TV VIEWING INTERVAL ESTIMATION FOR PERSONAL PREFERENCE ACQUISITION

Hiroaki Tanimoto, Naoko Nitta, and Noboru Babaguchi

Graduate School of Engineering, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
tanimoto@nanase.comm.eng.osaka-u.ac.jp, {nitta, babaguchi}@comm.eng.osaka-u.ac.jp

ABSTRACT

The importance of personalized information services has been increasing. Description of personal preferences needs to be prepared beforehand to realize such services. We propose a system for automatically acquiring personal preferences from TV viewer's behaviors. Considering "when" a viewer is watching TV is highly related to the viewer's preferences, we focus on estimating the time interval during which a pre-registered viewer is watching TV. In this paper, we firstly describe the outline of the personal preference acquisition system, and address a method for estimating the TV viewing intervals based on the appearance of frontal faces. Experiments resulted in a precision rate of 97.1% and a recall rate of 70.6% on average for TV viewing interval estimation.

1. INTRODUCTION

Recently, the importance of personalized information services has been increasing. In order to realize such services, it is essential to prepare description of personal preferences beforehand. Conventionally, personal preferences have been described directly by users or acquired by user feedback. However, this can be an excessive burden to users since the data has to be always updated to conform to continuous changes of preferences over time. Therefore, we propose a system for automatically acquiring personal preferences by observing viewer's behaviors in front of TV[1].

The proposed system records TV viewers by cameras and microphones and recognizes their identification and behaviors by analyzing the recorded video and audio. The TV viewer's personal preferences are estimated by temporally associating his/her recognized behaviors with the content information of the corresponding video segments.

There are some previous works to automatically acquire personal preferences from user's behaviors. Web browsing histories[2], operation records of remote controls[3], and viewing histories of TV programs[4] have been used as sources to analyze user's behaviors. Since they have been considering only explicit inputs such as keywords to search when browsing Web and fast-forwarding when watching TV, these systems have only enabled us to obtain personal preferences that users have already been aware of. On the other hand, since we consider passive behaviors such as laughing, clapping, and humming to music, even personal preferences which viewers are still unaware of are expected to be obtained. Moreover, traditional personal preference acquisition method from TV viewing histories have only examined viewer's behaviors for TV program and have not considered the differences of degree of interest toward the content. The proposed system tries to acquire more detailed personal preferences by examining TV viewer's behaviors for video segments and by considering differences of the interest inferred by emotional differences represented by his/her behaviors.

Since there can be several people in front of the TV, identifying each viewer is necessary to individually acquire their personal preferences. Normally, at home, where the proposed system is expected to be used, viewers are limited to specific people such as family members. Therefore, viewers can be registered beforehand to create their face models for identification. However, a viewer might not be interested in TV even when he/she is in front of the TV. Therefore, "when" he/she is watching TV is considered to be highly related to his/her preferences. Here, assuming that viewers face the TV when they are watching TV, we try to estimate their TV viewing intervals by identifying only viewers who are facing the TV.

2. PERSONAL PREFERENCE ACQUISITION SYSTEM

2.1. Outline of the system

Fig.1 shows the outline of the proposed system. First of all, viewers register themselves before using the system. Cameras and microphones are set up in the TV viewing space to record viewers who are watching TV. Whenever the registered viewers are in the TV viewing space, they are identified and their behaviors are recognized by analyzing the recorded video. Their personal preferences are then estimated by associating their recognized behaviors with the content information of the corresponding video segments. The estimated personal preferences are stored as viewer profiles, which are composed of keywords and their scores. The score represents the degree of the viewer's interest toward the content represented by the keyword. Personalized services such as personalized video abstraction[5] can be provided by using the viewer profiles. Here, we assume that all videos have the MPEG-7[6] metadata which describes their content in a text form. The flow of the proposed system is shown below.

