Interaction Model for Emotive Video Production

Hiroko Mitarai and Atsuo Yoshitaka

***Abstract*—Video images are capable of expressing various kinds of information; however special knowledge and techniques are required for authoring quality video content. In order to represent impressions, proper camerawork is required for delivering understandable content. However, amateur users often have difficulties in shooting images which appropriately reflect their expressive intentions. In this paper, we propose an incremental interaction model which supports the user's shooting activity and aims to take shots that fit the expressive intention. The model helps to build interaction incrementally between the user and the system. Experiments were carried out to see how amateur users express emotive information using video cameras without any prior information. Results showed that the users with more experience could recognize if they have shot footage in an appropriate way; however they could not always shoot appropriately. It was also indicated that their self-evaluation of the shots does not always reflect the actual suitability of the shots.**

***Index Terms*—Emotive, video production, film grammar**

I. Introduction

Image technologies have been evolving rapidly. The broadcasting media have branched out too; besides TV stations and movie theaters, the internet has opened up various opportunities, such as cell phones and internet broadcasting. Technical advances in video cameras have enabled instant image correction such as autofocus and auto white balance. In video editing, increasing hard disk space has yielded more editing space; people can handle non-linear editing more easily. Compared to text or still images, video can convey a vast amount of audiovisual information which includes temporal transitions. Along with dissemination of audiovisual equipment, home video production has penetrated to ordinary families. However, footage shot by amateur users on occasions such as weddings or a baby's first steps is often left without editing and seldom watched [1]. Adams states that inexperienced users tend to move cameras in random directions [2]. This is one of the causes for a gap between user's shooting intention and authored contents.

One of the possible reasons for this is that the users shoot subjects too carelessly, and do not consider enough if the footage shot is worth watching, or reflects their intention properly. Even if the footage was about someone revealing strong emotions, these emotions can not be expressed effectively with dull wide-angle shots. In order to express emotive information effectively, appropriate camerawork is necessary. This is not easy for inexperienced users, and this difficulty prevents smooth communications between the user and the audience.

In this paper, utilizing professional knowledge such as film production techniques [3] and film grammar [4] that have been adopted by industry professionals, we utilize emotive information such as impressions to convert to camerawork. Most present video shooting support technologies are limited to optical or digital auto correction such as exposure and focus, or displaying guides such as rule of thirds, therefore they do not explicitly assist users with shooting footage. An experiment was carried out to see how amateur users express emotive information using video cameras without any prior information. Results suggested that random shooting experience did not always help to acquire the shooting skill to emphasize intended impressions. In order to shoot appropriately, the system is required to suggest the appropriate shooting method, and support the user according to the shooting situation of the user. Utilizing a knowledge base of professional production techniques, the proposed model instructs the user the appropriate shooting method, matching current conditions. The user enters desired emotive information when shooting. The system then analyzes images during shooting and supports the user by suggesting camerawork or shot sizes visually. In this paper, we carried out experiments to analyze how ordinary users handle these techniques without any prior knowledge. Upon evaluation, an interaction model for video shooting support system is proposed.

II. Technologies Related to Video Shooting and Editing

A. Past Research and the Advantage

At the time of shooting, three factors must be considered: the size of subject on screen (shot size), the angle of subject being shot (camera angle), and the camera motion. For example, close-up shots are effective for emphasizing emotions, and strength can be expressed through low-angle shots, in which the camera is angled upward [3]. Tables I and II are lists of popular shot sizes and camera angles based on emotional expression.

TABLE I: Shot Sizes

Shot type	Details	Impression
Long shot	Includes full human body or more	Smaller
Medium shot	Includes waist up	Neutral
Close-up shot	Includes top shirt button up	Details, emotional

In long shots, subjects appear smaller. Medium shots include the subject from waist up. Since this distance evokes

Manuscript received May 21, 2012; revised June 12, 2012. This work is partially supported by Grant-in-Aid for Scientific Research (C), 21500197.

The authors are with School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292 Japan (e-mail: hmitarai@jaist.ac.jp, ayoshi@jaist.ac.jp).

DOI: 10.7763/IJIEE.2012.V2.183

the distance when two people are talking, it gives a neutral impression. Close-up shots give emotional impressions, since they give an impression of the subject standing close by.

Low-angle shots represent strength or threat of the subject in frame, because they give an impression of looking up at the subject. On the other hand, high-angle shots represent weakness or threatened atmosphere, because of the impression of looking down at the subject. Oblique shots represent madness or distorted world around the subject. Point-of-view shots represent viewpoint of a specific character, and are frequently adopted in horror films to express madness of the villain.

TABLE II: CAMERA ANGLES

Shot type	Camera position	Direction	Impression
Low-angle shot	Below subject	Up	Strength, intimidation
High-angle shot	Above subject	Down	Weakness, threatened
Eye-level shot	Subject's height	—	Neutral
Oblique shot	—	Diagonal	Distortion, unbalance
Point-of-view shot	Subject's perspective	—	Sympathy

Camera movements can also represent certain impressions, e.g. fast zoom-in for tension rising. Table III is a summary of camera motion based on reference [4].

