Effects of Emotion-induction Words on Memory of Viewing Visual Stimuli with Audio Guide

Mashiho Murakami[1], Motoki Shino[1], Katsuko T. Nakahira[2][a] and Muneo Kitajima[2][b]

[1]*Department of Human & Engineered Environmental Studies, The University of Tokyo,*
Kashiwanoha, Kashiwa, Chiba, Japan

[2]*Department of Information & Management Systems Engineering,*
Nagaoka University of Technology, Nagaoka, Niigata, Japan

Keywords: Memory, Audio Guide, Emotion, Omnidirectional Watching, Information Acquisition, Cognitive Model.

Abstract: The goal of this paper is to examine the possibility of using emotion-induction words in audio guide for the learning of visual contents by extending the study that focused on the provision timings of visual and auditory information (Hirabayashi et al., 2020). Thirty emotion-induction words were extracted from the database and categorized into positive, negative, and neutral words. Three experiments were carried out. The first experiment was conducted to confirm the reliability of the emotional values. The result showed a good consistency between the values on the database and the ratings given by the participants. The second experiment was for examining whether the consistency is maintained if the words appeared in sentences. The result confirmed the expectation but showed larger individual differences compared with the first experiment. The third experiment was conducted to examine the effect of emotion-induction words used in audio guide for explaining the visual contents on memory. The results showed that the participants who were exposed to the positive and negative emotion-induction words, remembered the content better than those who were presented with neutral triggers. Through the three experiments, the emotion value of the neutral words were found to be sensitive to the context in which they were embedded, which was confirmed by observing the changes of pupil diameter. Suggestions for designing audio and visual contents by using emotion-induction words for better memory are provided.

1 INTRODUCTION

Effect of an excess of information on human beings' cognitive processes has been pointed out in a variety of contexts. Simon suggested that an excess of information should cause lack of attention (Simon, 1971). Bitgood focused on the way museum visitors were able to learn given the excess of information they were presented with, pointing out "During museum visits, learners may fail to understand the exhibits deeply because of the abundance of exhibits and time limitations leading to information overload" (Bitgood, 2002). As the development of ICT, the amount of information that can transmit from the artifacts to human beings continues to increase. On the other hand, human beings, who receive the information, are equipped with limited perceptual and cognitive capabilities for processing the rich information. As described above, the excess of information would cause undesirable effect on human beings such as lack of attention, fatigue, and ineffective learning. In order to balance the excess of information with human perceptual and cognitive capabilities, Pierdicca et al. proposed the methodologies that enable the use of both novel IoT architectures and suitable algorithms to derive indicators concerning visitor attention with a significant degree of confidence in the concept of "ubiquitous museum" (Pierdicca et al., 2019).

Lifelong learning has become an important part of our daily life and visiting museum is considered as a typical method for it. The development of ICT has changed the style of exhibition toward the direction of information excess. The more information transmitted from the exhibit does not necessarily mean the more knowledge the viewer acquires due to the problems involved in information excess. In the information excess, learning methods should be designed considering human perceptual and cognitive capabilities.

[a] https://orcid.org/0000-0001-9370-8443
[b] https://orcid.org/0000-0002-0310-2796

Murakami, M., Shino, M., Nakahira, K. and Kitajima, M.
Effects of Emotion-induction Words on Memory of Viewing Visual Stimuli with Audio Guide.
DOI: 10.5220/0010348800890100
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 2: HUCAPP*, pages 89-100
ISBN: 978-989-758-488-6

In the context of lifelong learning, the knowledge that is acquired while appreciating exhibits in museum visit has to be remembered. Hirabayashi et al. took omnidirectional movie as an example of exhibition that utilizes the modern advanced technologies and examined the effect of auditory information presentation timings on memory (Hirabayashi et al., 2020). Their idea was that a stronger memory would result if an auditory information for a particular object is provided after having the viewer attend to it, which is projected on the surface of the dome. For directing the viewer's eyeballs to the object to be explained, they used the appearance of the object to be announced as a part of the audio guide (appearance information). The information for explaining the object (contents information), which is not directly accessible from the appearance of the object, was provided as audio guide after the appearance information using a variety of intervals. They found experimentally that the intervals of 2 $\sim$ 3 seconds was most effective for creating memory of the object. They argued that, in the best presentation interval condition, the visual process and the auditory process that are carried out for comprehending the object should have jointly activated part of long-term memory to generate most richly connected network.

This paper extends Hirabayashi et al.'s study by shifting the focus the research from the richness of the network connection to the contents of the richly connected network. The idea is to strengthen the constituent nodes by manipulating the words used in the audio guide while maintaining the topology of the network generated by processing visual and auditory information by using the best interval of 2 $\sim$ 3 seconds between the timings of provision of appearance information and content information. Namely, on the assumption that generation of richly connected memory network is assured by the best interval condition, this paper investigates the possibility of making the network stronger in terms of the total amount of activation the network holds by manipulating the concrete words used in the contents information.

For this purpose, this paper utilizes the finding that emotion enhances episodic memory by strengthening the constituent nodes. Deborah et al. (Talmi et al., 2019) proposed an extension of the Context Maintenance and Retrieval Model (CMR) (Polyn et al., 2009), eCMR, to explain the way people may represent and process emotional information. The eCMR model assumes that a word associated with emotion, e.g., spider, is encoded with its emotional state in working memory (they called it "context layer") and the presented emotional word establishes stronger link than neutral words. This paper realizes the same effect operationally by attaching a larger strength to the emotional node in the network and examines the effect of emotion-induction words used in audio guide on memory of movie viewing experience. It is likely that viewers' reaction to emotion-induction words may reflect their personal experience or knowledge. Therefore, this paper also utilizes the finding that pupil dilation reflects the time course of emotion recognition (Oliva and Pupil, 2018; Henderson et al., 2018; Partala and Surakka, 2003) to gather the evidence that the manipulation of emotion induction is successful.

