Estimation of User Interest Using Time Delay Features between Proactive Content Presentation and Eye Movements

Jean-Baptiste Dodane, Takatsugu Hirayama, Hiroaki Kawashima, and Takashi Matsuyama
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, Japan
jbdodane@vision.kuee.kyoto-u.ac.jp, {hirayama, kawashima, tm}@i.kyoto-u.ac.jp

Abstract

Human-machine interaction still lacks smoothness and naturalness despite the widespread utilization of intelligent systems and emotive agents. In order to improve the interaction, this work proposes an approach to estimate user's interest based on the relationships between dynamics of user's eye movements, more precisely the endogenous control mode of saccades, and machine's proactive visual content presentation. Under a specially-designed presentation phase to make the user express the endogenous saccades, we analyzed delays between the saccades and the presentation events. As a result, we confirmed that the delay while the user's gaze is maintained on the previous presented content regardless of the next event, called resistance, is a good indicator of the interest estimation (70% success, upon 20 experiments). It showed higher accuracy than the conventional interest estimation based on gaze duration.

1. Introduction

1.1. Mind Probing

Over the past few years, machines have spread in numerous aspects of everyday life. We frequently deal with them as much as with our partners. However, we still behave with those machines as a user who has to execute specific commands to get a desired result. We think of them as passive and reactive objects. Human-machine interaction is hence not as smooth as human communication despite its evolutions in affective computing and other related domains.

So as to make machines' behavior closer to humans' one, we draw our inspiration from basic interaction scenes that happen in everyday life. Actually, in order to smooth a conversation with a partner, we naturally adopt proactive behaviors. In other words, we try to probe the partner's mental state, by bringing up a general topic or asking a friendly question, while showing expectation for a reply with eye contact or voice intonation. Such proactive approaches make the partner reveal his/her internal (mental) state through words, tone, facial expressions, gaze, and so on. As a consequence, we are able to make the conversation evolve smoothly toward a topic suitable for both parties. This is all the more true in a situation where one individual tries to provide information or a service to an undecided person: the former should probe the latter's mind casually and understand the latter's needs by analyzing reactions. This is just what is happening when a customer is seeking for a present, or when a tourist is looking for a destination in a travel agency. We believe that the machines should acquire this behavior of (1) proactively approaching the user and (2) estimating his/her internal state based on reactions for it, without waiting for any commands from the user, in order to move on toward a smoother and more efficient HCI. We call that concept *Mind Probing*.

Figure 1. A prototype information service system with a large display and 3 cameras (circled in white).

Our goal is to build an information service system able to probe the user's mind and estimate his/her interest in some presented contents. This is a first step toward realizing intelligent recommendation devices which provide relevant contents according to the situation after understanding the user's interest. Figure 1 shows our prototype system.

1.2. From mental states to eye movements

Although some researchers lately consider interest as an emotion [1], it is basically a mental state that causes attention to focus on something. We are attracted to its intimate connection with the attention. Degrees of attention are measured mostly by (1) the effort that the person provides to focus on his/her goal and (2) the resistance that the person opposes any influences [2]. We use those measurements of attention to evaluate the user's interest.

How can we measure the attention? We focus on overt attention which humans express in their eye movements. Based on the eye-mind hypothesis [3], eyes are often a window into the mind. We humans turn our gaze (central fovea of the retina) on an object with interest to get detailed information of it. The gaze action has jerky fast movements. They are called *saccades* and are closely relevant to attention [4]. The saccades are programmed under two different control modes: exogenous (bottom-up, stimulus-driven, depending on the saliency of objects in the visual field) and endogenous (top-down, goal-directed, depending on volition of the person). We have to extract specifically the saccades of the latter type for analysis, which express the overt attention occurring in top-down process, whereas the saccades of the former type have no relationship with goals of the user [5][6]. For that purpose, we design a proactive content presentation scenario to trigger the exogenous saccades and the endogenous saccades separately.

In our assuming situation, the information service system displays some contents on a screen and estimates the user's interest through the eye movements. We need a precise and reliable "bridge" revealing a connection between the eye movements and the contents. We consider that the bridge is their dynamics. The user will respond to proactive and dynamic content presentation. And the response patterns of eye movements are sure to reflect the interest. We especially focus on delays between presentation events and gaze switches, that is, the saccades. Furthermore, we consider that dynamics of humans are most likely to reveal the true nature of them: they are often unconscious and can hardly be simulated.

