

The Easy Scoring System: A Tool for Scoring Subjective Impressions of Content

Yasushi TAKAHASHI*, Mikio KAMADA**

* Chiba Institute of Technology C/o Department of Design 2-17-1 Tsudanuma, Narashino-shi, Chiba 275-0016 Japan, Takahashi.Y@pf.it-chiba.ac.jp

** Broadband Applications Laboratories, Sony Corporation 6-7-35 Kitashinagawa, Shinagawa-ku, Tokyo, 141-0001 Japan, micky@aic.sony.co.jp

Abstract: This is a report of our developing and evaluating the easy Scoring System (eSS), a system for easily collecting subjective impressions of content such as TV programs. In the information age where we are inundated with vast amounts of information, the chronological externalization and description of people's reactions, actions, and values vis-à-vis content is essential to the selection and summarization of information suited to individuals as well as interpersonal communication about content. This system offers one method for effectively and efficiently describing such information. There has not been a tool for easily entering and automatically tabulating evaluations. That is why we have developed the easy Scoring System (eSS), a simple evaluation system that uses a server to centrally manage and tabulate data and enables the easy entry of evaluations anytime, anywhere, using mobile phones. We have derived terms for impressions appropriate to evaluating various kinds of content and have verified the effectiveness of the eSS through real-time impression entry tests that employ an entry method for any point in time regardless of scene divisions for broadcasted TV programs. Based on the results of having test subjects living in the Tokyo area watch a 7 PM variety program while relaxing and entering their impressions into their mobile phones, we discovered the following. First, the eSS system can collect accurate impression data without burdening test subjects. Second, the impressions "amusing," "boring," "huh?," and "aha!" are effective in ascertaining the characteristics of variety programs. Third, peak data for impression frequency on a time scale is effective in extracting program highlights. Finally, principal component analysis, in which the number of entered impressions for segmented plots of the content is used as data, can identify test subject clusters that have the same interest in program content.

Key words: impression, evaluation, content, TV program, mobile phone

1. Introduction

This is a report of our developing and evaluating the easy Scoring System (eSS), a system for easily collecting subjective impressions of content such as TV programs. In the information age where we are inundated with vast amounts of information, the chronological externalization and description of people's reactions, actions, and values vis-à-vis content is essential to the selection and summarization of information suited to individuals as well as interpersonal communication about content. This system offers one method for effectively and efficiently describing such information.

The protocol method[1], in which experiences are freely uttered and recorded, has been used thus far for describing people's experiences along a time scale. This method, however, is qualitative and depends on the nature of the utterances, and is thus unable to statistically process many tests. The semantic score (SS) method