This paper is organized as follows. Section 2 describes cognitive framework that shows the effect of timings and contents of audio guide on memory. Section 3 describes three experiments that investigate the effect of emotion-induction words on memory. The first one is for examining the participants response to emotion-induction words, the second one is for sentences with emotion-induction words, and the last one is to measuring memory for movie with positive, negative, and neutral emotion-induction words. The following sections, Section 4 and Section 5, provide the results of the experiments and discusses them from the viewpoint of pupil diameter changes.

2 INTEGRATION OF VISUAL INFORMATION AND AUDITORY INFORMATION

Hirabayashi et al. studied the importance of timing of providing auditory information while watching movies to make the experience memorable (Hirabayashi et al., 2020). This paper extends their findings, i.e., effective provision timing of auditory information for memory formation, by focusing on the effect of emotion-induction words in the auditory information. This section outlines Hirabayashi et al.'s model that explains the effective timing auditory information while processing visual information along with necessary modifications for dealing with the effect of emotion-induction words in auditory information. Starting from the introduction of cognitive model of memory formation, this section discusses why it is essential to take timings into account.

2.1 Memory Formation by Integrating Visual and Auditory Information

Figure 1 illustrates the perceptual and cognitive processes while acquiring visual information with the support of an audio guide, incorporating the find-

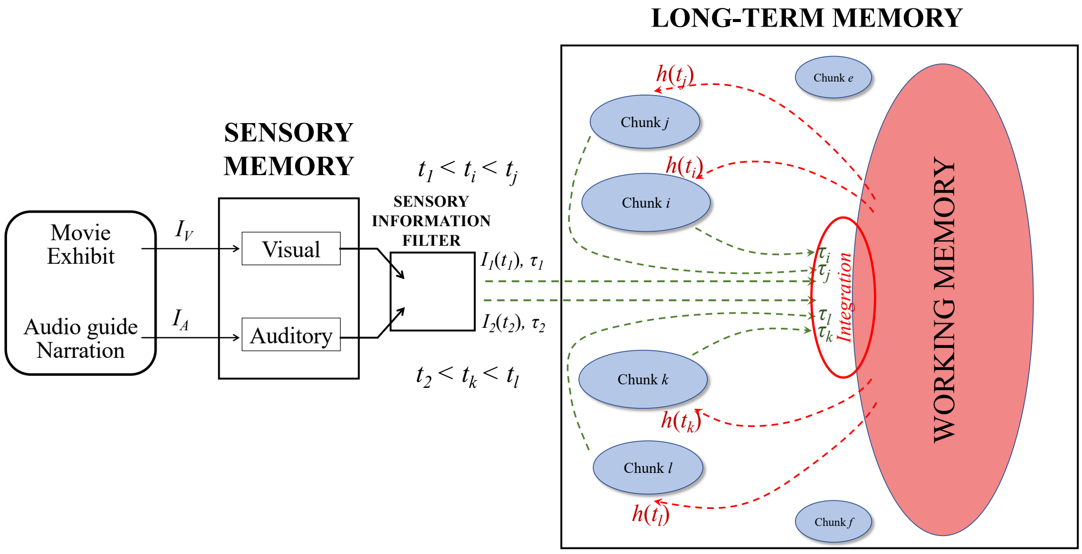


Figure 1: A cognitive model on memory formation.

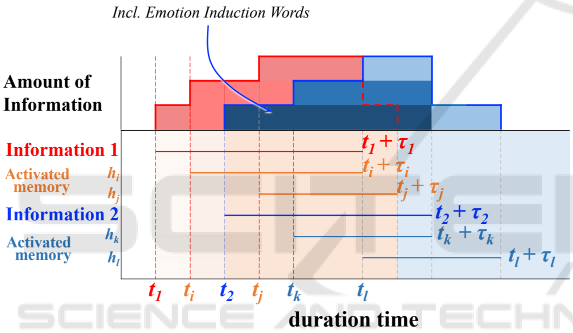


Figure 2: Timeline of memory activation.

ing concerning effective timing of provision of audio guide (Hirabayashi et al., 2020).

In Figure 1, visual information and auditory information are represented as “movie exhibit” and “audio guide narration”, respectively. Consider the situation where visual and auditory stimuli that have the amount of information of I_V and I_A are fed into sensory memory, respectively. Part of the information stored in the sensory memory, I_V and I_A, is passed to working memory (WM) via the sensory information filter as I_1 and I_2 at the time t_1 and t_2, respectively. This paper assumes that $\Delta t = t_2 - t_1 \approx 2 \sim 3$ seconds following Hirabayashi et al.’s finding that their participants showed the best performance of memory for the movie exhibit when auditory information was provided $2 \sim 3$ seconds after the corresponding visual information was provided. Visual information $I_1(t_1)$ and auditory information $I_2(t_2)$ are present in WM for the duration time of τ_1 and τ_2, respectively.

The visual information and auditory information in WM activate part of long-term memory (LTM) via a resonance mechanism (Kitajima and Toyota, 2013). The activated portion of LTM is incorporated in WM, which serves as the next source of activation as long as it exists in WM. In this paper, WM is considered as activated portion of LTM, which is called Long-Term Working Memory (Ericsson and Kintsch, 1995). This process is expressed as follows:

Only Visual Information Is Available $(t_1 \leq t < t_2)$**.**

1. Visual information I_1 is stored in WM at t_1, which is present in WM for the duration time of τ_1.
2. At $t = t_i (> t_1)$, a chunk in LTM C_i is activated by the current contents of WM.
3. The activated chunk C_i is incorporated into WM. The information thus incorporated in WM at t_i is denoted as $h(t_i)$.
4. $h(t_i)$ is present in WM for the duration time of τ_i.
5. During the overlapping period of $(t_1, t_1 + \tau_1)$ and $(t_i, t_i + \tau_i)$, I_1 and $h(t_i)$ serve as the WM contents to activate LTM further. In Figure 1, the chunk C_j is activated at t_j and incorporated in WM as $h(t_j)$ with the lifetime of τ_j

Auditory Information Is Available $(t \geq t_2)$**.**

1. Auditory information I_2 is stored in WM at t_2, which is present in WM for the duration time of τ_2.
2. At $t = t_k (> t_2)$, a chunk in LTM C_k is activated by the current contents of WM.
3. The activated chunk C_k is incorporated into WM as $h(t_k)$ with the lifetime of τ_k.