1-4244-0367-7/06/$20.00 ©2006 IEEE

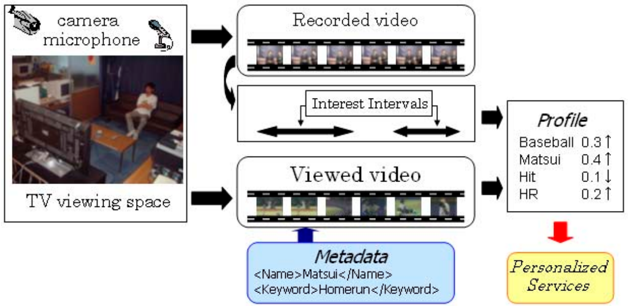

Fig. 1. System configuration

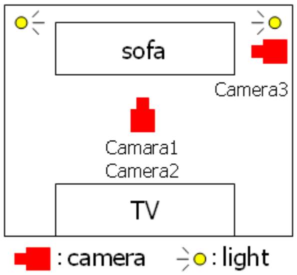

Fig. 2. TV viewing space

I. **Viewer Registration**: Viewers to acquire their personal preferences are registered.

II. **Video Recording**: Viewers in front of the TV are recorded by cameras and microphones.

III. **Viewer's Behavior Recognition**: After the registered viewers are identified, their behaviors are recognized by analyzing the recorded video.

IV. **Personal Preference Estimation**: Personal preferences of each viewer are estimated by temporally associating his/her recognized behaviors with content information of corresponding video segments.

V. **Viewer Profile Creation**: The viewer profile is created for each viewer and updated over time.

2.2. TV viewing space

Fig. 2 shows our experimental setup of the TV viewing space. There is a sofa in front of a TV and three cameras are installed to record people on the sofa. Camera 1 and 2 are installed in front of the sofa and Camera 3 is installed on the left side of the sofa. Camera 2 records viewer's face in close-up and Camera 1 records the entire sofa. There are two lights on both sides of the sofa. In this paper, the video with Camera 1 is used for TV viewing interval estimation. The frame rate of the video is 30fps and the image resolution is 176×120.

3. TV VIEWING INTERVAL ESTIMATION BASED ON VIEWER IDENTIFICATION

TV viewing is the viewer's basic behavior that infers his/her interest toward the content. Considering that a viewer faces the TV when he/she is watching TV, we try to identify him/her and estimate the time intervals when he/she is actually watching TV.

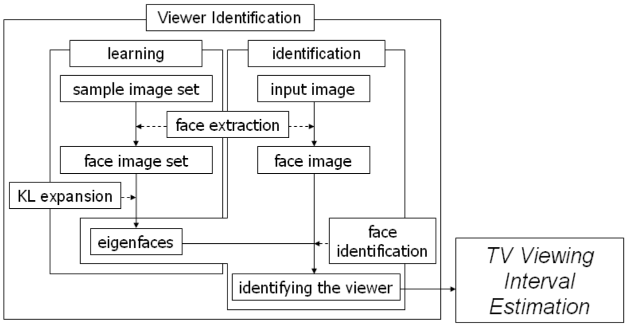

Fig. 3. A flowchart of TV viewing interval estimation

Viewer identification consists of two steps: model learning and identification. In the model learning step, a face image set is created from a sample image set for each viewer. Then, eigenfaces[7] are obtained as a face model of the corresponding viewer by applying KL expansion to the face image set. The identification step also creates a face image from an input image and checks if the face image can be approximated with the obtained eigenfaces. TV viewing intervals are then estimated by temporally examining the identification results. Fig.3 shows the flow of the proposed method.

3.1. Viewer identification

We proceed to describe the details of three processes in viewer identification: face extraction, eigenface construction, and identification with eigenfaces.

3.1.1. Face extraction

A face image is defined as an $n \times n$ image locating the face in the center. It is extracted from an image recorded with Camera 1 as follows. The system has an image of the TV viewing space without any person as a background image.

I. The pixels which satisfy the following conditions are extracted as skin color regions: $max(R, G, B) > RGB\_th$, $H\_low < H < H\_high$, and $S\_low < S < S\_high$. Here R, G, B are the RGB differences between the input and the background images, H and S are the hue and the saturation of the input image. $max(R, G, B)$ represents the maximum value among R, G, and B. RGB_th is the threshold of RGB difference, and H_high, H_low, S_high, and S_low are the upper limits and the lower limits of hue and saturation, respectively.