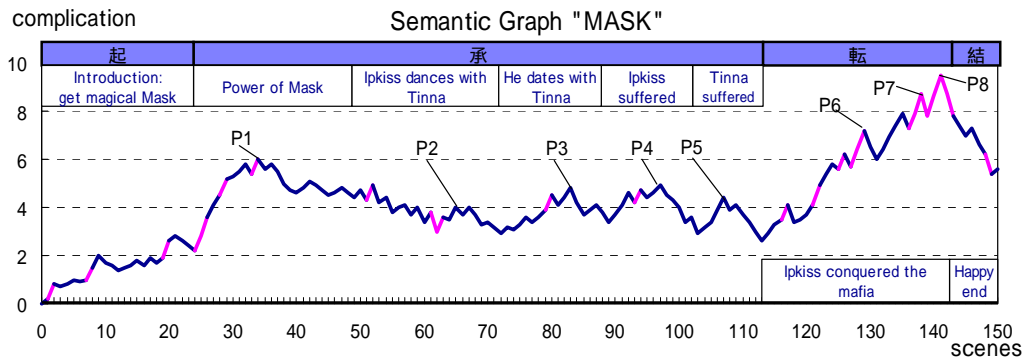


Fig. 1 Story Structure of the Movie “Mask” Integrated from Its Semantic Scores

[2][3][4] is a quantitative technique that uses an evaluation scale for complication and resolution to perform subjective evaluations of audio/visual experiences of story-based content such as movies quantitatively. (Figure 1) This evaluation method, however, is limited by the fact that it evaluates content on a scene-by-scene basis. Whether the same method of evaluating complication and resolution is applicable to non-story content has yet to be verified. There has not been a tool for easily entering and automatically tabulating evaluations. That is why we have developed the easy Scoring System (eSS), a simple evaluation system that uses a server to centrally manage and tabulate data and enables the easy entry of evaluations anytime, anywhere, using mobile phones. We have derived terms for impressions appropriate to evaluating various kinds of content and have verified the effectiveness of the eSS through real-time impression entry tests that employ an entry method for any point in time regardless of scene divisions for broadcasted TV programs. Based on the results of having test subjects living in the Tokyo area watch a 7 PM variety program while relaxing and entering their impressions into their mobile phones, we discovered the following. First, the eSS system can collect accurate impression data without burdening test subjects. Second, the impressions “amusing,” “boring,” “huh?,” and “aha!” are effective in ascertaining the characteristics of variety programs. Third, peak data for impression frequency on a time scale is effective in extracting program highlights. Finally, principal component analysis, in which the number of entered impressions for segmented plots of the content is used as data, can identify test subject clusters that have the same interest in program content.

2. Derivation of Evaluation Terms for Content Impressions

The semantic score method has thus far been limited to story-like content such as movies and TV dramas. In our latest effort, we decided to examine evaluation terms for impressions that could also support a variety of other media content, primarily TV programs.

Using TV guides and program rating columns as our source, we started by deriving pairs of terms that express impressions toward content. The eight pairs we obtained after organizing the terms using the KJ Method were “like/dislike,” “happy/sad,” “beautiful/ugly,” “amusing/boring,” “cute/not cute,” “good/bad,” “approve/disapprove,” and “meaningful/meaningless.” In addition to these, we added “complication/resolution,” the rating scale we have used to date, and “keep/discard,” which questions whether the content itself is desirable, for a total of 10 pairs. We employed paired concepts in order to elucidate the chronological evaluation structure, just as in

the semantic score method. In contrast to the Semantic Differential (SD) method, which employs numerous adjective pairs to statically structure a target in multidimensional psychological space and is widely used in psychological evaluations, we aimed to narrow down as much as possible the number of evaluation terms to make it easy to evaluate impressions in real time while watching content, thereby revealing chronological changes.

To find out which evaluation terms suited content in a wide variety of genres, we showed test subjects video clips about 3 minutes in length sampled from 12 programs with high viewer ratings. The eight genres comprising the programs were news/reports (3 programs), sports (1 program), documentaries (2 programs), variety shows (2 programs), dramas (1 program), animated cartoons (1 program), music (1 program), and image videos (1 program). We then examined which of the previously mentioned 10 pairs and other terms applied. Multiple answers were allowed for each title, and test subjects were instructed to mark the most suitable answer with a double circle symbol. There were 10 test subjects in all, comprised of university students and faculty.

Tabulated results for impression terms disjunctively selected for each program revealed that “amusing/boring” was the most prominent, followed by “complication/resolution” and “meaningful/meaningless.” (Figure 2) In the results for multiple responses, “amusing/boring” was also the most prominent, followed by “complication/resolution” and “meaningful/meaningless,” as well as “like/dislike” and “happy/sad.” (Figure 3)

We carried out analyses based on quantification method type 3 using the multiple response data for the 10 impression terms. Figure 4 is the result of cluster analysis using an up to 2-axis solution determined from an eigenvalue graph. This revealed Group 1, in which “happy/sad,” “beautiful/ugly,” “good/bad,” “keep/discard,” and “approve/disapprove” are all in close proximity, centered around “like/dislike” and “amusing/boring.” In contrast, “complication/resolution” is the impression term most divergent from the rest, followed by “cute/not cute” and “meaningful/meaningless.” Figure 5 shows the arranged cluster analysis results for the category scores of quantification method of type 3. From this figure, we can see that on axis 1, plus (+) is the process and minus (-) is the result, and on axis 2 plus (+) indicates the meaning of content for the subject himself, while on the other hand, minus (-) places greater weight on the target. As you can clearly see in Figure 2 and 3, “amusing/boring” was selected the most for both single and multiple responses, and thus in Group 1, it forms the core of the group and is thought to play a role in expressing the concepts of the other impression terms.

Based on the observations we have discussed thus far, it is clear that “amusing/boring,” “complication/resolution,” “cute/not cute,” and “meaningful/meaningless” hold importance, in that order, as impression terms for a variety of genres.