TABLE III: CAMERA MOTIONS

Speed	Camerawork	Impression
Fast	Zoom-in, dolly-in	Tension, excitement
	Zoom-out, dolly-out	Liberation
Slow	Zoom-in, dolly-in	Closeness, Intimacy
	Zoom-out, dolly-out	Loneliness

A dolly is a carrier with wheels to move a camera smoothly. Technical terms such as *dolly-in* or *dolly-out* mean that the camera moves toward or away from the subject while shooting them. Practically, it is different from the shots applying zoom, but many directors use them interchangeably [3]. By using camera motion, impressions such as tension or liberation can be expressed.

B. Related Work

It is said that there are two levels of content understanding in the field of video content retrieval: cognitive level and affective level [5]. Similarly, two types of assistance exist in content creation: *technical assistance* and *affective assistance*. Technical assistance includes mechanical or electric image stabilization, correction of focus or exposure. Affective assistance includes support for users who want to express certain emotions or atmosphere. A number of studies are found in the field of technical assistance, especially in editing, however less are found regarding affective assistance. Research related to home video shooting can be classified into directing and cinematography. In direction, shooting instruction is given, and in cinematography, technical aspects of shooting are explained.

In direction assistance, active capture is a system which directs performances [6]. Notion of capture intention was proposed when shooting home videos, emphasizing the importance of the intention of the user [7]. MediaTE is a system which gives shot suggestions according to annotations on shot location, cast and subject [8]. It attempted to integrate shooting and editing processes. In order to create a more comprehensible story, a system utilizing commonsense was proposed [9].

In cinematographic assistance, besides the alleviation of shaky shooting with image stabilizer and lighting artifacts [10], a navigation system which detects inappropriate camera movements by analyzing shots was implemented [11]. Our research is classified into cinematography assistance; however, it supports the shooting process based on affective assistance. In editing, many studies have been done on technical assistance. AVE is a system which automated home video editing by synchronizing images and analyzing music tempo [12].

Hitchcock is an editing support system which searches for inappropriate sections in a clip by detecting improper camera motion such as inappropriately fast pan [13]. Zodiac is a video editing system which improves video editing manipulation by controlling edit history [14]. In terms of affective assistance, Yoshitaka proposed an automated editing system based on film grammar [15]. This system automatically edits arbitrary sequences by extracting temporal differences in the clip based on the selected emotional expression. Table IV is a summary of the studies mentioned above.

TABLE IV: SUMMARY OF STUDIES

	Technical Assistance	Affective Assistance
Production	**Directing:** Davis(2003), Barry(2003), Adams(2005)	Adams (2005)
	Cinematography: Yan(2002), Kumano(2007)	Our study
Postproduction	Chiueh(2000), Girgensohn(2000), Hua (2003), Adams(2005)	Yoshitaka (2009)

Effective editing helps the media transfer information more effectively. However, this only hides away the mistakes made in shooting. We carried out experiments to see if the ordinary users are capable of expressing emotional information without any prior knowledge, and evaluated the resultant shots.

Fig. 1. Capture environment.

In order to investigate how amateur users shoot their footage, we carried out an experiment requesting each examinee to emphasize the atmosphere corresponding to affective words (Fig. 1).

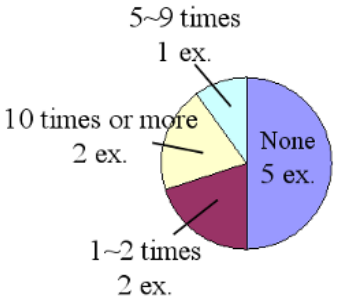

Fig. 2. Shooting experience.

The examinees were ten graduate students in their twenties, five with no prior video shooting experience, and two with one to two times, one with more than five times, two with more than ten times, as in Fig. 2. They were mainly home videos, and others were research demonstration movies or commercial video for an open-air café at a school festival. During the experiment, the examinees were directed not to use any other functions except record, zoom-in, and zoom-out.

Seven affective words, *emotion*, *strength*, *weakness*, *tension/excitement*, *closeness/intimacy*, *loneliness* and *liberation*, were shown to each examinee, and the examinees were requested to express each affective word by shooting the subject sitting on a chair with a consumer video camera using camera positions or camerawork of their choice. Table V indicates the assignments presented to each examinee, and camera placement or movement was set according to the film production techniques stated in the previous section.

