Doing so ensures a natural judgment by a human observer. In Figure 6 sample images taken from our video database are shown. The upper row images were used to train the network. The lower row images have been used for verification.

Figure 6 Sample images of our video database

For training the network we used the ‘Resilient Propagation’ method (Rprop). ‘Rprop’ adapts its learning process to the topology of the error function. The weight-update and weight-adaptation are performed after the gradient information of the whole pattern set has been computed. For the weight-update only the sign of the gradient will be considered to determine the direction of update. A detailed description can be found in [11]. The following parameters have been set :

- initial update value $\Delta_0 = 0.1$
- maximum weight step $\Delta_{max} = 50.0$
- weight-decay exponent $\alpha = 4.0$

Figure 7 shows the resulting error graph for 5000 cycles.

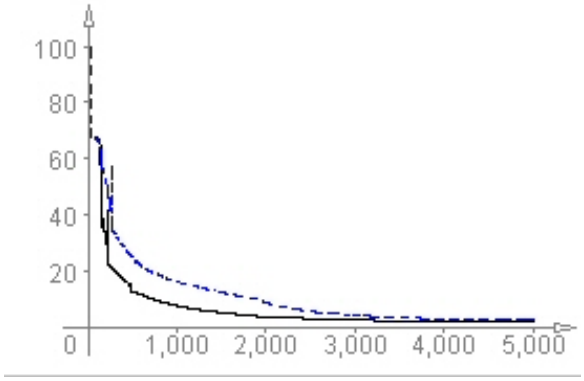

Figure 7 Resulting error curve during training the NN over 5000 cycles

The dotted line is the resulting error curve for presenting the patterns in a fixed order to the network. The closed line shows the resulting error for presenting the patterns randomly. Random presentation leads to a faster convergence of the system.

4. Experiments and Results

After training the network classifier the video sequences of previously unknown persons have been presented to the system. The processing time for classifying the extracted mouth features into smile/non-smile patterns has been measured and takes 3 ms in average

per video frame. In Figure 8 sample pictures taken during the verification process can be seen. All persons are not included in the training set. The intensity of the smile recognized by the system is represented by the blue bar in the right images. The yellow polygon is the shape of the mouth formed by the tracked features.

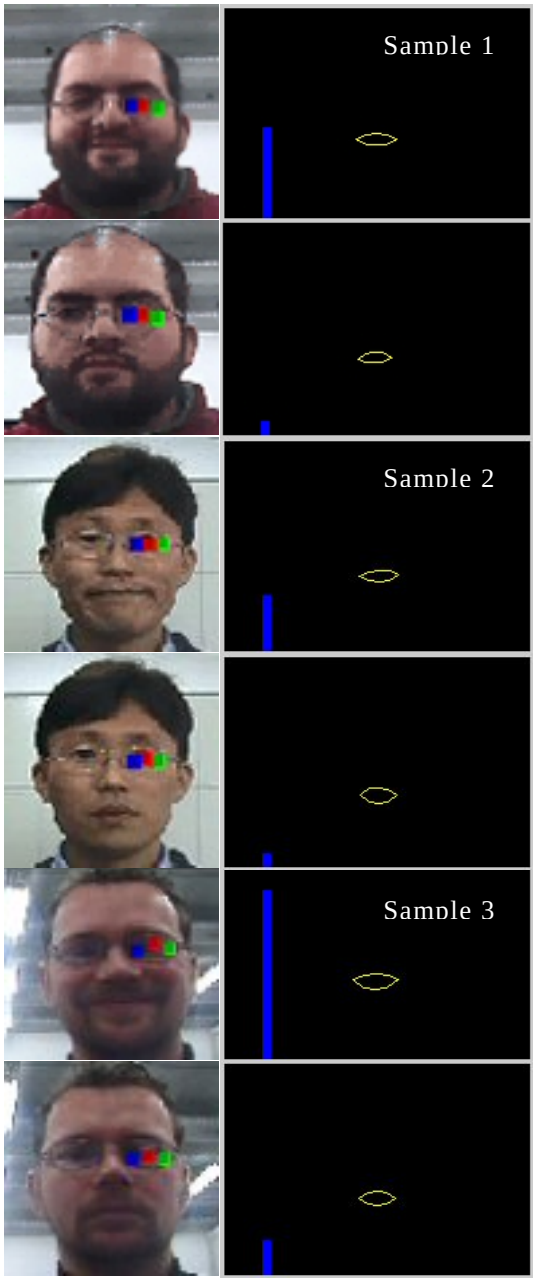

Figure 8 Resulting smile intensities perceived by the system

The smile intensity is scaled between 0-100% of the average maximum smile intensity learned by the system. Since the neural network tries generalizes the smile patterns there is always a minimum smile intensity depending on the mouth shape perceived. This minimum is specific for each person and can be seen as a individual measure. The same applies to the maximum value. In the case of sample 2 a human would probably not recognize it as a real smile as the system does. This is a hint, that actually more facial features are involved in forming a smile expression than used in our system.

5. Conclusion and Future Work

In this paper we have presented our approach for an automatic smile-detection system, that can measure the intensity level of smile in presented images of human faces. An feed-forward ANN has been trained with the 'Rprop' learning algorithm to classify locations of mouth shape features regarding the presence of smile expression. Classification errors can be further minimized by taking more facial features into account (e.g. eyes, texture). However, the classification has a very low computational complexity and is therefore useful for real-time applications while providing a reasonable classification result. Based on the results presented in this paper we will investigate the incorporation of above mentioned facial features in our future work. Also the extension of our video database for training the network is an open issue to achieve a better generalization of the classification result.

References

[1] P.Ekman and W.V. Friesen, Unmasking the Face. New Jersey: Prentice Hall, 1975

[2] P.Ekman, Emotion in the Human Face. Cambridge Univ. Press, 1982

[3] P.Ekman, Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life
[4] Facial Action Coding System Affect Interpretation Dictionary (FACSAID), http://face-and-emotion.com
[5] P.Ekman and W.V.Friesen, Facial Action Coding System (FACS): Manual. Palo Alto: Consulting Psychologists Press, 1978
[6] S.Singh, D.S. Chauhan, M.Vatsa, R.Singh, A Robust Skin Color Based Face Detection Algorithm, Tamkang Journal of Science and Engineering, Vol.6, No.4, pp227-234, 2003
[7] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, ``Face detection in color images,'' IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 696-706, May 2002
[8] J.Buhmann, J.Lange, C.von der Mahlsburg, Distortion invariant Object Recognition Matching Hierarchically Labeled Graphs, Proc. Int. Joint Conf. Neural Networks, pp. 155-159, 1989
[9] Y.Shinohara, N.Otsu, Facial Expression Recognition Using Fisher Weight Maps, Sixth IEEE International Conference on Automatic Face and Gesture Recognition, Soul, 2004
[10] U.Kowalik, T.Aoki, H.Yasuda: “A TV Rating System Using Facial Expressions”, FIT2004, Kyoto, 2004
[11] M Riedmiller. Untersuchungen zu Konvergenz und Generalisierungsverhalten überwachter Lernverfahren mit dem SNNS. Proceedings of the SNNS 1993 workshop, 1993