According to the above argument about measurements of attention, we compute two indicators for the delays called *reaction* and *resistance*.

- The *reaction* represents the response time to switch the gazing to the next presentation event.
- The *resistance* represents the duration keeping the user's gaze on the previous presented content regardless of the next event, more or less stimulating, which occurs in a different part of the user's visual field.

We make hypotheses that relate the delays for the proactive content presentation to the interest and test the hypotheses through some experiments.

1.3. Related researches

Many researchers have proposed how to estimate the user's interest and mental states. Picard *et al*. analyzed the passively sensed physiological behaviors in order to recognize user's emotions and estimate the interest [7]. Mota and Picard applied the Hidden Markov Model to physiological information (posture features using pressure sensors mounted on a seat) to estimate the interest of children during a constraint satisfaction game [8]. In the continuity, Kapoor *et al*. used multiple modalities (facial actions, postures and game state) probabilistically combined to classify 3 mental states for the same game task [9]. But these researches were based on the passive sensing. Later, Onishi *et al*. analyzed the timing structures of proactive *face turning* behavior in human-human consensus building communication [10]. It would be worth checking if the results are similar in human-machine communication.

As far as we know, there are no previous researches analyzing the dynamics of eye movements in response to machine proactive behaviors in order to estimate the user's mental state. We have proposed the *Mind Probing* which combines interest, eye movements and proactive behaviors. In our preliminary work, we divided user's state into two phases called "input" and "evaluate" through proactive content presentation, to estimate the user's interest. The former is the state where the user reads all information of contents and the latter is another state where the user compares some contents. In order to induce the user to the "input" phase, the contents are exclusively presented by turns. On the other hand, the "evaluate" phase is triggered by redisplaying all contents at the same time. We focused on the frequency and the duration of user's gazes during the "evaluate" phase, but did not consider the dynamics between presentation events and user's gaze reactions. The weak point is that those two phases are not clearly separated. The user often reads again to remember the information during the "evaluate" phase. Such behaviors are not related to the interest.

2. Proactive presentation and hypotheses

2.1. Situation description

The machine side is a system providing information with the purpose of helping a user. It is a visual pool of new or rare contents. The user makes a choice from them under his/her lack of knowledge. They face each other and interact at a distance of approximately 1 meter. The machine proactively approaches the user by presenting contents likely to spark interest on a large display.

The user interacts with the machine by using only his/her eyes, who acquires information by the vision sense and talks by means of eye movements. On the other hand, the system

uses two different devices to interact, the large display to send information and cameras to receive user's signals (eye movements). Indeed, the user is not asked to talk or point an area. As shown in figure 1, the screen of the display is divided into several areas. A content which consists of some objects, is presented in each area. The system estimates which content the user is most interested in.

2.2. Scenario

In order to extract the endogenous saccades, we design a proactive content presentation that separates the two saccades control modes. The proactive content presentation consists of two phases:

- The **glimpse phase**: new objects are displayed with a very fast update rate on the screen, one area after another. Its purpose is to make the user realize the exogenous saccades. This phase will also let the user aware about what kind of objects are likely to be displayed in which part of the visual field. The user builds the cognitive map, which is a mental representation of spatial locations, in his/her mind [11]. The presented information is stored in the cognitive map. Its purpose is to connect the environmental information to an image that is built inside the mind for achieving an user's goal [12]. It is literally the "mind's eye". The quick update of the glimpse phase is set to prevent the user from reading and understanding all the information and to "tease" his/her interest.

- The **observation phase**: the same objects are displayed once again, in the same area as in the glimpse phase, but for a longer period. This time, the user can fulfill his/her interest by sufficiently reading some contents that just "teased" him/her in the glimpse phase. As the user had already seen them before, we believe that the saccades will be mainly endogenous, by referring to the cognitive map. The important detail is that the system redisplays each object in random order. The user cannot guess where the next object will be "redisplayed". But the user can pay attention to contents areas with interest thanks to the cognitive map.

Figure 2 and 3 profile the proactive content presentation scenario.

2.3. Hypotheses: dynamical relations between proactive presentation and user's interest

We evaluate the user's interest by calculating the delays between the object presentations and the following gaze switches. In the frame of the proposed scenario, we make the following hypotheses:

- During the *glimpse phase*, we believe that some contents, or some objects will stimulate the user's attention, and *he/she will be eager to see them again because of interested in their details*.