4. During the overlapping period of $(t_2, t_1 + \tau_2)$ and $(t_k, t_k + \tau_k)$, I_2 and $h(t_k)$ serve as the WM contents to activate LTM further. The chunk C_l is activated at t_l and incorporated in WM as $h(t_l)$ with the lifetime of τ_l

As these processes proceed, the visual information I_1 at t_1 and the auditory information I_2 at t_2 are elaborated through the cascade of activation of chunks in LTM using the dynamically updated contents of WM. The activated chunks are then integrated to make sense of the visual information and the auditory information. In text comprehension research in cognitive psychology, these processes are modeled as the Construction-Integration process (Kintsch, 1998; Kintsch, 1988).

Figure 2 schematically shows the process how the visual information provided at t_1 collects information in LTM by activating relevant chunks at t_i and t_j. The vertical axis represents the number of chunks incorporated in WM. At $t = t_2$, where $t_1 < t_2 < t_1 + \tau_1$, auditory information is incorporated in WM, and collects information in LTM by activating relevant chunks at t_k and t_l. The information originated from the visual information and the auditory information are represented as red rectangles and blue rectangles, respectively. The number of information is the largest between the time $t_k < t < t_1 + \tau_1 = t_i + \tau_i$. It is five, three and two originated from visual and auditory information, respectively.

Assuming that the auditory information provided after the provision of visual information should be used for helping the viewers comprehend the movie better, the overlapping part of the diagram should direct to the common areas of the memory network in LTM. Hirabatashi et al. (Hirabayashi et al., 2020) examined how the intervals between the provision of visual information and auditory information should affect the number of links that could be established through these processes. Through these processes, the activated chunks establish links with the existing memory networks and as a result, it is memorized.

Figure 2 is the best timing for creating a richly connected network for better memory. If $t_2 - t_1$ gets longer or shorter, the area of overlap of the read area and the blue area becomes smaller than the case shown by Figure 2. Therefore, not only the input information itself but also the other pieces of information that is available at the same time plays an important role in how memorable the input information is. Even if the same pieces of information are presented, if the pieces of information are perceived in different timings, it directly affects the quantity of information available for integration and organization of the acquired information.

2.2 Emotion-induction Words for Better Memory

This paper examines the possibility of the effect of the overlapping part of Figure 2 by using emotion-induction words in the auditory information. In Figure 2, the blue part that corresponds to the auditory information is represented by different color values which corresponds the strength of the activated chunks.

According to the theory of cognition, Adaptive Control of Thought - Rational (ACT-R) (Anderson and Lebiere, 1998), the more strong the chunk becomes, the more probable the chunk is retrieved. In ACT-R, the activation level of the i-th chunk, A_i, is defined by the following equation:

$$A_i = B_i + \sum_{j=1}^{N} W_{ji} \times A_j.$$

In this equation, B_i, W_{ji}, and N denote the base level activation of the chunk C_i, the strength of the link between C_j and C_i (from j to i),and the number of chunks that are connected to the chunk C_i. The base level activation decays overtime. But it gets larger when the chunk is used, or activated. This paper assumes that emotion-induction words should activate stronger chunks than emotionally neutral words. Figure 2 depicts the situation where the auditory information contains emotion-induction words and stronger chunks are activated and incorporated in WM. This should cause stronger memory trace than neutral words are used in the auditory information.

3 THREE EXPERIMENTS FOR INVESTIGATING EFFECTS OF EMOTION-INDUCTION WORDS

Three experiments were conducted to understand the relationships between emotion-induction words and memory when watching movie. Figure 3 outlines the relationships between the experiments. The first experiment (Word Impression Evaluation Experiment, Exp-W) was conducted to confirm the reliability of the emotional values. The second experiment (Sentence Impression Evaluation Experiment, Exp-S) was for examining whether the consistency is maintained if the words appeared in sentences. The third experiment (Video Appreciation Experiment, Exp-V) was conducted to examine the effect of emotion-induction words used in audio guide for explaining visual contents on memory. Thirty participants took part in all

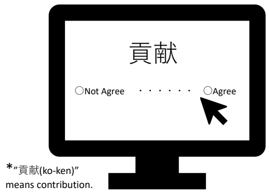
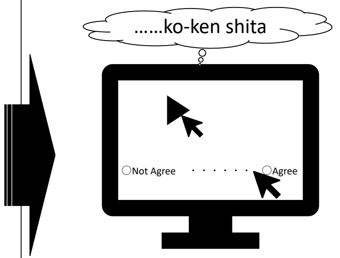
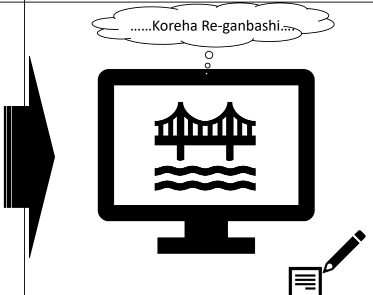
	Experiment 1 : Word Impression Evaluation	Experiment 2 : Sentence Impression Evaluation	Experiment 3 : Video Appreciation
purpose	confirming participants' characteristic for affective value and check the discrepancy between affective value DB and participants' feature	confirm emotion-induction words effects for one sentence	find out the role of emotion-induction words in the viewpoint of memory
content	word	sentence	movie
conditions	participant's PC at their home	PC at laboratory	display and speaker at laboratory
measuring data	impression evaluation	impression evaluation and pupil diameter	impression evaluation, pupil diameter, and memory test
stimuli	visual	audio	visual + audio
stimuli control	by participant	by participant	by operator
stimuli outline	 貢献 ○Not Agree · · · · · · ○Agree *"貢献(ko-ken)" means contribution.	 	
how to response	mouse click	mouse click	Impression : oral communication Memory : after video watching, write-down

