Using Automatic Facial Expression Classification for Contents Indexing Based on the Emotional Component

Uwe Kowalik, Terumasa Aoki, Hiroshi Yasuda

Research Center of Advanced Science and Technology
The University of Tokyo,
4-6-1 Komaba Bldg.#3, Meguro-ku Tokyo, 153-8904 Japan
{uwe, aoki, yasuda }@mpeg.rcast.u-tokyo.ac.jp

Abstract. Within the last decade the development of new technologies in the multimedia sector has advanced with stunning pace. Due to the availability of high-capacity mass storage devices at low cost private multimedia libraries containing digital video and audio items have recently gained popularity. Although attached meta-data like title, actor's/actress' name and creation time eases the task of finding preferred contents, it is still difficult to find a specific part within a movie one enjoyed before by remembering the time code. In this paper we introduce the BROAFERENCE system that provides a solution for the above problem. We propose meta-data creation based on recorded user experience derived from facial expressions containing joy, sadness and anger events as well as interest focus data. In the following the system layout, functionality and conducted experiments for system verification will be introduced to the reader.

Keywords: Facial expression, emotion, video retrieval

1. Introduction

Facial expressions are playing an important role in human communication. They convey information of emotional states of joy, anger, fear, disgust or surprise [1]. In [2] it is proposed, that this basic emotions have corresponding prototypic facial expressions. P. Ekman developed the 'Facial Action Coding System' (FACS). 'FACS' is a systematic approach of defining a general, atomic set of facial action units (AU) based on visually perceived muscle movements in human faces. Any possible facial expression can be described by a combination of multiple AUs [2] [3]. Although the process of FACS-coding is still a manual task for certified FACS-coders, several automatic approaches have been presented lately [4]. In front of this background we have developed the BROAFERENCE system. BROAFERENCE provides a platform for delivering interactive multimedia applications over IP based networks. One or more terminals may be connected to the system. Each terminal is equipped with a module for tracking, analyzing and recording facial expressions of the user in front of the terminal's screen. Our approach for indexing multi media contents is based on the

temporal correlation between emotional triggers caused by a screen event at a certain media time and the related facial expression. The expression can be automatically perceived in the user's face taken by a video camera connected to the BROAFERENCE terminal. The facial expression parameters will be synchronously stored along with the media time, while the user is watching e.g. a movie. This data form a set of meta-information that is associated with the contents. Due to the synchronization between media time and occurrence time of a certain facial action event it is possible to find a specific part of a movie later by looking for e.g. the highest intensity of a 'smile' expression in order to identify funny parts in the movie. A gaze detection module estimates the screen area visited by the user's eye. This information can be used to perform an analysis of user's focus of interest during the contents consumption. In the following we will first describe the system layout and functionality of the modules. In order to verify the systems performance we conducted user experiments with the specific goal to prove the reliability of the real-time smile detection module. The experimental setup, execution and results will be explained later. Finally we will summarize the paper and give an outlook to the future work.

2. Related Works

In [7] news video archives are indexed with a lexicon of 100 semantic concepts, e.g. sports, people, etc. Users may query by concepts or combinations of them and access the resulted video on a semantic level. In [8] video documents are automatically indexed with 17 specific concept detectors for speech and noise analysis [9]. Keywords are automatically derived from the speech recognition result. The keywords forming a 400 dimensional vector after a PCA has been applied. The keyword vector represents a conceptual description for each video document. In addition a low-level color histogram based indexing is performed on the key frames. Above works are using a contents based indexing approach. In [10] it is suggested to take the user behavior while watching a video from a private video data base into account. The user actions, e.g. *pause*, *stop*, *fast forward*, etc. are stored in a log-file for later examination. Different behavioral types are defined and a specific user's type is derived from the log-file of his/her watching behavior. Whereas the approaches of contents indexing presented in [7] [8] are contents oriented, the approach in [14] is more near to our approach i.e. user oriented. In this paper we present our idea of exploiting emotion data derived from facial expressions for video indexing in large private video data bases. With our approach a user can find specific contents based on how much oneself (or others) enjoyed the contents previously.

3. System Overview

In Fig. 1 the general network structure of the BROAFERENCE system is shown. A video server (broadcast station) sends the movie or TV program to an IP multicast address that is known to the clients (participants). Using IP broadcasting ensures that

the timestamps of the media stream packets received by a client terminal are consistent through all clients. The advantage by doing so lies in the fact, that the recorded meta-data sets for a certain transmission are comparable between different users later. The BROAFERENCE system also supports connections between different client terminals, which allow the establishment of individual video chat sessions while the user is watching the program.