- Then, during the *observation phase*, we predict that *the user's gaze will be switched proactively to interesting contents*. We expect to observe a shorter delay for switching the user's gaze to the redisplayed object of content with interest, or a longer delay if a gazed object is more attractive than the next redisplayed object. This is our main hypothesis.

3. Apparatus and software architecture

3.1. System overview

The proposed system remotely measures the eyes movements with no intrusive devices, so as to achieve natural human-machine communication. It is composed of a 50 inch large display [1], 3 synchronized and calibrated cameras [2] and 2 backup lights, shown in figure 1.The objects of contents are presented and controlled by a display software coded using WinAPI.

3.2. Gaze estimation

The gaze estimation is performed in four steps: face detection, estimation of the face orientation, iris detection and estimation of the gaze direction.

First of all, user's face is detected with the Intel OpenCV library, using Haar-like filters. Facial features (45 points) are extracted by Active Appearance Model (AAM) algorithm [13], which is a statistical subspace model of shape and appearance, previously trained with a set of user's pictures captured in a preliminary experiment. The AAM is trained using 15 pictures with various head rotations for each user and each camera.

Then, a 3D face shape model, calibrated with stereo cameras in the preliminary experiment, made of 45 feature points, eyeball centers and iris radius, is fitted on the AAM by the bundle adjustment [14]; in fact, the optimization of translation and rotation parameters using the steepest descent method. Consequently, the 3D position of user's face is estimated. The irises are extracted by matching iris templates generated from the iris radius, and their 3D positions are estimated from the eyeball center and the iris radius.

After computing the straight line running through both the eyeball center and the iris center, its intersection with the display plane informs about the gazed point. Finally, estimation results of 3 cameras are integrated. The gaze estimation accuracy is about 5 degrees (= about 10 cm on the screen).

[1] 1106 mm height, 622 mm width

[2] Point Grey Research Grasshopper: UXGA, 30 fps, 8 bits gray scale.

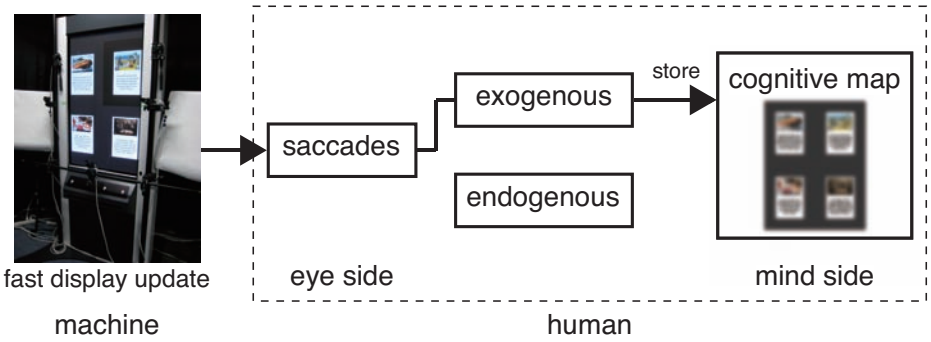

Figure 2. Interaction scenario and information flow in the glimpse phase.

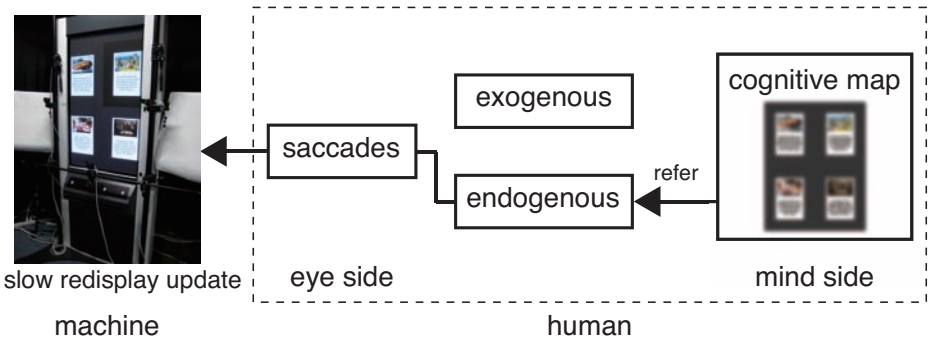

Figure 3. Interaction scenario and information flow in the observation phase.

