2004－AVM－47（6）
2004／12／9

Direct Feedback Method for Measuring the Level of Attention of TV Customers

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 Television plays an important in the area of information and entertainment of the masses. TV program makers are interested in a realistic feedback about the quality of their program. Today indirect methods of statistical investigations are used to retrieve an approximation of the user's satisfaction level. The indirection of today's methods leads to the fact that it is rather unclear, whether the test person really watches the program or not. The proposed system deals with this uncertainty by introducing a direct method for measuring the point of user's attention. The view direction is automatically detected by the system. This information can be recorded or continuously transmitted to the investigating authority for later inspection by an expert. We believe that this approach provides a key technology for statistical investigations about the customers watching behavior and it will help to improve the TV program quality.

Keyword Human Machine Interface, Eye Tracking, User Behavior

1. Introduction

The current technology provides means to focus on the 'who is watching when which channel' problem. A representative amount of households is provided with special equipment to measure statistical watching behavior by tracking channel switch events synchronously along with the program. A weak point of the current technology is the fact, that the level of attention that is actually paid to the program by the audience cannot be verified (e.g. people might leaving the room or doing something else than watching, while the program is still running).

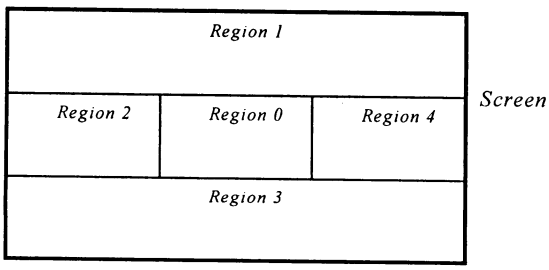

Figure 1 The screen is divided into five regions of interest

In this paper we introduce an implementation study for a gaze detection system (ROI tracker) that differentiates five regions of interest based on the user's eye pose. The screen area is divided into regions as can be seen in Figure 1. The analyses of the user's face as well as the pose classification are performed in real-time. The classification result can be locally stored or transmitted over a network to a server for later examination. An expert may use this information as an indirect measure of 'how interesting' a presented program is to the audience.

2. Related Work

In [1] a gaze tracking system has been introduced that is stable against head motion. The system facilitates a combination of vision based and infrared tracking system. The vision-based head tracking system compensates head motion. After the compensation step the local eye movement is measured based on the reflection of infrared light from the retina. In [2] a communication system for ALS patients has been introduced that is able to differentiate between twelve screen regions based on eyeball movement. Whereas that system provides a higher spatial resolution, head motion cannot be compensated.

The system we are presenting in this work is fully vision based and also stable against head translations. In the typical use-case of our system, a person would sit in front of a TV screen. This implies that the movement of the head is restricted. Although head

motion is not a restriction of the approach presented in this paper, it improves the stability of the tracking result. Besides a camera and a common PC no other additional expensive hardware is required. We believe that our approach could be realized at low cost with sufficient performance for collecting statistical data of people's TV watching behavior. In the following sections first an overview of the system layout and functioning is given. At next implementation issues will be pointed out and future steps and extensions will be discussed.

3. System Overview and Function

3.1. Block Diagram

In Figure 2 the block diagram of the realized ROI tracker can be seen. It consists of the following parts:

1. Video capture unit
2. Face Detection/Face Feature Tracking (FD/FFT)
3. Eye Refinement
4. Local Feature Detection
5. Eye pose classification

3.2. Function

A camera captures the face of the person in front of the screen with a resolution of 640x480 pixels at 30fps.

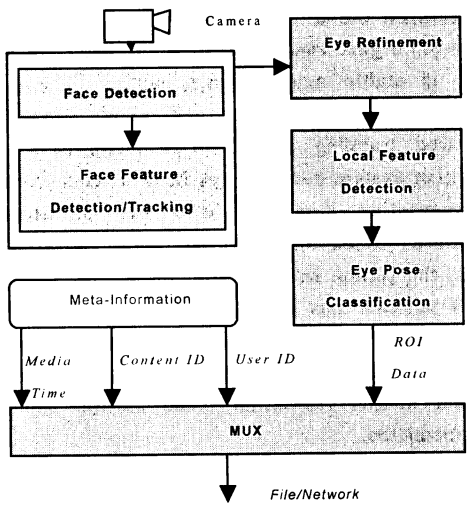

Figure 2 Block Diagram of the ROI Tracker

In a first step the video will be processed by the FD/FFT module at half resolution and full speed. The output of the FD/FFT unit is a set of twenty-two labeled feature points with reliability information attached. This data set contains the global head position information relative to the image coordinate system as well as the location of inner and outer corners for each eye. The positions of the two eye centers are also retrieved from this module (Figure 3). In this work a two-step approach has been realized. First the global head motion is compensated by the result of the FD/FFT module. The area of one eye will be extracted from the full resolution video frame (Figure 5). It has been found by experiments, that the examination of one eye is sufficient for the system's intention. Therefore only the image information of the right eye is further processed to reduce the calculation costs.