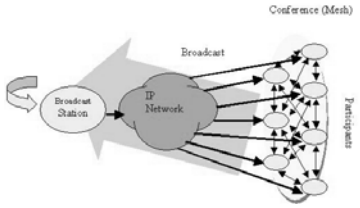

Fig. 1. Network structure of the BROAFERENCE system

Fig. 2 shows the block diagram of a terminal. The BROAFERENCE system facilitates a flexible scene composition paradigm. A compositor module is responsible for assembling the final screen presentation based on the given contents description. The description will be loaded at start up from either a local or remote location. The scene description defines a tree structure of 2D, 3D, AV and interactive objects. The description format is text-based and similar to the VRML97 markup language. This approach is widely used and provides a high level of flexibility to build multimedia applications.

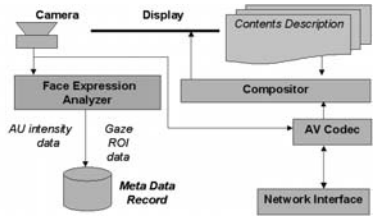

Fig. 2 Terminal block diagram

AV streams received/sent via the RTP network interface will be decoded/encoded by appropriate codecs. A face expression analyzer is connected to a local camera. This module automatically derives the intensity values of facial expressions from a set of tracked face feature points and stores them along with the media time of the received TV program building an index file that is used later for identifying parts of the contents that the user e.g. enjoyed.

3.1 Face Expression Analyzer

A commercial, vision based face tracker provided by 'NVision' is used to track the user's face and to extract relevant feature points. Video images taken from a video camera are fed into the face tracker module, which 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, i.e. a vector of 44 elements consisting of the x- and y-components of the features (Fig. 3).

Fig. 3 Face tracker output: 22 feature points

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. 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 as well. Always a sub-set of feature points is involved to form a specific facial expression.

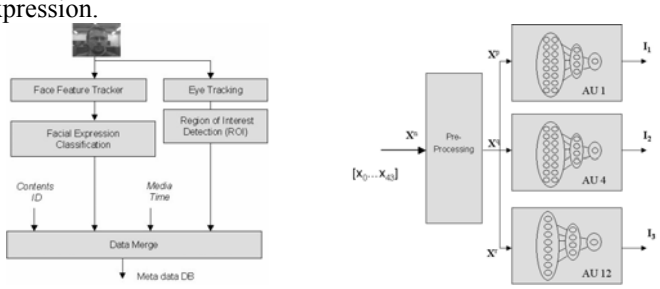

Fig. 4 Facial expression analyzer (left) and expression classifiers with preprocessing

The *facial expression classification module* (Fig. 4) exploits this correlation in order to perform the expression classification. The left part of Fig. 4 shows the block diagram of the facial expression analyzer. It consists of the *expression classifier module* and the *gaze detection module*. The *gaze detection module* will be described later. Currently tree classifiers for facial expressions have been implemented (AU1, AU12 and AU4). The intensity of the AU1 classifier output refers to the activity of the 'inner brow raiser' muscle. The activity of the 'zygmaticus major' muscle is measured by the intensity of the AU12 classifier output. The 'inner eye brow lowerer' muscle's contraction level is expressed by the AU4. The AU1 activity is related to the occurrence of sadness events experienced by a human. Activity of AU12 can be perceived when a person smiles. The 'inner eyebrow lowerer' muscle is incorporated in angry facial expressions. The feature vector will be preprocessed before passing it to the classifier. The preprocessing step reduces the dimension of the input vector from 44 parameters (all features) to the appropriate number of parameters used by the specific classifier. A subset of eight feature points is involved in our approach of smile-detection (AU12). The preprocessing step performs a selection of the features defining the mouth shape. For the classification of AU1 we are using the three features describing the position of nose root, left and right inner eyebrow corner. The AU4 intensity is calculated based on the same features as AU1, whereas the training data are different. A compensation of head translation, rotation and scale (z-component) is

performed for all data before applying a neural network classifier to the feature vector. Due to the real-time constraints of the system simple Artificial Neural Network structures (ANN) have been designed in order to reduce the processing time. The ANN consists of three layers. The input layer has a number of neurons that fits the number of elements of the feature vector. A hidden layer consisting of four neurons has been found to perform well. The output of each classifier is a single neuron that provides the intensity value for each classifier. For training the networks the 'Resilient Propagation' method (*Rprop*) was used. A detailed description can be found in [6].