II. Gray-scale opening and closing are applied to smooth the border.

III. The connected pixels are extracted as candidates of face regions.

IV. The connected pixels whose size, si, y-coordinate of the centroid, wy, and the degree of roundness($= (4\pi \times \text{pixels})/(\text{length})^2$), $ratio$, satisfy the following conditions are obtained as a face region: $si > size\_th$, $wy < wy\_th$, and $ratio > ratio\_th$, where $size\_th$, wy_th, and $ratio\_th$ are the thresholds. An $n \times n$ image is obtained as a face image with the centroid of the connected pixels as its center.

V. The face image is transformed into graylevel image in such a way that the average graylevel of all pixels is 0.

3.1.2. Eigenface construction

Eigenfaces are obtained from the extracted sample image set of each viewer. The face image set is represented by $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_m}\}$. The meanface is defined as $\boldsymbol{\mu} = \frac{1}{m}\sum_{k=1}^{m} \boldsymbol{x_k}$. The eigenfaces $\boldsymbol{u_k}$ are the eigenvectors of the covariance matrix $\boldsymbol{\Sigma} = \frac{1}{m}\sum_{k=1}^{m}(\boldsymbol{x_k} - \boldsymbol{\mu})(\boldsymbol{x_k} - \boldsymbol{\mu})^t$. The direction of viewer's face when he/she faces the TV depends on where he/she sits. Therefore, the eigenfaces are obtained for three different positions: on the left side, in the middle, and on the right side of the sofa.

3.1.3. Viewer identification with eigenfaces

The face image is identified with the eigenfaces as follows.

I. Determine which eigenfaces to use according to x-coordinate of the center of the input face image, wx. Use the eigenfaces of the viewer sitting on the left side of the sofa when $wx < wx\_l$, of the viewer sitting in the middle of the sofa when $wx\_l < wx < wx\_r$, and of the viewer sitting on the right side of the sofa when $wx\_r < wx$. wx_l and wx_r are the thresholds to determine which eigenfaces to use for viewer identification.

II. The face image in each frame is approximated with eigenfaces as follows. Let $\boldsymbol{x}$ represent an input face image. The difference between the input face image and the meanface is defined as $\boldsymbol{D} = \boldsymbol{x} - \boldsymbol{\mu}$. $\boldsymbol{D}$ is projected into the d-dimensional face space as $\boldsymbol{D_f} = \sum_{k=1}^{d} \omega_k \boldsymbol{u_k}$ with the weight vector $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_d)$, where $\omega_k = \boldsymbol{u_k}^t(\boldsymbol{x} - \boldsymbol{\mu})$.

III. The difference between $\boldsymbol{D}$ and $\boldsymbol{D_f}$ is defined as $\varepsilon^2 = \frac{\|\boldsymbol{D}-\boldsymbol{D_f}\|^2}{n^2}$. If $\varepsilon^2 \leq \varepsilon\_th^2$, the face image of the current frame is identified as the viewer. Otherwise, if the face image of the previous frame is identified as the viewer and more than id_th frames among the following N frames have the face images that satisfy $\varepsilon^2 \leq \varepsilon\_th^2$, the face image of the current frame is identified as the viewer. ε^2_th is the threshold of viewer identification using eigenfaces.

3.2. TV viewing interval estimation

TV viewing interval is estimated by temporally examining the identification results as follows.

I. The frames with the identified viewer are counted within $30 \times M$ frames, which correspond to M seconds.

II. If there are more than $view\_th$ frames with the identified viewer, we determine that the corresponding M seconds were viewed by the viewer.

III. I and II are repeated for the next $30 \times M$ frames.

4. EXPERIMENTS

We conducted the experiments with four viewers (Viewer A, B, C, and D). 50 sample images were provided to obtain eigenfaces for each viewer sitting on the left side, in the middle, or on the right side of the sofa.