4. Experimental procedure

The experiments evaluate the interest of several subjects by measuring the delays and durations of their eye movements in response to the content presentation, and comparing the delays with interview survey.

As recommended by Just and Carpenter [3], in such an experiment it is important to design a well-specified task. This is because we want to observe and measure the endogenous saccades, which are goal driven. A goal therefore has to be defined for the subjects to make sure they mentally focus during the experiment. In our experiments, we adopt a simple decision task which asks the subjects to choose among the presented contents, in our case unreleased movies. The task question is "If you are given a ticket, which movie would you want to see?".

4.1. Contents

Dispersion: Basically, we display information on the screen divided into 4 equivalent areas, each one can be controlled independently and display a content which consists of some objects called vignettes. We try to observe as clearly as possible the dynamic differences between the gaze movements to those 4 areas. We need to design 4 categories of contents (movies) that are most likely to separate interest of the subjects. Upon several preliminary trials, we discovered that the most important parameter is the proximity between those categories. Indeed, for 4 sets of vignettes (4 categories of contents) which share nothing in common, the subjects require a more intense reasoning to evaluate their differences, since correspondences between them are few. Moreover, there is no assurance that one of the sets stands out above the 3 other sets. On the other side, in the case where the 4 categories are really close and present a small dispersion, interest of the subjects is most likely to be the same for all of them, unless they have a very sharpened knowledge of the field.

We seize here the difficulty of finding the right dispersion of contents to observe a good difference. We adopt 1 genre of movies for 1 category. We humans often make a choice from some genres. The same could be said for movies. The subjects can evaluate the differences of movie genres based on their criterion.

Nature: All presented contents are the unknown movies for the subjects, which have not been released at the time of the experiments yet, to ensure the fairness among the 4 areas. Each vignette has a fixed size, 200 pixels width and 300 pixels height, and 24 bits color, which is made up of a picture (upper part) and text (lower part). The picture is

Figure 4. An example of vignettes (refer to Internet Movie Database, http://www.imdb.com).

a poster or a production still of movie whose area has 200 pixels width and range from 100 to 140 pixels in height. The text description is always made of 6 or 7 lines, written in English, so that it takes approximatively the same time to read it for all of the vignettes. An example of vignettes is shown in figure 4.

The picture information is as much as possible representative of the genre. The images of action movies show fights or cars, whereas those of comedy movies show smiles or goofy characters. The text information gives some clues about the storyline, the cast, and so on, written in the same style as movie reviews of specialized magazines.

We should not give the vignettes eye-catching appearance because the exogenous attention is very sensitive to sudden luminance changes. However, it is difficult to collect pictures without eye-catching appearance. We hence set the background color of the whole screen to dark gray and design white background area of the vignette to be larger than picture area, in order to increase visual fairness between vignette areas.

The subjects might miss the presentation events on the large screen when the system updates the diagonally opposed vignette for their gazing vignette. We need to work out how to make sure that they are always aware of the events, without extremely triggering their attention. This is realized by framing the last updated vignette with a black border.

4.2. Time line

The first phase (glimpse phase) is executed for two purposes: letting the subject build the cognitive map and arousing interest in the contents. All vignettes are displayed at 3 seconds intervals by turns. We set the intervals according to the weight of information of a vignette. In the second phase (observation phase), the vignettes are redisplayed in random order at 8 seconds intervals, in order to make the user read them sufficiently. The duration of an experiment is limited to about 4 minutes because the subject feels tiresome to give attention to the screen for a long time.

The time line of the experiment is shown in figure 5.

1. In the introduction phase, the system displays 5 frames explaining the task. Each of these display events occurs at the label *task*.

2. The label *title* denotes an event of displaying basic information, e.g. genre, poster, etc., about movie in each area.

3. In the glimpse phase, 5 vignettes (a, b, c, d, e) of each 4 contents are displayed at one area after another, from the top left to the bottom right ("*TL*" stands for Top Left, "*BR*" for Bottom Right, etc.).

4. In the observation phase, the label *rand* means that the vignette of contents is redisplayed in random order. The vignette remains in the area until the next update occurs in the same area.

4.3. Measurements

During the observation phase only, we define 3 dynamic indicators to measure the response by eye's saccades to the vignette redisplays. They are illustrated on figure 6.

- $Reaction(x)$: the response time of gaze switch from the previously gazed vignette to the next redisplayed vignette x.