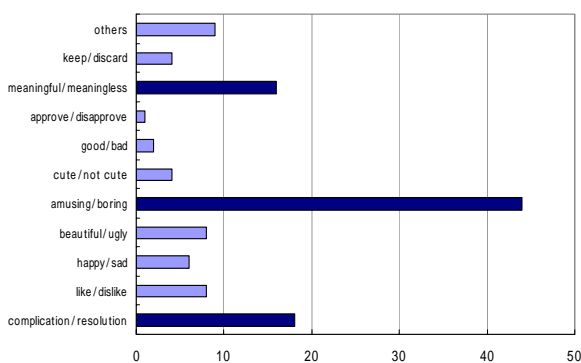


Fig. 2 Tabulation of Single Answers for Impression Terms

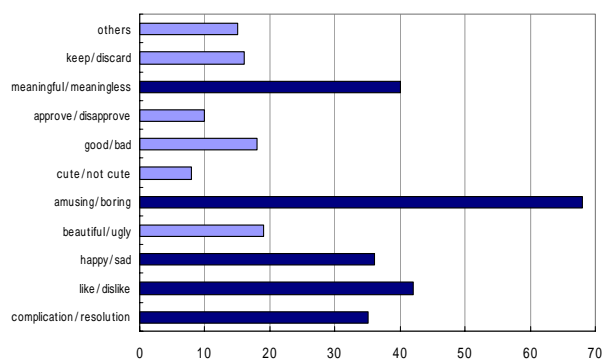


Fig. 3 Tabulation of multiple Answers for Impression Terms

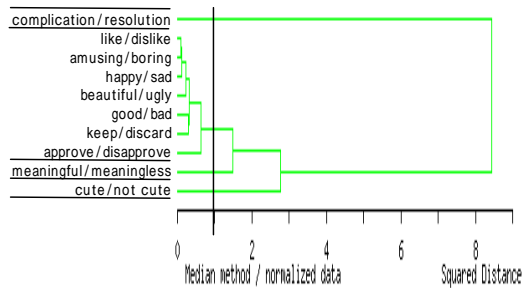


Fig. 4 Clusters Generated from the Category Scores of Quantification Method of Type 3

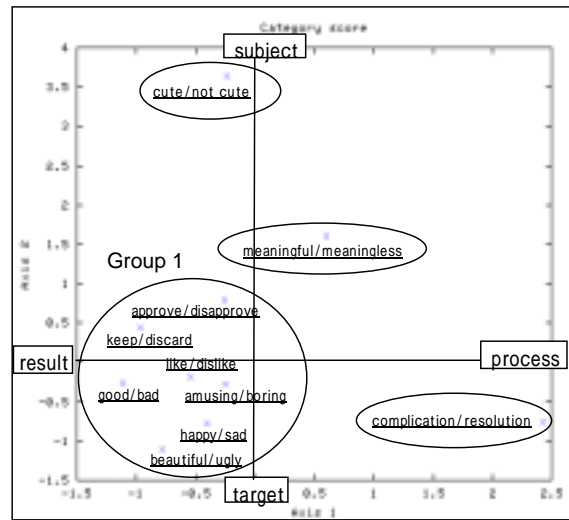


Fig. 5 Arranged Cluster Analysis Results for the Category Scores of Quantification Method of Type 3

3. Development of eSS and Real-time Impression Evaluation Test for Broadcast TV Programs

3.1. Development of the Easy Scoring System (eSS)

An individual’s mental experience is wide ranging. It begins with watching content and then diverges to interaction with a device/system and communication with other people. Chronologically recording such life experiences requires that individuals are always able to easily enter their impressions anytime, anywhere, through a handheld device and that the resulting data can be transmitted. The market penetration of mobile phones has exceeded 58% in 2002 in Japan, and they are the perfect information terminal for our purposes. We decided to develop HTML programs for three major carriers and to set up a Web site that can be accessed from mobile phones and a server that can collect data in real time from all over the country. Using the ten-key pad on a mobile phone, from one to three pairs of evaluation terms for impressions can be assigned. (Figure 6-1) It is also possible to enter a three-level evaluation value when evaluation terms can be narrowed down to one pair. (Figure 6-2) People who are well accustomed to the ten-key pad on their mobile phones can enter their impressions without even looking at the key pad, thereby enabling intuitive entry while viewing content.