Table 1. Parameters for face extraction

RGB_th	H_high	H_low	S_high	S_low	$size\_th$	wy_th	$ratio\_th$	n
30	50	5	0.5	0.15	250	0.5	0.3	15

Table 2. Parameters for TV viewing interval estimation

wx_l	wx_r	d	ε^2_th	N	id_th	M	$view\_th$
2/7	5/7	5	1200	60	15	1	15

Table 1 and 2 show the parameters determined experimentally for face extraction and TV viewing interval estimation.

The results were evaluated with the precision and recall rate which are defined as $Precision = \#3/\#2$ and $Recall = \#3/\#1$, where $\#1$, $\#2$, and $\#3$ are the number of the frames with the viewer watching TV, the number of the frames with the identified viewer, and the number of the frames with the correctly identified viewer, respectively, for viewer identification. For TV viewing interval estimation, $\#1$, $\#2$, and $\#3$ are the number of video segments viewed by the viewer, the number of video segments determined as viewed by the viewer, and the number of video segments correctly determined as viewed by the viewer, respectively, where the video segments correspond to M seconds.

4.1. Evaluation of viewer identification

Firstly, we tried to identify each of Viewer A, B, and C who was watching TV in the middle of the sofa alone. Fig.4 shows example images. Each viewer was recorded for 25 seconds, which correspond to 750 frames. Table 3 shows that the recall rate was 91.7% and the precision rate was 93.7% on average.

Secondly, we recorded Viewer A who was not watching TV in the middle of the sofa for about 13 seconds, which correspond to 400 frames, and examined whether these images were identified as either the positive viewing: he/she was watching the TV, or the negative viewing: he/she was not watching it. Our method correctly identified these images as the negative viewing with the accuracy of 99.0%.

Thirdly, we recorded Viewer A, B, and C who were watching TV together. They sat in the order of (B, A, C), (C, B, A), and (A, C, B) from the left side of the sofa, so that everyone sits on all three positions on the sofa. Fig. 5 shows example images. They were recorded for 30 seconds each time, which correspond to 900 frames. Tables 4, 5, and 6 show the results for each viewer sitting on each position. You can see that the recall rate for the viewers who sat on either side of the sofa was sometimes not satisfactory. Since the lights are set up right next to the sofa, the lighting condition was largely affected by even the slight changes of their sitting position and decreased the recall rate.

4.2. Evaluation of TV viewing interval estimation

We prepared a TV video composed of news of 1 minute 23 seconds, commercials of 45 seconds, and a weather forecast of 35 seconds. We asked two viewers (Viewer A and D) to watch the news and the weather forecast but not the commercials. We then examined if the video segments corresponding to the news and the weather forecast were determined as the

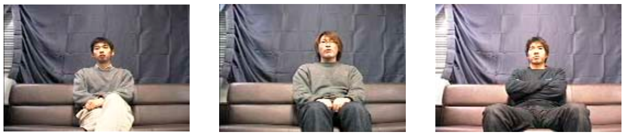

Fig. 4. Example images (one person is on the sofa)

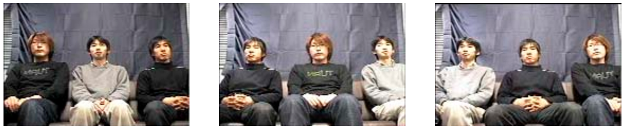

Fig. 5. Example images (three people are on the sofa)

Table 3. Viewer identification when one person is on the sofa

viewer	#1	#2	#3	precision (%)	recall (%)
A	750	891	750	84.2	100
B	750	692	670	96.8	89.3
C	750	644	644	100	85.9

Table 4. Identification of Viewer A when three people are on the sofa

position	#1	#2	#3	precision (%)	recall (%)
left	900	900	900	100	100
center	900	895	895	100	99.4
right	900	900	900	100	100

Table 5. Identification of Viewer B when three people are on the sofa

position	#1	#2	#3	precision (%)	recall (%)
left	900	547	547	100	60.8
center	900	655	655	100	72.8
right	900	672	672	100	74.7

Table 6. Identification of Viewer C when three people are on the sofa

position	#1	#2	#3	precision (%)	recall (%)
left	900	862	862	100	95.8
center	900	903	900	99.7	100
right	900	296	296	100	32.9

positive viewing. Table 7 and 8 each shows the results when each viewer was sitting in the middle of the sofa alone and the results when both viewers were sitting together, with Viewer A on the right and Viewer D on the left of the sofa. The face of the Viewer D sometimes inclined differently from the sample images, resulting in the lower recall rate. A few commercial video segments were identified as the positive viewing since he was still facing the TV even though he was not watching TV.

