BROAFERENCE - A Next Generation Multimedia Terminal Providing Direct Feedback on Audience's Satisfaction Level

Uwe Kowalik, Terumasa Aoki, and Hiroshi Yasuda

Research Center of Advanced Science and Technology,
The University of Tokyo,
4-6-1 Komaba, Meguro-ku, Tokyo, 153-8904, Japan
{uwe, aoki, yasuda}@mpeg.rcast.u-tokyo.ac.jp

Abstract. In this paper we present BROAFERENCE, a test bed for studying future oriented multimedia services and applications in distributed environments. The BROAFERENCE system provides a method of direct feedback of audience satisfaction level by means of real-time smile-detection. We describe the real-time system components and we will show the result of our approach to record and transmit the data of smile intensity as a hint for 'how much the audience enjoys' the TV program.

1 Introduction

Definitely interesting for content and service providers is the feedback of the audience' satisfaction level and watching behavior. Tracking this information will be means for providing higher quality and customized contents to the audience in the future. Against this background we present a study of a new type of multimedia terminal called BROAFERENCE. The system is equipped with a module, which tracks the user's face in order to detect facial expressions of 'joy'. We are using the 'smile'-intensity as an indicator for the 'joy' emotion. Current tracking systems for TV audience' watching behavior provide only manual feedback of appreciation. The observed subjects have to give the feedback by themselves and the awareness about giving the feedback is required. This will influence the preciseness of the result and moreover in will effect the user's experience negatively. The advantage of our approach within the BROAFERENCE system is deriving the information of appreciation from the user's face. The system provides a direct feedback of a spontaneous emotion compared to current methods. Based on a vision based tracking SDK' provided by NVision we developed a real-time smile detection module, which continuously outputs a value proportional to the smile intensity while the user is watching TV. The smile intensity value is synchronized with the program time, recorded or transmitted and can be used as a hint how much the audience enjoyed the program. In the following first an overview of related work will be given. Than our implementation approach of BROAFERENCE will be introduced to the reader and described in detail.

2 Related Work

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 [1]. The approach of [2] directly locates eyes, mouth, and face boundary based on measurements derived from the color-space components of an image. [3] 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 [4] combining Higher-order Local Auto-Correlation features and fisher weight maps to detect facial expressions of smile. The BROAFERENCE system uses a commercial face detection module that provides face features. In this paper we focus on deriving a facial expression of smile describing intensity value in real-time under near-reality conditions (i.e. lighting conditions may change).

3 System Overview and User Feedback

The block diagram of the BROAFERENCE terminal and the feedback module can be seen in Fig. 1. Media streams can be received over a network connection. The composition of media streams and other data is described by contents description, rendered by a compositor and displayed at the screen. A camera takes the video of the person in front of the terminal and hands the images to the feedback analyzer that analyzes the expressions shown at the user's face in real-time. The face tracker module finds the user's face and provide the geometric information of specific points (e.g. mouth corners, eye center, etc.). Fig. 2 visualizes the output of the tracker module. The 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 $0 \leq x, y \leq 1.0$ independent of the image size. A scale factor z with $z \sim d$, where d is the distance of the face to the camera plane. The Euler-angles for the rotation of the face around the x-, y- and z-axis are provided too. The smile detection module provides an intensity value of smile calculated from the tracked mouth points. The smile intensity value is in the range between 0.0 and 1.0, whereas a value of 0.0 means 'no smile' and a value of 1.0 is related to 'big smile' expression. A compensation of head translation, rotation and scale is performed before applying a neural network classifier to the mouth feature set. The resulting feature vector is 16-dimensional. 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 training patterns 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. 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.

Fig. 1. Block diagram of BROAFERENCE terminal and feedback system

Fig. 2. The system tracks 22 features, 8 mouth features are used

In Fig. 3 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. For training the network we used the ‘Resilient Propagation’ method (Rprop). A detailed description of the algorithm can be found in 5 .

Fig. 3. Sample images of our video database

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 has been measured and takes 3 ms in average per video frame. The performance of the used tracker is satisfying also under darker lighting conditions, but it heavily depends on the used camera model.

Fig. 4. Smile intensity values

In Fig. 4 a sample pictures taken during the verification process can be seen. The intensity of the smile recognized by the system is represented by the colored bar in the right images. The polygon is the shape of the mouth formed by the tracked features.

5 Summary

In this paper we introduced the new feature for providing direct feedback of audience' satisfaction level by automatically recognizing the facial expression of smile in order to provide a hint of how much the audience enjoyed the program. The future work will focus on facial expressions taking other emotions like boredom or sadness into account. We believe that our approach of direct feedback will help to improve the quality of future TV programs and multimedia presentations which will finally improve the user experience and contribute to a richer life by exploiting latest technology advances.

References

1. S.Singh, D.S. Chauhan, A Robust Skin Color Based Face Detection Algorithm, Tamkang Journal of Science and Engineering, Vol.6, No.4, pp227-234, 2003
2. R.-L. Hsu, M. Abdel-Mottaleb, A. K. Jain, Face detection in color images, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 5, May 2002
3. J.Buhmann, J.Lange, C.von der Mahlsburg, Distortion invariant Object Recognition Matching Hierarchically Labeled Graphs, Proc. Int. Joint Conf. Neural Networks, 1989
4. Y.Shinohara, N.Otsu, Facial Expression Recognition Using Fisher Weight Maps, Sixth IEEE International Conference on Automatic Face and Gesture Recognition, Soul, 2004
5. M Riedmiller. Untersuchungen zu Konvergenz und Generalisierungsverhalten überwachter Lernverfahren mit dem SNNS. Proc. of the SNNS workshop, 1993