Figure 3 The output of the FD/FFT module is a set of 22 feature points extracted from the face

Due to the high stability of the FD/FFT tracking-result against head movement and the change of lighting conditions, the local image information of the eye area can be assumed as quasi stationary. For this reason a low-complexity algorithm can be used for the following local feature detection. We apply a simple algorithm that compares the features of an eye model for each eye pose against the presented image. Each eye model is related to one region of interest on the screen. The model has been previously created manually based on eye templates. For each eye pose out of five a set of four features is used. Each feature is labeled and describes the local image properties of a small block located around the pupil as shown in Figure 5. The spatial relationship between the feature blocks is fixed and given by the eye model. During the feature search process a small variance between the model positions and the found features in the current frame is allowed in order to perform local shape adoption.

If appropriate candidates have been found, their distance to the center point of the inner edge of the eye area is calculated (marked by the cross in Figure 5). This difference between the position of the feature cluster center and the fixed point is used to classify the current eye pose. In Figure 6 a selection of eye templates can be seen that was used for model creation. The eye pose classification compares the estimated feature values of the current frame against all model features associated with each eye pose. A similarity measure is calculated between each model and the current data set.

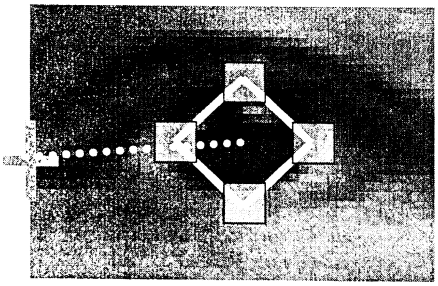

Figure 5 Features extracted to locate the pupil

The model with the feature set of the highest similarity to the current data set is chosen in a 'winner-takes-all' manner. Since each screen region out of five is associated with one model, an unambiguous decision can be made, which region has the current attention of the user.

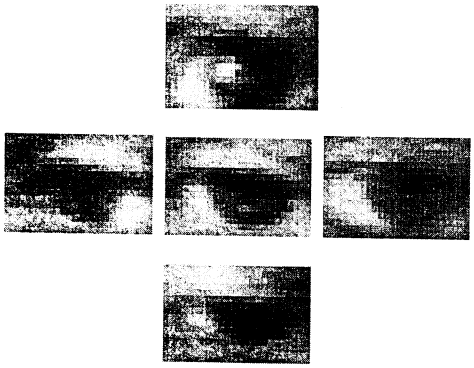

Figure 6 Selection of eye pose templates

The region information will be labeled with the current time stamp and user ID and send to a remote server as text stream. In Figure 4 the format of the transmitted TCP/IP packet is depicted.

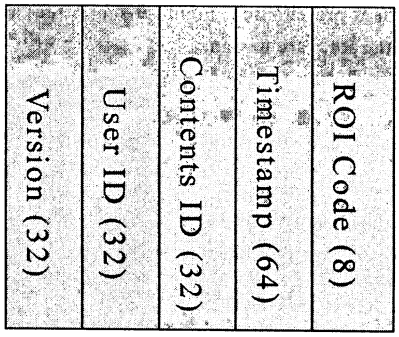

Figure 4 Layout of transmitted data packets (168 bits/sample)

The data starts with a version string followed by the user id, contents id and timestamp. The region id is represented by an integer between 0 and 4. A value of −1 indicates, that no eye pose model has been selected, since the similarity is smaller than a certain threshold. This happens usually, if the person does not look at the screen. In the current implementation the server simply writes a log file containing the received information. It could be also another application that makes use of the transmitted region information for other purposes than statistics.

4. Mock-Up and Implementation

Figure 7 shows the system configuration of the experimental system. It uses a PDP503CMX plasma display for content presentation. Since the eye-refinement step requires good image quality a SONY DCR VX2000 video camera is used to capture the face of the person in front of the screen. All image-processing modules are completely realized in software running on a PC with two Pentium 4 CPUs at a speed of 2.8 GHz. The FD/FFT module has been implemented using the 'Face Feature Tracking' SDK provided by N-Vision. The other components are mainly written in Java programming language in an object-oriented design. For specific functions with high computational costs (e.g. image filtering and similarity calculation) Intel's 'Integrated Performance Primitives' library is used to achieve a high performance for real-time processing. The

connection between Java and native code has been realized via the JNI interface. All processing modules have been embedded into the JMF 2.0 framework for easy integration into the BROAFERENCE system, which is used for content presentation. Figure 9 shows the layered structure of the software design.