For our first test of evaluating a broadcast program in real-time, we set two pairs of evaluation items (“amusing/boring” and “huh?/aha!”), developed HTML programs for NTT DoCoMo i-mode, J-Phone J-Sky, and au EZweb mobile devices with the operation flow shown in Figure 7. The key assignment was 1 for “amusing,” 3 for “boring,” 7 for “huh?,” and 9 for “aha!”(Figure 8) After the impression evaluation was completed, the LCD on the device prompted test subjects to respond to a simple questionnaire and enter their user profile. The data log for impression evaluations collected by the server recorded the test subject ID, impression key numbers, and the time of entry (in seconds). A pre-test inspection verified that the communication time delay was around 1 second for the most part with 3 seconds in the slowest of cases and that when test subjects repeatedly pressed keys within 1 second, only the first key press was transmitted.



Fig. 6-1 An Example of Three Pairs of Impression Terms



Fig. 6-2 An Example of a Three-level Evaluation Value

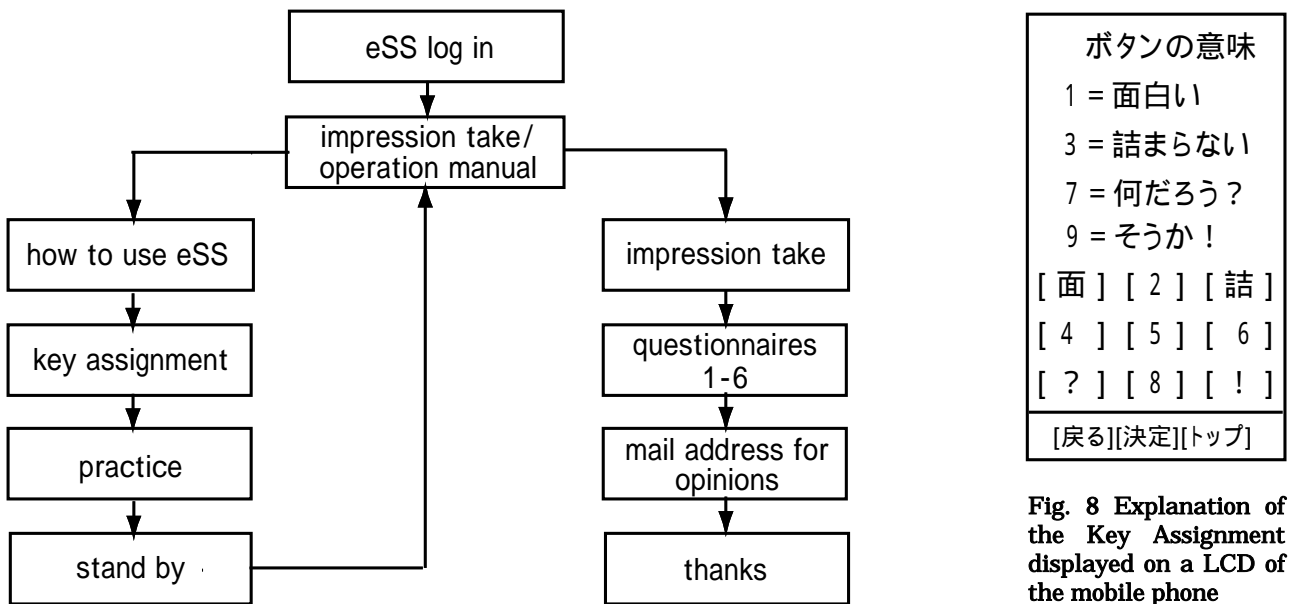


Fig. 7 Flow Diagram of the HTML Programs for Mobile Phones

3.2. Real-time Impression Evaluation Test for Broadcast TV Programs

For our real-time impression evaluation test for broadcast programs, we employed the hit variety program “Sanma no Super Karakuri TV,” broadcasted on TBS TV at seven o’clock Sunday evening, March 16th in 2003. We collected data for the first 30 minutes of the one-hour program. The volunteer test subjects from whom we collected data consisted of 24 people in all live in the Tokyo metropolitan area. Before the test, we sent test subjects an e-mail message that explained the gist of the test, noted the URLs where they could access the HTML programs, and reminded them that they could press keys any number of times, whenever they had an impression (“amusing,” “boring,” “huh?,” and “aha!”), while watching the program.

Figure 9-1 and 9-2 show simple tabulations of questionnaire results. Out of 24 subjects, 19 responded the questionnaires. The most watched program genre among them was by far news/reports, followed by documentaries, dramas, variety, sports, theatrical films, and music in that order. The Web was mentioned by the most subjects as the media they felt was important, followed by books/magazines, and TV in that order. From this, we can envision the character of the test subjects: regarding TV programs, they prefer informative programs, and regarding media, they prefer types of media that enable pull-type, on-demand information acquisition.