Figure 3: The outline of three experiments flow.

experiments. They first participated in Exp-S and Exp-V consecutively scheduled on a single day in the laboratory. After an interval of about 10 days, they participated in Exp-W on their own PC's. Different sets of induction words were used in Exp-S and Exp-V. Exp-W was conducted after Exp-S and Exp-V for the purpose of checking the appropriateness of the emotion-induction words used in the sentences and audio guides presented in Exp-S and Exp-V.

3.1 Measurement Data

Impression Rating. In this study, emotion-induction words were used to investigate the influence of emotion on memory. All of them were taken from the database (Gotoh and Ohta, 2001) being created as the following way: 618 participants were asked to rate their impressions to the presented words by using a 7-point Likert scale as follows: that were "very positive," "positive," "somewhat positive," "neither," "somewhat negative," "negative," and "very negative," After the rating experiment, a database of 389 words consisting of 122 negative, 146 positive, and 121 neutral valence words were constructed.

In this study, impression ratings were collected for words in Exp-W, sentences in Exp-S, and videos in Exp-V. The options for rating included the ones used for constructing the above-mentioned database and a new option, "I cannot catch the mean of this word/sentence/video." The last one was added considering the possibility of not knowing the word or not being able to get the meaning of the sentence.

Pupil Diameter. It is known that pupil diameter dilates when emotions move (Oliva and Pupil, 2018; Henderson et al., 2018; Partala and Surakka, 2003). From this reason, the participants' pupil diameters were measured in Exp-S and Exp-V. Tobii Pro2 with the sampling rate of 50 Hz was used as the equipment.

Memory. Memory was measured in Exp-V. A questionnaire was conducted to investigate the information that participants memorized when viewing movies. Since the questionnaire was conducted soon after viewing the movie, a recall test was chosen. To make the quantitative evaluation of memory, the responses from the participants were broken down to meaningful units using a morphological analysis technique. They were scored from the quantitative and qualitative perspectives, by giving special points to those units related to the targets and the narrative contents spoken in the audio guides (one point was given to a noun or a verb, two points to a pronoun). Those points were summed up to define "Memory Score." Table 1 shows an example of the responses from Exp-V and the calculated Memory Score.

Table 1: Examples of Memory Score (n/v:noun or verb, pn: pronoun, MS: Memory Score (points)).

Response	n/v	pn	MS
Statue	1	0	1
Statue of Messenger	0	1	2
replica of Statue of Messenger	1	1	3
Statue of Messenger was given as a proof of friendship	2	1	4

3.2 Word Impression Evaluation Experiment (Exp-W)

The purpose of Exp-W was to confirm that there was no discrepancy between the emotional value of the two-character words in the Japanese language in the database and the evaluations of the participants. Thirty participants evaluated their impressions of the words displayed on their own PC's by using the 7-point Likert scale. The number of words was thirty. Considering the order effect on evaluation, multiple patterns were randomly created, and experiments were conducted on 5 patterns for thirty words.

3.3 Sentence Impression Evaluation Experiment (Exp-S)

The purpose of Exp-S was to confirm that emotion-induction words affect emotion even when they are presented in sentence. The procedure was exactly the same as Exp-W except for the stimuli were sentences instead of words and presented in audio, and the location of the experiment.

The number of sentences was thirty and they were randomly presented as Exp-W in the laboratory using a laptop computer with a 13-inch display. The participants operated the touch pad of the laptop computer. Ten negative words, ten positive words, and ten neutral words were extracted from the database. Each sentence was a single sentence containing one of the extracted word. Each sentence was a single sentence lasting about 3-5 seconds, and was broadcast twice, read out loud by the same person in succession. The break between the two was clear. The participants listened to the thirty sentences and evaluated their impressions. The sentences were not shown visually on the display.

3.4 Video Appreciation Experiment (Exp-V)

Exp-V was performed in an appreciation environment similar to learning at a social education institution. Thirty participants took part in the experiments and no one had visual or health problem on taking the experiment. They watched three movies consecutively and evaluated impression of each movie using the 7-point Likert scale. Each movie was provided with audio guide that contained a few number of positive, negative, or neutral words extracted from the database. Memory test was conducted immediately after the viewing the movies by having the participants write anything they remembered. In Exp-V, biological information was collected as well as a potential objective indicator that should reflect the effect of emotion-induction words on the psychological states of the participants.

Table 2: Three presentation patterns for audio guides of three movies that contain positive, negative, and neutral emotion-induction words.

Pattern	Movie 1	Movie 2	Movie 3
1	Neutral	Positive	Negative
2	Negative	Neutral	Positive
3	Positive	Negative	Neutral

Exp-V was conducted in the laboratory. The participants operated the touch pad of the laptop computer. The participants evaluated their impressions of the sentences in the same way as Exp-W after watching three movies. Three patterns were considered for the presentation order of positive, negative, and neutral audio guide as shown in Table 2.