Compensation of Head Movement and Rotation

The goal of the compensation step is to separate feature movements caused by head motion from those caused by facial expression changes. The exploited tracker provides head center position, scale and rotation (Euler angles around x-, y- and z-axis) parameters. The compensation is done in two steps. First, translation, scale and in-plane rotation around the z-axis are compensated by simply applying the inverse transform to the feature point positions. The result is still not stable against out-of-plane head rotations. The impact of out-of-plane head rotations on the feature positions can be approximated by Eq.1 provided the angles around x-axis and y-axis (α, β) are known and assuming parallel projection. Than an adaptive depth map estimation algorithm is applied to each feature in order to compensate displacement errors caused by a feature depth $Z_0>0$

$$X' = X \cdot \cos\beta + Z \cdot \sin\beta$$
$$Y' = Y \cdot \cos\alpha - Z \cdot \sin\alpha \qquad \text{Eq. 1}$$

The adaptive depth map estimation for compensating out-of-plane head rotations works as follows.

1. Capture a reference shape S_0 of feature points at $\alpha = \beta = 0$
2. Calculate expected X'_0, Y'_0 in subsequent frames assuming a rigid planar face model where Z=0 for each feature point and $\alpha, \beta > 0$ using Eq.1
3. Calculate the displacement $dx = X' - X'_0$, $dy = Y' - Y'_0$
4. Calculate and store a compensation term z via Eq.1 assuming dx and dy are caused by unknown $Z_0 \neq 0$
5. Calculate dz in subsequent frames until $dx, dy < \varepsilon$ where ε is the maximum allowed displacement error.
6. Finish z-estimation, if $dx, dy < \varepsilon$ for all features

Fig. 5 shows a result of the compensated feature position (x-component) over frame number for the nose tip (left) and a depth map example.

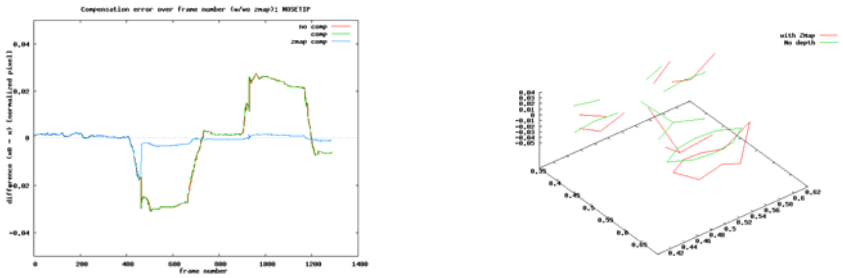

Fig. 5 Displacement error with/with out depth map estimation (left) and depth map example

3.2 Video Database for Classifier Training

A video database has been created for deriving the training data for the AU classifiers. It consists of assembled video sequences each about 2 minutes of length. Each sequence shows maximum (or minimum respectively) expressions taken from six different people. In order to get real (not posed) facial expressions, the people have been ask to perform different tasks while filming there face. Comedy strips have been presented to the persons for getting data for AU12 (smile intensity). The 'angry face' expression data (AU1) and data for AU4 have been extracted from a video taken while people were playing an online computer game. The final video data for extracting the training data set have been created by frame-based inspection and assembling of appropriate parts.

3.3 Gaze Estimation Module

The gaze estimation module was developed in order to gain information about the user's focus of interest while watching the TV program or movie. We define the focus of interest as the time that a user's gaze spends on a certain screen area. The gaze direction will be estimated based on the eye pose of the person in front of the screen. The gaze tracker incorporates a 3D head/eye model to simulate the physics of real head and eye movement (Fig. 6). The geometric parameters provided by the face tracker module are used to estimate the resulting view angle taking into account the connection of 3D transforms between head and eye rotation. While the head rotation parameters provided by the tracker module are used to transform the head model, the estimated eye rotation is applied to the 3D eye models. Fig. 7 shows a screen shot of the gaze estimation output.

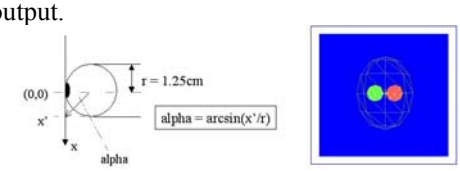

Fig. 6 3D eye model used for gaze estimation

Two cone geometries reaching from the eyes to the screen's image plate are simulating the gaze ray. The cone's tops are pointing into the area that is currently visited by the user's eye. The values for the angles of eye rotation are calculated separately, taking possible differences in left and right eye movement into account. By the approach presented here the BROAFERENCE system can automatically distinguish between nine different screen areas. An overlaid rectangle marks the region of interest (ROI) that has been decided based on the pointing cone tops.