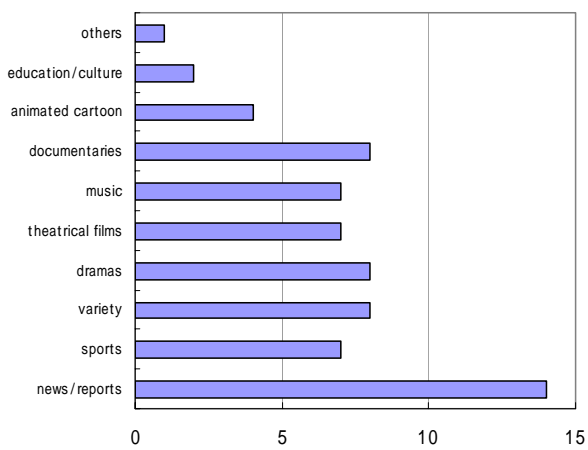


Fig. 9-1 Most Watched Program Genre

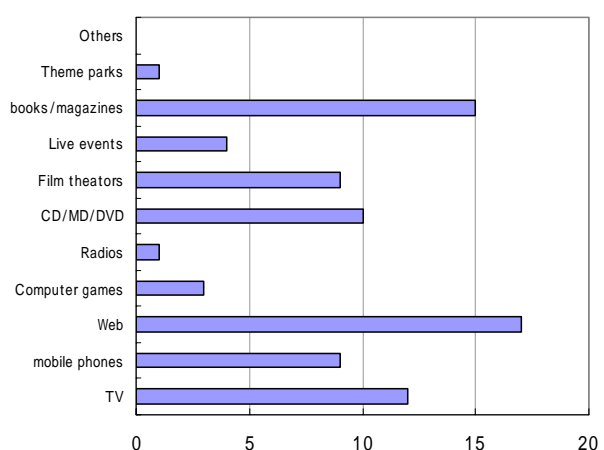


Fig. 9-2 Media They Felt Important

4. Analysis of Data from Real-time Impression Evaluation Test

4.1 Analysis of Graph Showing Chronological Frequency of Impressions

The number of entries by impression for the 24 test subjects is as shown in Figure 10. A breakdown by impression item reveals that “amusing” was the most prominent impression (entered an average of 14.2 times), followed by “boring” (5.2), “huh?” (4.6), and “aha!” (3.7). The number of entries overall averaged 27.7 and ranged widely between 71 at the top and 3 at the bottom.

Figure 11 shows the chronological frequency for the impressions “amusing/boring” and “huh?/aha!” respectively while watching the 30-minute program. To contrast paired impression terms, that graphs assign a positive value to “amusing” and “huh?” and a negative value to “boring” and “aha!.” And to make the synchronization of impressions between test subjects easier to see, the graph adds 2 seconds more or less to the time of entry in order to create 5-second intervals and then integrates the data in seconds.

The peaks wherein three or more test subjects had the same impression numbered 41 for “amusing,” 13 for “boring,” 11 for “huh?,” and 7 for “aha!.” A look at the characteristics of the video clips extracted from these peaks reveals that test subjects selected “amusing,” for the unexpected and funny scenes in “Masterpiece Video Corner” and the amusing remarks during little Taiki’s (5-year-old) karate practice scene in the “Ganbare Taiki-kun” and pottery class scene in the “Thane’s Funniest Language School.” “Boring” was selected for the silly events in “Masterpiece Video Corner” and the scenes showing failures and problems in the “Pottery Class Corner.” “Huh?” was selected for weird quiz answers and strange remarks. “Aha!” was selected for the unexpected conclusions to scenes in “Masterpiece Video” and the punch line for strange remarks. There were cases where “huh?” was followed several seconds later by “aha!” for some video clips.

The above-mentioned examination clearly shows that the peaks of each impression graph accurately reflect the characteristic video highlights that form the impressions held by test subjects.