Visual and Auditory Stimuli: Three movies were prepared. Each movie had a respective audio guide with one of the three conditions of attributes of emotion-induction words. The movies showed the landscape taken from a slow-paced boat going down the Sumida River in Tokyo. A movie taken from a slow-paced boat was chosen as a stimulus for this experiment because it is likely to contain scenes or targets that satisfy the following conditions:

1. The target in the movie should move in a slow pace. This condition was needed to make the target appear and stay in the field of view long enough for a viewer to take needed visual information of the target object.
2. The target should not be easy to notice without a guidance. This condition was needed to refrain viewers from paying attention to the target beforehand and to see the effect of audio guide clearly.
3. Scenes should contain many objects to look at throughout the movie. This condition was needed to simulate the situations where audio guide is in need.

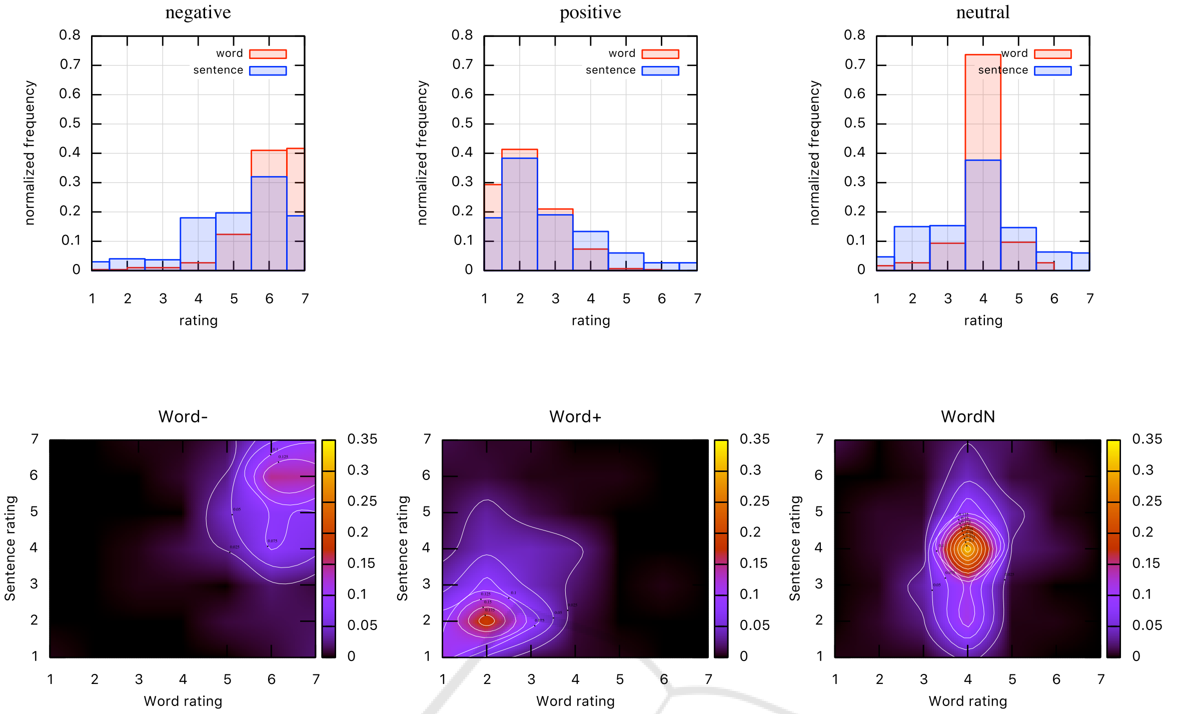


Figure 4: Upper three graphs : frequency of ratings for “word” and “sentence”. Left side is for negative, middle is for positive, right is for neutral mode.Lower three graphs : probability distributions for “word” versus “sentence”. Word $-$ is for negative, Word $+$ is for positive, and WordN is for neutral.

Experiment System. Viewing behavior including pupil diameter was recorded using a wearable eye tracker (Tobii Pro Glasses 2) at a sampling rate of 50 Hz. The experiment was carried out with one participant at a time. The participants were seated on a chair and their head positions were located approximately 0.8 meters from the display.

Procedure. Before viewing the movie, participants were told to make themselves comfortable and to view the movie freely in order to simulate the actual viewing behavior. Each movie was approximately one minute long, and intervals were inserted between the movies. Considering the order effect, each participant was presented with a randomly chosen pattern from the three shown in Table 2. After the participants finished viewing the movies, they were asked to complete the questionnaire.

4 RESULT

This section starts by showing the results of Exp-W and Exp-S concerning impression evaluation of words that appeared in isolation and in context. It is followed by the results of Exp-V concerning memory score for the movies that used positive, negative and neutral emotion-induction words.

4.1 Impression of Emotion-induction Words

In the top half of Figure 4, normalized frequency in Exp-W and Exp-S are shown. At a single glance, the ratings for emotion-induction words that appeared in isolation or in sentence were consistent irrespective of the nature of the words, i.e., positive, negative, or neutral. For the negative words, the mode of the ratings for the negative words in isolation and in sentence was 6, that for the positive words was 2, and that for the neutral words was 4.

The results shown by histograms are further decomposed by focusing on the ratings for individual words. Each word was rated in isolation and in sentence. The bottom half of Figure 4 shows the normalized frequency of the data points of the two dimensional space, i.e., rating of word in isolation vs. rating of word in sentence. The number of data points is 300 (10 words by 30 participants) for negative, positive, and neutral conditions. As shown in the figures, the ratings for the emotion-induction words were consistent whether they were rated in isolation or in sentence in general. However, the neutral

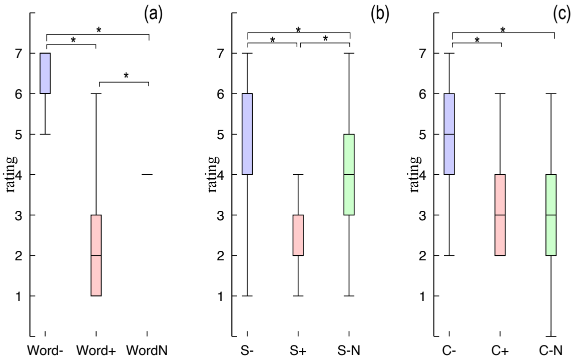


Figure 5: Box-plots of ratings for (a) emotion-induction words in Exp-W, (b) sentences that emotion-induction words are embedded in Exp-S, and (c) visual contents with audio that includes emotion-induction words in Exp-V. The signs '+', '-', and 'N' stand for 'positive', 'negative', and 'neutral', respectively.

emotion-induction words showed a significant difference in terms of the degree of variance in the sentence rating, as shown by the bottom-right plot in Figure 4. This indicates that the neutral emotion-induction words were rated as neutral when they were presented as single words in isolation but the rating of emotion value had fluctuated when they were presented in sentence. This suggests that the neutral emotion-induction words cannot be emotionally neutral when they appear in sentence.