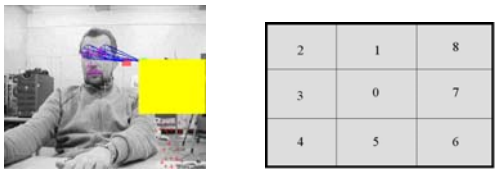

Fig. 7 Gaze estimation screenshot (left) and 'Region of Interest' areas on the screen derived from 2D coordinates of eye center feature points

3.4 Index Data File Format

A simple proprietary file format has been developed to store the intensity values of AU1, AU4, AU12 and the region code of the estimated region of interest along with the media time. The file format is packet oriented. Each packet contains the indexing information related to a specific media time stamp. The beginning of a packet is indicated by a packet start code followed by the media time. This allows fast and random access to the index data. The ROI code is stored as an 8 bit integer value. The intensity values, indicating the *joy*, *anger* or *sadness* factor, are stored as 32bit floating-point values subsequently. The packet structure would allow a streamed distribution of the meta-data, which could be interesting in the context of media research in a future ubiquitous network environment for e.g. collecting online user feedback on delivered multimedia contents.

4. Experiments and Discussion of the Results

In order to validate the function of the BROAFERENCE system approach for indexing of digital video contents delivered over IP network experiments have been carried out. We describe the experiments for all three classifiers in the following.

3.1 AU1 and AU4 Classifier

The performance tests for AU1 and AU4 have been carried out on video data that have been recorded for the classifier training, whereas this data where not part of the training set. A simple test application has been written in order to browse the created indexing data manually (Fig. 8)

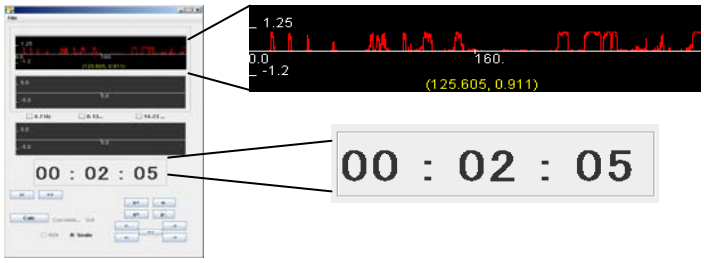

Fig. 8 Media time browser application

The black area displays the recorded intensity values along with the media time. The time of occurrence of a high intensity for AU1 and AU4 has been selected manually by pointing to the peak with the mouse pointer. The application translates the mouse position to the appropriate media time of the movie that shows the facial action of the test person. It has been found that the AU4 classifiers for anger detection performed quite well. We achieved 83% correct classification results for angry faces from the untrained video data base. In contrast the classification result of the AU1 classifier showed a lower reliability (high intensity, but actually no sadness event has recognized by a human observer). We achieved a correct classification of sad face expressions of 68%. This is due to the fact that the displacement of inner eyebrow feature is very small. It seems to be not sufficient to rely only on the inner eyebrow feature points for classification of AU1. An incorporation of texture properties of the area above the inner eyebrow features may lead to more stable results.

3.2 AU12 Classifier

For evaluation of the AU12 classifier performance participants have been asked to watch comedy scenes by using the BROAFERENCE system. A direct feedback method was used to evaluate the automatically derived intensity values of facial expression of smile against the true user experience. The evaluation setup is shown in Fig. 9.

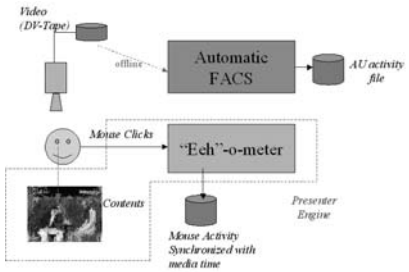

Fig. 9 Setup of the direct feedback evaluation method

A camera records the user's face while watching comedy scenes with the BROAFERENCE system. The recorded video data where used to derive intensity data of AU12 offline. Users have been asked to give feedback by clicking a mouse