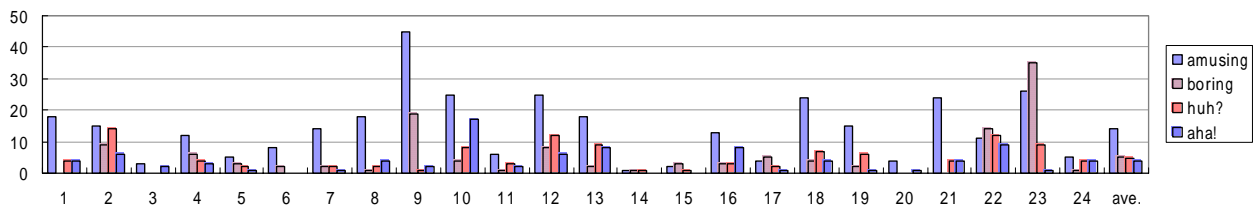


Fig. 10 Number of Entries by Each Impression “Amusing,” “Boring,” “Huh?” and “Aha!” for the 24 Test Subjects

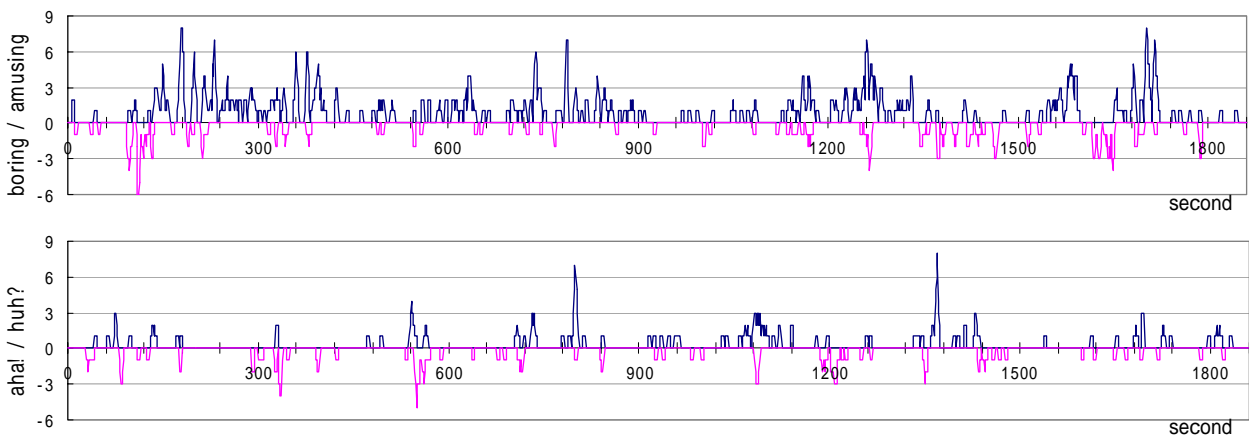


Fig.11 The Chronological Frequency for the Impressions “Amusing/boring” and “Huh?/aha!” (a data width = 5 seconds)

4.2 Analysis of Test Subjects' Impression Characteristics

Using the chronological impression data collected from test subjects, we extracted clusters of test subjects who had similar impressions to the program and then explored their significance.

We started by dividing the 30-minute program into 21 video plots based on its content (Figure 12) and derived sum frequencies for four impressions per test subject on each video plot. We then applied principal component analysis to this data and interpreted the significance of each principal component based on the content of video plots with a principal component score of 0.600 (absolute value). (Table 1) Based on our results, the first principal component was “Masterpiece Video” and the quiz chat, the second the kids’ “Masterpiece Video” and “Pottery Class,” the third “Ganbare Taiki-kun,” the fourth the opening “Masterpiece Video,” and the fifth the quiz about little Taiki.

We focused on the principal component scores up to the third principal component as determined from the eigenvalue graph (Figure 13) and then applied cluster analysis. Figure 14 shows the results overlaid onto a scatter chart of the principal component scores. The characteristics of the test subject clusters can be interpreted as follows based on the cluster positions on the scatter chart:

C1: Mainly reacted to the kids’ “Masterpiece Video” and “Pottery Class.”

C2: Did not react to “Ganbare Taiki-kun.”

C3: Mainly reacted to “Masterpiece Video” and quiz chat.

C4: Specifically reacted to “Masterpiece Video” and quiz chat.

C5: Mainly reacted to “Ganbare Taiki-kun.”

C6: Specifically reacted to “Ganbare Taiki-kun.”

C7: Specifically reacted to the kids’ “Masterpiece Video” and “Pottery Class.”