Box-plots of the ratings for the emotion-induction words, the sentences that the emotion-induction words are embedded, and the visual contents with audio that includes the emotion-induction words are shown in Figure 5 (a), (b) and (c). The ratings were significantly different across conditions, i.e., positive, negative, and neutral except for the ratings between neutral and positive emotion-induction words in Exp-V. More specifically, by comparing the results of Exp-V shown in Figure 5 (c) with the results of Exp-W and Exp-S, it is found that only in negative condition, the impression ratings were in accordance with the attributes of the emotion-induction words, and the impression ratings for the positive and neutral conditions were not consistent with the attributes of the emotion-induction words. The ratings for the positive condition in Exp-V shifted to the region of neutral impression, and those for the neutral condition shifted to the region of positive impression. This observation is consistent with the result shown by the bottom-right plot in Figure 4. It is likely that the ratings of emotion value had fluctuated when the emotion-induction words were presented in context.

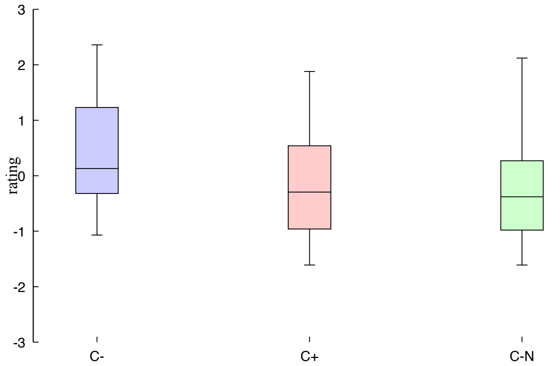


Figure 6: Memory scores for the three movie categories; the movies in the C-, C+, and C-N categories consisted of the negative, positive, and neutral emotion-induction words, respectively.

4.2 Memory Score

A standardized summary of memory score in each video is shown in Figure 6. Using *t* test, a statistically significant difference was seen between C- and C+ conditions and between C- and C-N conditions. However, there was no statistically significant difference between C+ and C-N conditions. This result is parallel with the the result concerning Figure 5 (c) for Exp-V described in Section 4.1. In the condition where the emotion-induction words appeared in audio guide for providing information of the movies, it is likely that positive and neutral words were more affected by the context where they appeared and they could change their context-free emotion values, which were maintained even if they appeared in sentences.

5 DISCUSSION

This section starts by analyzing the results of Exp-V concerning relationship between emotion-induction words included in movies and memory score. It follows pupil diameter dilation analysis, as an objective indicator of emotional changes. Finally, this section ends by discussing the possibility of implementing a design method based on the above-mentioned two points.

5.1 Memory Score Analysis

Section 4 showed that 1) the memory scores were high in the negative condition, and 2) the attributes of emotion-induction words in the impression evaluation were only maintained i the negative condition. As shown in Figure 4, the emotion-induction words should also work in sentences regardless of they are

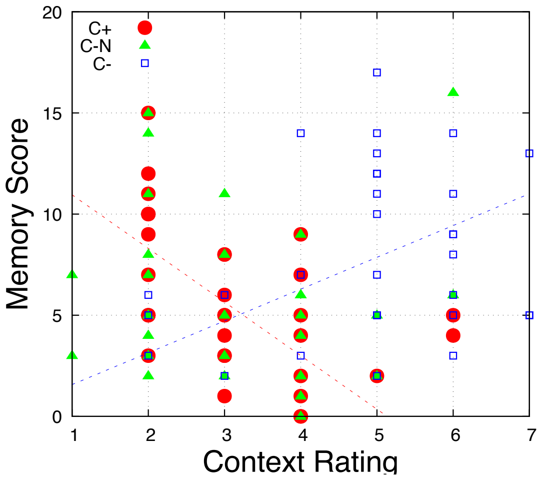


Figure 7: Relation between context (movies) rating for emotion and memory scores. C+ is for the movies which were constructed with the positive emotive-induction words. C- is for the movies which were constructed with the negative emotive-induction words. The dotted lines are linear regression for each movie. The blue dotted line is for C- with the R^2 value of 0.79. The red dotted line is for C+ with the R^2 value of 0.33.

positive or negative. When they were added as an audio guide to the video, it is likely that for some reason the positive emotion-induction words did not manifest their expected effect when added to the video. Therefore, it is not possible to discuss the relationship between positive emotion-induction words and memory. Further analysis is needed in the future. However, the results suggest that negative emotion-induction words also influenced the evaluation of impressions and contributed to the improvement of memory in the movies.

Next, the relationship between impression evaluation and memory scores is discussed. As shown in Figure 1 and Figure 2, the presence of emotion-induction words in a sentence should enhance memory by overlapping visual information as objects with auditory information at appropriate times. In addition, as shown in Figure 4, the influence of neutral words on emotion varied more when they were presented in sentence than when they appeared in isolation. This suggests that this tendency is more pronounced when the words are presented in context.