- $Resistance(x)$: the delay interval while the gaze is maintained on vignette x regardless of the next redisplayed vignette. It is the same value as the *reaction*, but credited to the previously gazed vignette x.

- $Duration(x)$: the *duration* of gazing on vignette x.

In the purpose of verifying the validity of our hypotheses, we correlate the dynamic indicators with subjective evaluation (choice of a movie) through interview survey after the experiment.

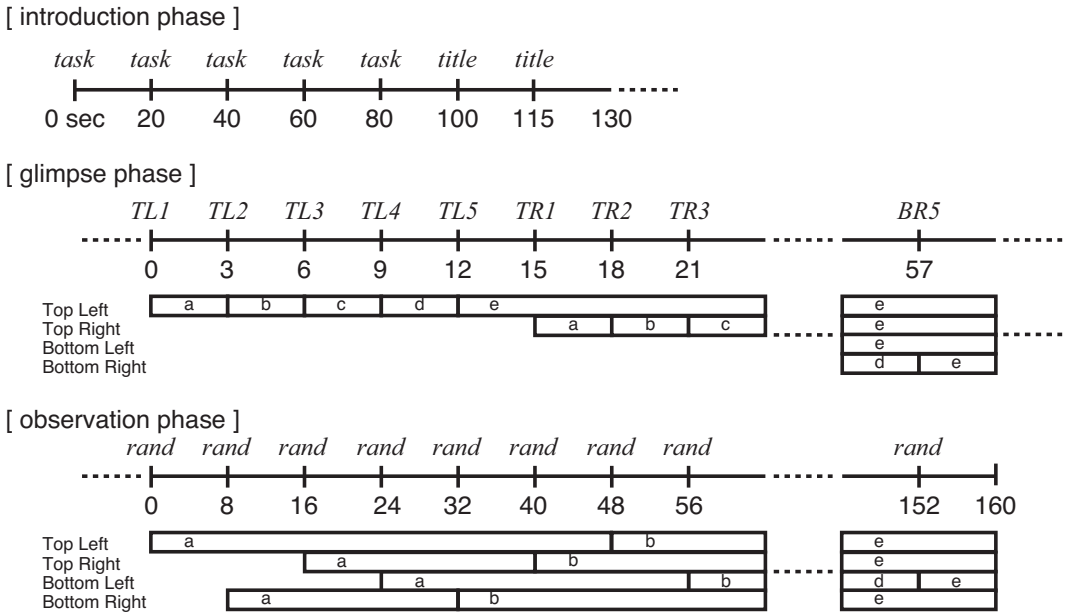

Figure 5. Time line of the experiment. A rectangular bar which includes symbol "a", "b", "c", "d", or "e" represents duration displaying a vignette. The label *task* and *title* denote the events of displaying the task instruction and the basic information about 4 contents, respectively. "*TL*" denotes a vignette presentation in Top Left area, "*TR*" in Top Right, and "*BR*" in Bottom Right. The label *rand* means that the vignette of contents is redisplayed in random order.

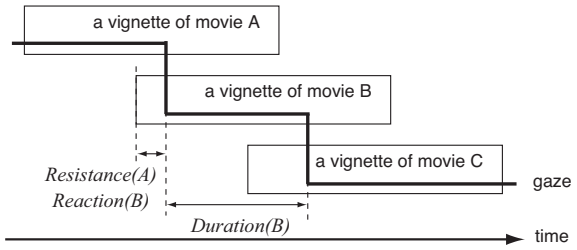

Figure 6. Examples of the 3 dynamical indicators during the vignette redisplay of 3 movies A, B, C.

Duration	Reaction	Resistance
35.0%	20.0%	70.0%

Table 1. Interest estimation accuracies of the 3 indicators.

5. Experimental results

5.1. Data analysis

We conducted 4 experiments for each of 5 subjects. We selected action, animation, comedy, drama movie as the 4 movie genres of each experiment. The display area of each genre was different between all experiments. A content consisted of 5 vignettes. Hence, an experiment featured 20 vignette redisplays (4 movies * 5 vignettes). We did not measure the 3 indicators when a vignette update happened at an area where the subject was already looking. We computed the averages per movie of each 3 indicators.