button in the case they see some funny part of the content. The so called "Eeh-o-meter" module in Fig. 9 integrates the number of mouse clicks over a period of 300ms. Each click restarts the timer interval. The click time was synchronized with the media time of the presented content. The same contents have been shown to all subjects. Fig. 10 shows the resulting feedback timing charts of four different people (inter-person, left). Four different areas of main activity can be distinguished (dotted lines). Inside these areas the timing of the mouse feedback is almost synchronous between all subjects (closed lined arrows). This indicates that all persons experienced the same parts of the content as 'funny'. The right diagram of Fig. 10 shows an example result of the timing between manual feedback and AU12 classifier output for one subject (intra-person). It can be seen that the manual feedback starts after the smile-onset (onset=time from minimum to next first maximum of intensity) and lies completely inside the range of maximum occurrence of AU12. This interestingly applies for about 90% of the recorded data. The reason is that the subjects seem to be completely immersed into the contents during this time and do not control their environment. The onset-time has been measured and lies between 0.5 to 1.5 seconds in the recorded data. This value differs usually amongst people. Due to the above results, the AU12 intensity output was perfectly applicable to identify funny parts over the content timeline. The same application as the one used for AU1/AU4 verification was taken to map intensity peaks to media time.

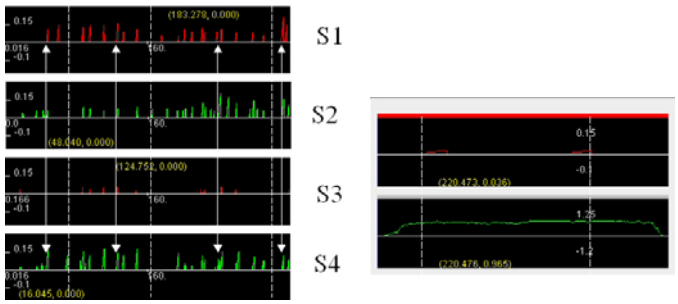

Fig. 10 Inter-person feedback timing (left) and intra-person AU12 intensity vs. feedback timing

4. Conclusion and Future Work

In this paper we have presented a new approach of indexing digital contents based on emotional events that normally occur while watching it. The idea is to exploit facial expressions since they are cues for 'how much' a person is riveted on the contents. The intensity of those facial expressions is automatically measured based on the position of tracked facial feature points. Currently three descriptors referring to the action units AU1, AU4 and AU12 in the Facial Action Coding System have been implemented. While AU12 is related to the smile intensity (*joy* emotion), the intensity of AU1 indicates the emotional state of *sadness*. AU4 indicates the experience of *anger*

events. Each descriptor's output is an intensity value accordingly to its AU. The intensity values are recorded during watching the contents and a search engine may access this data in order to extract media time stamps that are related to the user's previous *joy* or *sadness* experience. The results of conducted user experiments have shown, that automatically derived intensity values for AU12 expressions are reliable enough to serve as indexing hints for identifying contents based on the emotional component of *joy* experience. Future work will concentrate on the reliability improvement for the AU1 classifier. Further interest lies in analyzing dynamics of more complex facial expressions based on the here presented automatic approach.

References

1. P. Ekman, Facial Expressions, in T. Dalgleish, M. Power (eds.), Handbook of Cognition and Emotion, New York, John Willey&Sons Ltd., 1999
2. P. Ekman, Facial expression and emotion, AmericanPsychologist,48, 384-392
3. Paul Ekman, Wallace V. Friesen, and Joseph C. Hager: The new (2002) Facial Action Coding System (FACS)
4. Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joshua Susskind, Javier Movellan. "Dynamics of Facial Expression Extracted Automatically from Video," *cvprw*, p. 80, Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 5, 2004.
5. FACSAID, http://face-and-emotion.com
6. M Riedmiller. Untersuchungen zu Konvergenz und Generalisierungsverhalten überwachter Lernverfahren mit dem SNNS. Proceedings of the SNNS 1993 workshop, 1993
7. C.G.M. Snoek, M. Worring, J. Van Gemert, J.M. Geusebroek, D. Koelma, G.P. Nguyen, O. De Rooij, F. Seinstra, MediaMill: exploring news video archives based on learned semantics. Proc. of the 13th ACM international conference on Multimedia, Singapore, November 2005
8. M.Worring, G.P. Nguyen, L. Hollink, J.C. Gemert, D.C. Koelma, Accessing video archives using interactive search, Proceedings of IEEE International Conference on Multimedia and Expo, IEEE, Taiwan, June, 2004.
9. A. Hauptman, R.V. Baron, M.-Y. Chen, Informedia at TRECVID 2003 : Analyzing and Searching Broadcast News Video
10. S. Mongy, F. Boulali, C. Djeraba, Analyzing user's behavior on a video database, Proc. of Workshop on Multimedia Data Mining. Chicago, IL, USA, 2005