The relationship between the ratings and the memory scores for each of the three videos obtained from the participants is shown as a scatter plot in Figure 7. The dotted lines are linear regression for each movie. The blue dotted line is for C- with the R^2 value of 0.79. The red dotted line is for C+ with the R^2 value of 0.33. The blue dotted line increases as ratings become large. This means memory score becomes large as the movie in C- condition were rated in accordance with the negative emotion-induction words that appeared in audio guide. Similarly, the red dotted line decreases as ratings become large. This means memory score becomes large as the movie in C+ condition were rated in accordance with the positive emotion-induction words that appeared in audio guide. When the emotion-induction words used for providing explanation of the movies in audio guide function as they were expected, memory scores should increase. On the other hand, when they don't function as they should do, memory scores should not become large.

5.2 Pupil Diameter Analysis

To understand the effect of emotion-induction words included in audio guide on memory, pupil diameter changes in Exp-S was analyzed as an objective evaluation index. To analyze the relationship between the attributes of emotion-induction words which each participants heard and pupil diameter changes, the participants were screened according to the following conditions.

1. There is no problem in the ratings in Exp-W and Exp-S.
2. There is no problem in the pupil diameter data.

Of those who satisfied these conditions, two participants were selected for a preliminary analysis of pupil data considering the degree of accordance with their ratings for the sentences used in Exp-S with the attributes of the sentences, i.e., S+, S-, and S-N. One participant, ut33, showed the highest correlation of 0.92 (consistent-participant) and the other, ut19, showed the lowest correlation of 0.18 (inconsistent-participant). Since the pupil diameter dilation to emotional changes is smaller than the change to light changes, the dilation to emotional changes is considered to be captured by lower envelope for pupil diameter changes. Figure 8 is a graph showing the approximate lines of the lower envelopes. The difference between the value of pupil diameter curve and the value of the lower envelope was added by the time when the sentence was played, and the average was calculated by using the following formula:

$$\frac{1}{n} \times \sum_{i=1}^{n} (T(t_i) - L(t_i)),$$

where T is the trajectory, L is the lower envelope, t_1 is the start time of the i-th sentence in audio, and t_n is its end time.

Figure 9 shows pupil size changing distribution for ut33 and ut19. This result indicates that ut33 has a difference in the effect of emotions between emotion-induction words and neutral words, and ut19 has a small difference between emotion-induction words

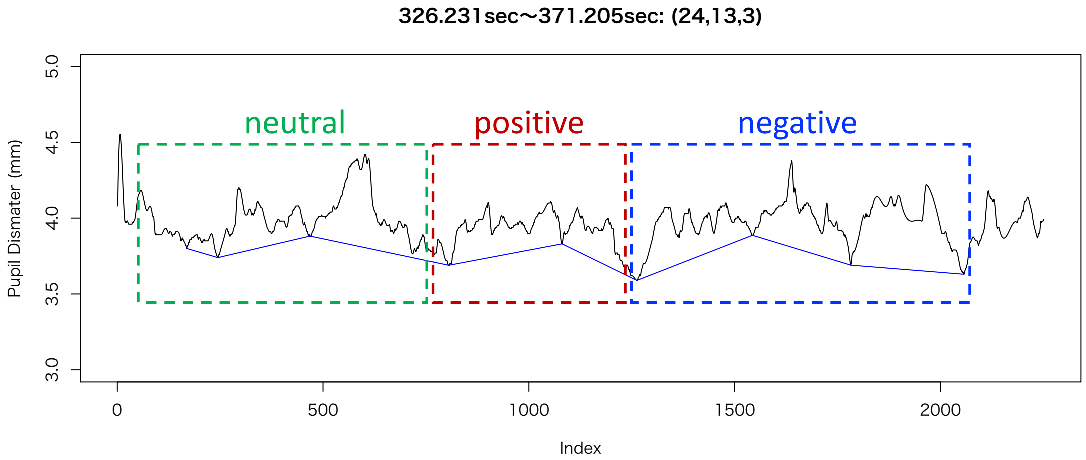


Figure 8: The black lines represent the trajectory of the size of pupil diameter. The horizontal axis is the sampling number (500 corresponds to 10 seconds). The widths of the three rectangles correspond to the duration of the presentation of sentences with neutral, positive, and negative emotional values. The blue solid lines represent the lower envelope for the trajectory.

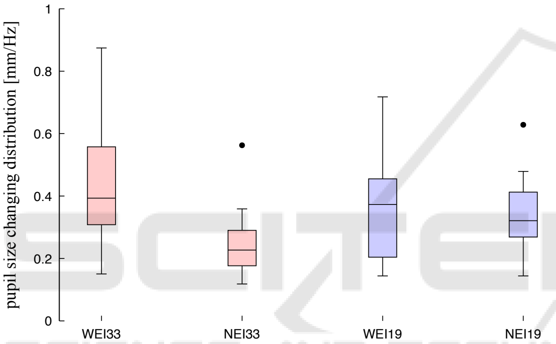


Figure 9: Participants' pupil size changing distribution for ut33 (2 red box-plots to the left) and ut19 (2 blue box-plots to te right). WEI stands for "emotional-induction words included", NEI stands for "emotional-induction words not included (i.e., neutral words)" .

and neutral words. This result corresponds to the characteristic of correlation with the database. In other words, the pupil diameter variability was greater for those who rated impressions according to the database when they heard emotion-induction words, while those who rated impressions not according to the database showed no difference in pupil diameter variability when they heard emotion-induction words compared to when they heard neutral words. The results suggest that there are individual differences in the emotional changes caused by emotion-induction words, which can be captured by pupil diameter analysis. Further pupil diameter analysis will support this finding.