Accuracy of interest estimation using each 3 dynamic indicators: Table 1 shows accuracies of interest estimation calculated as follows: we counted 1 point for the indicator *duration* if a movie with the longest *duration* was the selected one with the most interest in the interview survey. Else, we counted 0 point. The counted points were then divided by the number of experiments to yield success rate of the matching. The rate is the accuracy of interest estimation. We also applied the procedure to the other indicators, the *reaction* (in the case where a movie with the shortest *reaction* corresponds to the selected one, we counted a point to the indicator) and the *resistance* (in the case where a movie with the longest *resistance* corresponds to the selected one, we counted a point).

The table reveals that the *reaction* is a rather poor indicator of the interest estimation. Although we expected it to be shorter toward more interesting movies, it was not the case at all. However, on the opposite, the *resistance* was very often (70.0%) associated with the interest. It raises the *resistance* as the best dynamic indicator for the subject's interest estimation, and shows that the more the users are interested in a object, the less they switch gaze toward the next object presentation.

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5	
	Interest	Others	Interest	Others	Interest	Others	Interest	Others	Interest	Others
Experiment 1	962	-321	379	-126	27	-9	410	-137	574	-191
Experiment 2	-359	120	3213	-1071	396	-132	650	-217	54	-18
Experiment 3	478	-159	740	-247	-22	7	251	-84	-342	114
Experiment 4	436	-145	2190	-730	-170	57	1067	-356	376	-125
Ave.	379	-126	1631	-544	58	-19	595	-198	165	-55
SD	547	182	1313	438	241	80	355	118	401	134

Table 2. Average *resistances* (msec) of the selected (most interesting) movie in the interview survey ("Interest") and the 3 other movies ("Others"). Their average and standard deviation for each subject are shown at the bottom 2 lines of this table.

Duration	**Reaction**	**Resistance**
43.8%	25.0%	81.3%

Table 3. Interest estimation accuracies except the 3rd subject.

Analysis of relation between *resistance* and interest: Table 2 shows the average *resistances* for each 5 subjects, each one having undertaken 4 experiments. The experiments followed the same time line, only the movies were changed. The longer the *resistance* of a movie is, the more interesting the vignettes of the movie are, because the subject ignores the vignette update of the other movies. The "Interest" column is the average *resistance* of the selected (most interesting) movie in the survey. The column "Others" is the average *resistance* of the 3 other movies. Although we tried to have a fair competition between the movies by making similar vignettes for the 4 genres, the vignettes still contained different image colors and words. As a consequence, the standard deviation of *resistance* within the 4 experiments became large.

From table 2, we consider the 3rd subject is an outlier. His gaze response pattern was very different from the other subjects because he switched his gaze immediately to almost all of the redisplays. The interest estimation accuracy of *resistance* except the 3rd subject result was 81.3% as shown in table 3.

5.2. Discussion

As the *reaction* indicator could not estimate the interest satisfactorily, a good question to ask ourselves is whether the cognitive map was actually built inside the subjects' mind during the glimpse phase. We must evaluate the question by comparing the results of our presentation scenario to those of scenario without the glimpse phase. As a reason of the poor results for the *reaction*, we consider that the subjects did not struggle for reacting quickly because they knew that they have enough time to switch gaze later, since the next redisplayed vignette does not disappear quickly. They therefore kept on gazing to the contents with interest, that is, the *resistance* reflected the subject's interest.

Among the measures of the delays, there were large standard deviations. Moreover, the more the subject advanced in the time line, the longer the delay became. In the case where the subject lagged on a vignette, the delay of the following vignette summed up with the lag because all vignette redisplays were separated by a constant gap, 8 seconds. Indeed, it can be noticed on several results. Figure 7 shows an example of it. The obvious solution would be taking into consideration the lag when the gaze switches to the next redisplayed vignette. With an online real-time machine, it is possible to redisplay the next vignette after 8 seconds from the time the user switches gaze. This would ensure that the previous latencies do not have a negative influence on the rest of experiment.

6. Conclusion

6.1. Findings

We can conclude that according to the evaluation of our proactive behavior model as the following,

- user's interest can be estimated via the dynamics of eye's response by managing the 2 phases presentation. The *resistance* indicator, the delay interval while the gaze is maintained on a object without regard to the other events, is efficient for the estimation.

Dynamics will create a new world where humans and machines can interact together.

Although not including the results in this paper, we also verified the contraposition of our hypothesis, *i.e*. whether the least interesting movie corresponds to the shortest *resistance*. It was not true. We consider that interest and disinterest may not belong to a single dimension. "Intensity" of interest therefore must be multidimensional.