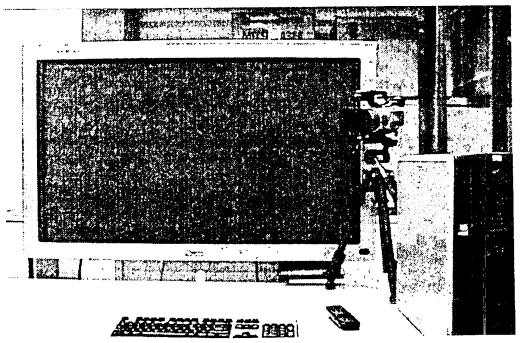

Figure 7 System Mock-up

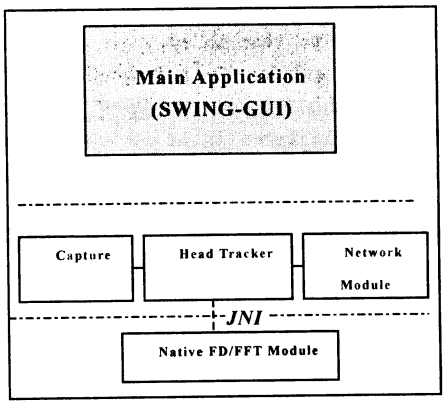

Figure 9 Layered Soft ware Structure using JMF-filter graph

In Figure 8, Figure 10 and Figure 11 screenshots of the GUI for controlling the capture process can be seen. The main controls provide a user interface to mainly adjust the neutral eye pose. The result of region classification can be visually checked by a moving rectangle that indicates which region has been selected. An eye model creation tool is also provided by the current GUI. It can be used for recording eye templates and selecting the area of feature point blocks for creating the models.

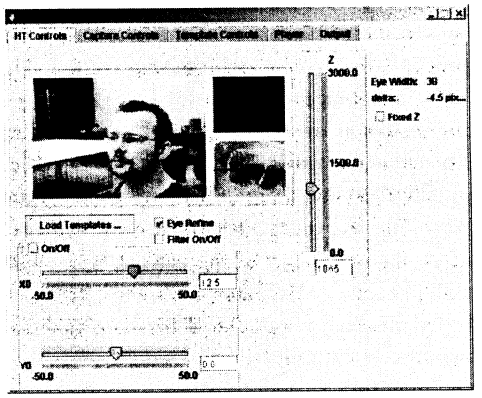

Figure 10 Main Control Panel

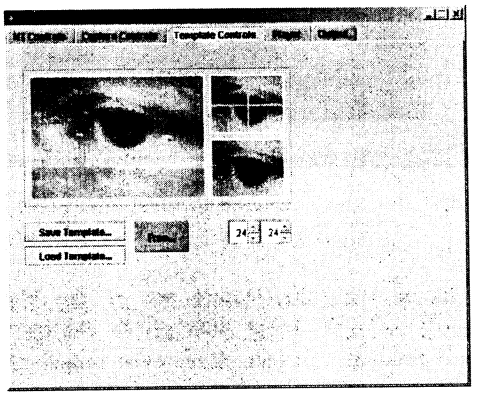

Figure 8 GUI for Eye pose Model creation

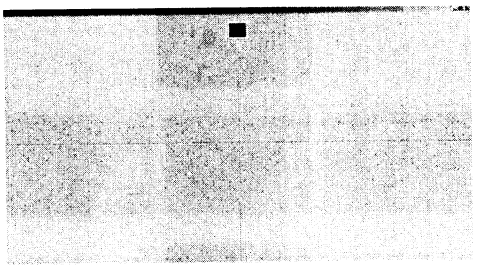

Figure 11 Example of feedback screen for region classification output

5. Conclusion and Future Work

A lot of effort is done to investigate the desires and interests of TV program consumers in order to provide high quality contents and keep the audience satisfied. The methods commonly used provide means of indirect measurement of statistical data (e.g. how many people watching which channel). In this paper we have introduced a system that provides a method for direct measurement of the attention level of the audience in front of a TV by detecting the region of interest on the screen from eye motion. The screen will be virtually divided into five regions, i.e. center, top, left, bottom and right region. The system can detect, which region is currently visited by the person in front of the screen. It is possible to estimate the ROI in real-time. The system is also able to deal with slightly changing lighting conditions, which may apply to a 'common living room situation'. The performance of detection is sufficient in the case of large size displays, since the eye pose changes are significant in this case. The performance when using smaller screens has to be evaluated. Only one user is currently supported by the system. In the case where the person in front of the screen wears glasses, the reflections may cause miss-judgment. We currently investigate a different approach that can detect more regions. It is planned to integrate the ROI tracker system into the BROAFERENCE platform – a multimedia distribution system that is currently under development. With doing so, it will be possible to measure the point of attention while presenting different contents to gain information about the interest of the examined person. By refining the approach by the introduction of expression detection, it may be possible to use our system for deeper analysis of the psychological impact that different multi media contents have onto humans emotions [3].

References

[1] Y.Sawahata, K.Komine, N.Uratani: "A Gaze Measurement System Permitting a Head Motion", FIT2004, Kyoto, 2004

[2] T.Katada, K.Tanaka, Y.Mizukami, Y.Wakaza, H.Okano: "A communication system for ALS patients using eyeball movement", FIT2004, Kyoto, 2004.

[3] U.Kowalik, T.Aoki, H.Yasuda: "A TV Rating System Using Facial Expressions", FIT2004, Kyoto, 2004