5.3 Design Implications

This subsection discusses designing for multi-modal information for better memory. First, as showed in Figure 4, it is found that impression ratings corresponded to the attributes of emotion-induction words. Second, it is known that emotional changes are captured by pupil diameter dilation. As shown in Figure 9, for the same stimuli, the consistent-participant and the inconsistent participant showed significantly different reaction in terms of pupil diameter reactions. This indicates that the measure of pupil diameter could be used to monitor how the participant reacted to the stimuli. As discussed in this paper, ratings of impression, which could be subject to individual differences as evidenced by the existence of inconsistent participants, should correlate with memory score. Monitoring participants emotional state could be a promising method for designing auditory information for helping construct better memory. In addition, as shown in Figure 6, memory scores should be higher when negative emotion-induction words were acquired. This finding is also applicable to designing auditory information.

6 CONCLUSION AND FUTURE WORK

This paper investigated the effect of emotion-induction words in audio guide on memory. As a result, it was found that the auditory information with negative emotion-induction words is easy to remember. It was also suggested that emotional changes during appreciation behavior might be measured by pupil diameter.

For now, this study only focused on the two-dimensional movie viewing behavior to simulate the real environment. However, for applying it to our everyday life, such as museum, gallery, guided tours, etc., it is important to apply and examine what this paper found in the omnidirectional situations such as a dome theater. Also, pupil diameter analysis was done for only two participants to examine the feasibility of the research direction. The results were promising. We plan to continue this approach and try to find the way to objectively estimate the viewer's cognitive state that should enhance learning of visual contents accompanied with auditory information.

ACKNOWLEDGEMENTS

This work was partly supported by JSPS KAKENHI Grant Number 19K12246 and 20H04290.

REFERENCES

Anderson, J. R. and Lebiere, C. (1998). *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, NJ.

Bitgood, S. (2002). Environmental psychology in museums, zoos, and other exhibition centers. In *In Handbook of environmental*, pages 461–480.

Ericsson, A. K. and Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102:221–245.

Gotoh, F. and Ohta, N. (2001). Affective valence of two-compound kanji words. *Tsukuba psychological research*, 23(23):45–52. in Japanese.

Henderson, R. R., Bradley, M. M., and Lang, P. J. (2018). Emotional imagery and pupil diameter. *Psychophysiology*, 55(6):e13050.

Hirabayashi, R., Shino, M., Nakahira, K. T., and Kitajima, M. (2020). How auditory information presentation timings affect memory when watching omnidirectional movie with audio guide. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 2: HUCAPP,*, pages 162–169. INSTICC, SciTePress.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review*, 95:163–182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press, Cambridge, UK.

Kitajima, M. and Toyota, M. (2013). Decision-making and action selection in Two Minds: An analysis based on Model Human Processor with Realtime Constraints (MHP/RT). *Biologically Inspired Cognitive Architectures*, 5:82–93.

Oliva, M., A. and Pupil, A. (2018). dilation reflects the time course of emotion recognition in human vocalizations. *Scientific Reports*, 8:4871.

Partala, T. and Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, 59:185–198.

Pierdicca, R., Marques-Pita, M., Paolanti, M., and Malinverni, E. S. (2019). Iot and engagement in the ubiquitous museum. *Sensors*, 19(6):1387.

Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, 116:129–156.

Simon, H. A. (1971). Designing organizations for an information rich world. In Greenberger, M., editor, *Computers, communications, and the public interest*, pages 37–72. Johns Hopkins University Press, Baltimore.

Talmi, D., Lohnas, L. J., and Daw, N. D. (2019). A retrieved context model of the emotional modulation of memory. *Psychological Review*, 126:455–485.

APPENDIX

The emotion induction words and the sentences used for the experiment.

No.	Word	Score	Sentence including the word
Positive			
1	[murder]	6.51	Murder is one of the worst sins a person commit.
2	[tragedy]	6.38	A tragedy happened in the baseball tournament final.
3	[assassinate]	6.36	Ryoma Sakamoto was assassinated on his birthday.
4	[banish]	6.33	Unable to understand the situation, he was banished from the dinner party.
5	[prejudice]	6.15	It is difficult to be aware of unconscious prejudice.
6	[transfer school]	6.14	The first day of transfer school is full of anxiety.
7	[worst]	6.13	Enter the site assuming the worst situation.
8	[dismiss]	6.13	One of the college students working part-time must be dismissed.
9	[penalty]	6.10	Nothing is as boring as paying a penalty.
10	[fall]	6.06	I have fallen since I became a college student.
Neutral			
1	[free]	1.93	Rice is often included free of charge at Iekei ramen shops.
2	[contribute]	1.93	How much do I contribute to the sales of the nearest convenience store?
3	[love]	1.90	I cook curry with plenty of love.
4	[experienced]	1.87	Rookie is good, but experienced veteran is also good.
5	[courage]	1.80	I gave up my seat with courage.
6	[holiday]	1.77	When I was in elementary school, I loved holidays.
7	[plenty]	1.53	There are plenty of drink bars here, so I'll follow you.
8	[fortunate]	1.50	I was fortunate to meet you.
9	[achieve]	1.47	I am good at achieving goals one by one.
10	[clear day]	1.47	Laundry progresses on a clear day.
Negative			
1	[field]	4.0	I think I know more about this field than most people.
2	[loading platform]	4.00	I'm watching the cardboard boxes pile up on the loading platform.
3	[railway route]	4.00	There are so many railway routes in Tokyo that you can't compare to the countryside.
4	[job seeker]	4.07	Job seekers were in line.
5	[address]	3.97	When I write my address, I'm wondering whether to write it from the prefecture.
6	[jacket]	3.93	It is difficult to choose a jacket because the temperature difference between day and night is large.
7	[seal]	4.11	I always carry my seal.
8	[budget]	4.07	You can't decide anything else unless you decide on a budget.
9	[next time]	4.1	Next time I will try to order a different menu.
10	[clerk]	3.98	Turn right at the end and a clerk is standing.