Finally, we recorded Viewer A watching a 30-minute Japanese animated comedy series "Chibi Maruko chan" to examine the relations between TV viewing intervals and personal preferences. Fig. 6 shows the results of TV viewing interval estimation. The black segments represent the time intervals determined as Viewer A's positive viewing. Almost all parts of the program were determined as the positive viewing, while the commercials were determined as the negative viewing since he took different actions such as looking away and scratching his face. These results indicate the mutual relation exists between TV viewing intervals and personal preferences since the viewer was not interested in the commercials. The video segments around 23 minutes after the program started was determined as the negative viewing; however, he was actually laughing, indicating strong interest in the scene designated by * in Fig. 6, where one of the main characters, Maruko, was reading out her funny essay. The viewer identification method should be improved to handle changes in the facial expressions to accurately acquire personal preferences.

Table 7. TV viewing interval estimation when one person is on the sofa

viewer	#1	#2	#3	precision (%)	recall (%)
A	118	91	90	98.9	76.3
D	118	72	72	100	61.0

Table 8. TV viewing interval estimation when two people are on the sofa

viewer	#1	#2	#3	precision (%)	recall (%)
A	118	88	85	96.6	72.0
D	118	92	86	93.5	72.9

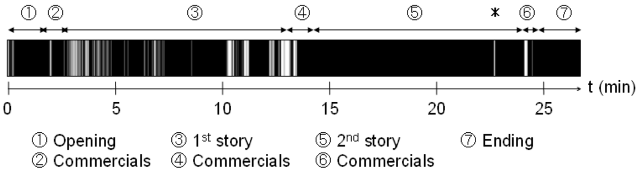

Fig. 6. A result of TV viewing interval estimation

5. CONCLUSIONS

This paper proposed a system for acquiring personal preferences from TV viewers' behaviors. Considering TV viewing as a basic behavior related to personal preferences, we proposed a method of estimating the TV viewing intervals by identifying viewers facing TV. The proposed viewer identification method obtained the recall rate of 91.7% and the precision rate of 93.7% on average when a user was watching TV alone. However, the lighting condition sometimes affected the results when three viewers were watching TV together. The TV viewing intervals were estimated with the recall rate of 97.1% and the precision rate of 70.6% on average. In order to acquire accurate personal preferences, considering changes in facial expressions will be our future work. This work was partly supported by HBF Inc.

6. REFERENCES

[1] H. Tanimoto, N. Nitta, and N. Babaguchi, "Viewing Interval Estimation for Personal Preference Acquisition in TV Viewing Environment," Technical Report of IEICE, PRMU 2005-151, pp.13-18, Jan. 2006.

[2] T. Tsandilas, and M.C. Schraefel, "User-Controlled Link Adaptation," Proc. of the 14th ACM Conf. on HT'03, pp. 152-160, Aug. 2003.

[3] K. Masumitsu, and T. Echigo, "Video Summarization Using Reinforcement Learning in Eigenspace," IEEE ICIP-2001, Vol. 2, pp. 267 - 270, Sep. 2000.

[4] Z. Yu, and X. Zhou, "TV3P: An Adaptive Assistant for Personalized TV," IEEE Trans. on Consumer Electronics, Vol.50, No 1, pp. 393-399, Feb. 2004.

[5] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, "Personalized Abstraction of Broadcasted American Football Video by Highlight Selection," IEEE Trans. Multimedia, Vol. 6, No. 4, pp. 575-586, Aug. 2004

[6] "Overview of the MPEG-7 Standard (version 6.0)," ISO/IEC JTC1/SC29/WG11 N4509, Dec. 2001.

[7] M. Turk, and A. Pentland,"Eigenfaces for recognition," J. of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86, Mar. 1991.