Fig. 12 Thumbnails of the 21 Plots

Table 1 Interpretation of the Principal Component Analysis

meaning of PC	Plot	Principal Component				
		PC1	PC2	PC3	PC4	PC5
"Masterpiece Video" and the quiz chat	P-20	0.901	0.123	0.069	-0.072	-0.056
	P-09	0.834	0.100	-0.158	0.033	-0.151
	P-11	0.809	0.267	-0.124	-0.178	0.082
	P-21	0.799	0.225	0.314	-0.045	0.205
	P-06	0.789	0.315	-0.087	-0.044	-0.156
	P-18	0.733	-0.040	0.458	0.116	0.241
the kids' "Masterpiece Video" and "Pottery Class,"	P-08	0.632	0.069	0.554	-0.124	-0.215
	P-05	0.223	0.916	0.062	0.080	-0.048
	P-19	0.138	0.903	0.203	-0.119	0.167
	P-17	0.239	0.753	0.267	-0.192	0.224
"Ganbare Taiki-kun"	P-04	0.107	0.750	-0.122	-0.441	-0.270
	P-13	-0.117	-0.168	0.844	0.250	-0.242
	P-14	-0.118	0.309	0.804	-0.084	-0.048
the opening "Masterpiece Video"	P-15	0.297	0.364	0.751	-0.099	-0.056
	P-01	0.169	0.015	0.138	-0.828	0.110
	P-02	0.018	0.101	-0.016	-0.829	-0.164
the quiz about little Taiki	P-03	-0.166	0.089	-0.172	-0.843	0.056
	P-12	0.091	-0.032	0.278	-0.029	-0.843
	P-16	0.380	0.533	0.172	-0.497	-0.066
	P-10	0.467	0.425	-0.193	0.223	-0.078
	P-07	0.503	0.495	0.258	0.032	-0.172
		5.291	4.020	2.980	2.785	1.214

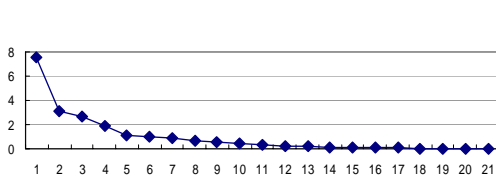


Fig. 13 Eigenvalue Graph of the Principal Component Analysis

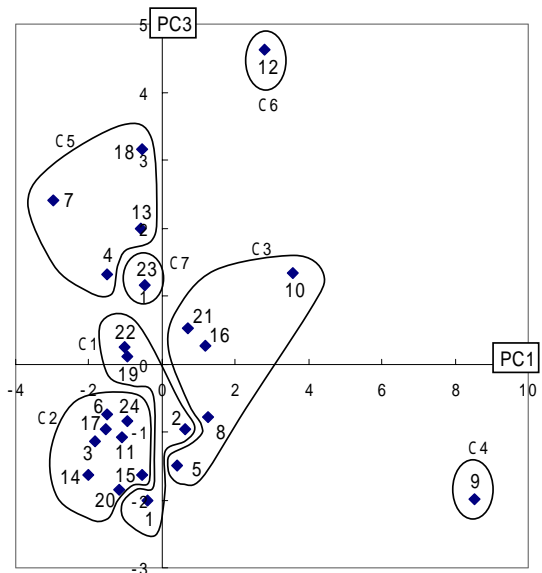
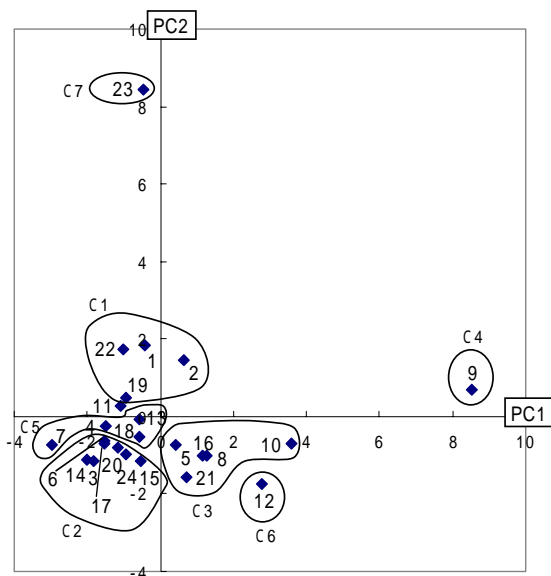


Fig. 14 Test Subjects' Clusters Overlaid onto a Scatter Chart of the Principal Component Scores