6.2. Future works

Our presentation design to frequently update objects must be better for interest estimation. However, we had one outlier among 5 subjects, who did not adapt to the presentation design. And, it may not be comfortable design for

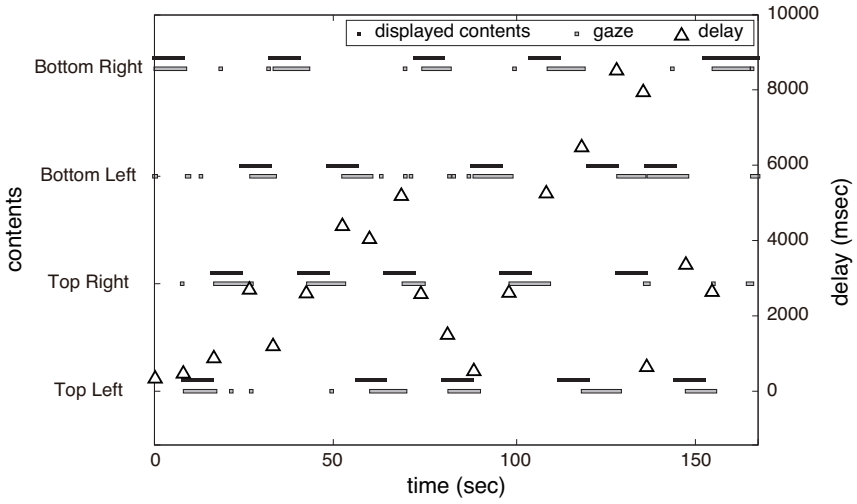

Figure 7. An example of the measured delays (subject 4, experiment 3).

users. We consider that the system should switch some presentation designs adapted to each user's personality. We also need to evolve the design to a new one with balance between interest estimation performance and usability.

The next step after estimating the interest is undoubtedly to use it and to provide an adapted response by interactive system. Instead of just estimating the mental state on a real-time measuring system, we can try to make an impact on it, making it evolve to a position with more benefits. The future system will be able to skillfully provide sensible contents to user, with the objective of increasing satisfaction of the user. For example, if the system notices that among the presented four movies, two of them attract the user stronger than the two others, the system can give to the user a deeper insight of the movies, or replace them to provide contents related only to interest. In that way, the system would perform a recommendation based on the user's interests and feelings, differences for every personality, and increase its integration with humans.

Acknowledgments.

This work is in part supported by Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan under the contract of 18049046.

References

[1] P.J. Silvia. Exploring the psychology of interest. *Oxford University Press*, 2006.

[2] W. McDougall. An outline of psychology. *Sigaud Press*, 2007.

[3] M.A. Just and P.A. Carpenter. Eye fixations and cognitive processes. *Cognitive Psychology*, 8:441–480, 1976.

[4] L.E. Sibert and R.J.K. Jacob. Evaluation of eye gaze interaction. *Proceedings of the SIGCHI Conference on Human Factor in Computing Systems*, 281–288, 2000.

[5] H. Deubel and W. Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, 36,12:1827–1837, 1996.

[6] M. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32,1:3–25, 1980.

[7] R.W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23,10:1175–1191, 2001.

[8] S. Mota and R.W. Picard. Automated posture analysis for detecting learner's interest level. *Computer Vision and Pattern Recognition Workshop*, 5:49, 2003.

[9] A. Kapoor, R.W. Picard, and Y. Ivanov. Probabilistic combination of multiple modalities to detect interest. *Proceedings of the 17th International Conference on Pattern Recognition*, 3:969–972, 2004.

[10] T. Onishi, T. Hirayama, and T. Matsuyama. What does the face-turning action imply in consensus building communication? *Proceedings of the 5th International Workshop on Machine Learning for Multimodal Interaction*, 26–37, 2008.

[11] E.C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55,4:189–208, 1948.

[12] R.M. Downs and D. Stea. Cognitive maps and spatial behavior: Process and products. *Image and Environment*, 8–26, 1973.

[13] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *Proceedings of the 5th European Conference on Computer Vision*, 2:484–498, 2001.

[14] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – A modern synthesis. *Vision Algorithm: Theory & Practice*, (B. Triggs, A. Zisserman, and R. Szeliski, eds.), Springer-Verlag LNCS 1883, 2000.