5. Examination of eSS Effectiveness and Practical Applications

5.1. Versatility and Effectiveness of eSS Using Mobile Phones

This latest test verified that the eSS using mobile phones is extremely effective as a tool for entering intuitive impressions about TV programs on real-time. Mobile phones can be easily operated by anyone, anytime, anywhere. Once impressions are entered, servers can collect accurate, time-coded (in seconds) data from across a wide area in Japan. The broadcast we used this time enabled the entry of data while test subjects were at home. Our system is effective for impression entry at event sites in addition to broadcasts. And for streaming content on the Internet and packaged content, it allows the comparison of the content and impression data at a later time, as long as the starting time is precisely controlled. In addition to content, the eSS will likely have applications as an input terminal for subjective evaluations of interpersonal communication and interaction with devices/systems. In other words, it is a versatile system for entering time-coded data for subjective evaluations involving an individual's experiences. It was revealed that we were able to extract program highlights from the impression data of audience. These impression data will be widely used as meta-data not only to making video highlights but also to information retrievals, automatic content recommendation, and so on.

5.2 About the Establishment of Impression Terms for Intuitive Entry

The SS method thus far has employed an evaluation system based on “complication/resolution” at several-second pauses per scene division. The impression entry using the latest eSS calls for the determination and entry of impressions in real time at any time while watching content. The selection of impression terms is crucial for ensuring that test subjects are not burdened. As was made clear by the latest analysis of the 10 pairs of impression terms, “amusing/boring” encompasses the meaning of a great many other impression terms and enables an intuitive determination in a wide range of situations. Test subjects seemed to have no qualms about it, particularly for the variety program in this latest test. This also can be determined based on the overwhelming number of times “amusing” was entered compared to the other impression terms. “Huh?/aha!” were easier-to-understand replacements for the concepts of “complication/resolution” that are used in ascertaining story structure, but there is a crucial difference between the two. We defined “complication” as “uncertainty as to how the story will develop from the present point” and “resolution” as “scenes that make the viewer say, ‘So that’s it!’ in view of story development so far.” In short, they seek to evaluate a scene while keeping in mind the entire story that has played out so far. On the other hand, “huh?/aha!,” which we used this time, are for the test subject’s impressions at the present moment in time. Of course, test subjects are still keeping the development of content thus far in mind, but ultimately the emphasis on the current impression differentiates these terms from “complication/resolution.” The same can be said for “amusing/boring.” The localized, intuitive determination of “huh?/aha!” reveals the structure of viewer’s interim questions and discoveries about the content or arousal and release of interest in a video production. A variety program of the sort used this time in particular was itself highly segmented and was comprised of mysteries and punch lines designed to capture the viewer’s attention at every turn. Its structure was not necessarily related to the story development of the overall program. Consequently, “huh?/aha!” functions effectively even for variety programs.

5.3 Characteristics of Content and Test Subjects

Impression data for content can reveal the characteristics of both content and test subject.

For the content in our latest test, we studied the chronological impression graph of the program and then

extracted highlight videos for four impressions. These videos can probably aid in saving program summaries and searching for archived programs. DVD recorders with built in hard disks and TVs with that same functionality have been showing signs of increased popularity in recent years, and if it were possible to obtain impression data from large numbers of viewers over the Internet, a person could roughly determine how good or bad a program was by previewing its highlights before watching it. In addition, it would even be possible to employ impression data for the browsing of content. A person's own impression data could be put to work in future searches as a bookmark function based on videos already seen. The algorithms to use for automatically extracting video for such purposes are an issue for future study.

We successfully uncovered the characteristics of test subjects through principal component analysis and cluster analysis using impression data obtained for the program plot segments. By creating highlight videos using impression data by cluster, we can expect video that is even better matched to cluster members. We would like to continue with our ongoing impression evaluations based on this point.

6 Acknowledgements

We would like to thank President Naotaka Ando and Keiji Kinoshita, Zeta Bridge Corporation, for the collaborative works of HTML programming development and Web server operation.

References

1. H. Kaiho, E. Harada; Introduction to Protocol Analysis (Japanese), Shin-yoh-sha (1993).
2. Y. Takahashi, Y. Shibata, M. Kamada, H. Kimura; The Semantic Score Method for Quantitative Evaluation of Video Content by Viewers, International Conference on Consumer Electronics/IEEE, 358-359(2001)
3. Y. Takahashi, K. Sugiyama, M. Watanabe; The Semantic Score Method for Describing Story Structures of Movies, 5th ADC(2001)
4. Y. Takahashi; A Video Summarization Algorithm Using the Semantic Score Method, and Its Applications, NICOGRAPH International 2002, The Society for Art and Science, 121-126(2002)