TABLE V: EXPERIMENTAL CONDITIONS AND PROPER SHOOTING METHODS

Assignment	Restriction	Shooting method
Emotion	Camera position	Close up
Strength	Camera position	Low-angle shot
Weakness	Camera position	High-angle shot
Tension/excitement	Camera motion	Fast zoom/dolly in
Closeness/intimacy	Camera motion	Slow zoom/dolly in
Loneliness	Camera motion	Slow zoom/dolly out
Liberation	Camera motion	Fast zoom/dolly out

The examinees were requested to shoot the subject expressing emotion, strength or weakness using camera positions, then tension/excitement, closeness/intimacy, loneliness and liberation using camera motion (including zoom-in, and zoom-out). We investigated if the examinees could shoot appropriately conforming to each affective word without any prior knowledge. After the experiment, each examinee was inquired about their video experience, their planned shooting method to meet the requirements, and their self-evaluation of each shot.

III. EXPERIMENTAL RESULTS AND ANALYSIS

We scored the shots taken by the examinees and calculated ratio of appropriately answered questions. If the shots were taken corresponding to the methods specified in Table V, we defined them as appropriate. For the camera motion assignments, we gave them partial credit for suboptimal results.

A. Relation between Experience and Cinematographic Skill

Examinee's intention at shooting (conceived plan to shoot the footage) and the actual shots were not always the same. After shooting, we inquired the examinees about the shooting plan in order to express the specified impressions only with camera position or camera motion. If the answers were consistent with the shooting methods summarized in Table V, we evaluated them as appropriate.

Fig. 3 indicates the ratio of appropriately shot footage and appropriately answered questions from the questionnaire for evaluating cinematographic skill and knowledge. In the figure, *shot content* indicates the score of the appropriately shot footage; *answered content* indicates the score of appropriately answered questions in the questionnaire. Shot content score was higher than the corresponding answered content score except for examinees E and H. Others shot more appropriately without understanding how to shoot appropriately. Examinee F, I and J were evaluated comparatively high in shot content even though the examinee F was not an examinee with expertise. On the contrary, an experienced user, examinee H, obtained a fairly low score. It indicates that having more shooting experience does not always lead to more appropriate shooting skill.

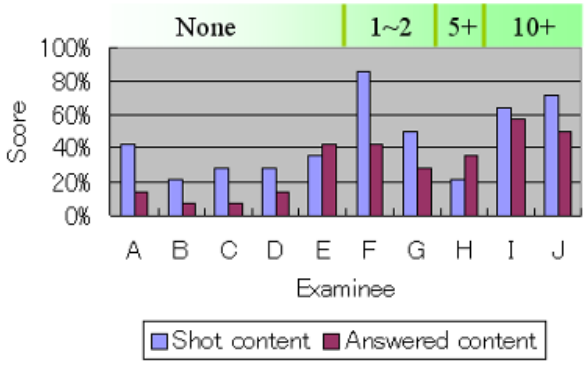

Fig. 3. Score on cinematographic skill and knowledge.

We then analyzed characteristics of examinees with more shooting experience.

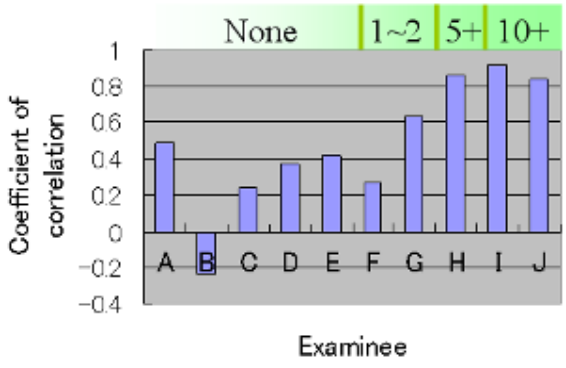

Fig. 4. Correlation of knowledge and actual skill.

Fig. 4 indicates coefficient of correlation for properly and improperly shot content, and answered content from the questionnaire. Since the experienced examinees G, H, I, and J showed an especially high correlation, it is assumed that examinees with more shooting experience are capable of recognizing the relationship between filmic expression and affective information more appropriately, even if they cannot always shoot appropriately.

B. Relation between Knowledge and Cinematographic Skill

We analyzed if the examinees understood the appropriate way of shooting corresponding to the purpose when shooting properly or improperly in each assignment. There are 70 combinations (10 examinees * 7 words) classified by shot content and answered content. Table VI indicates the details of the classification. We named them according to their characteristics. *Skilled* indicates that the category where shots were taken appropriately and questions were answered appropriately. *Unskilled* indicates that shots and answers were both inappropriate; *instinctive* indicates that shots were taken properly in the outcome but questions were wrongly answered, *unpracticed* indicates that the improper shots were taken, however indicates correct knowledge on cinematography at the same time.

TABLE VI: UNDERSTANDING OF SHOOTING METHOD AND CATEGORY

	Category	Shot Content	Answer content	Meaning
1	Skilled	A	A	Understands shooting method, and can shoot properly
2	Unskilled	N/A	N/A	Does not understand shooting method, and therefore cannot shoot properly
3	Instinctive	A	N/A	Does not understand shooting method, but can shoot properly
4	Unpracticed	N/A	A	Understands shooting method, but cannot shoot properly

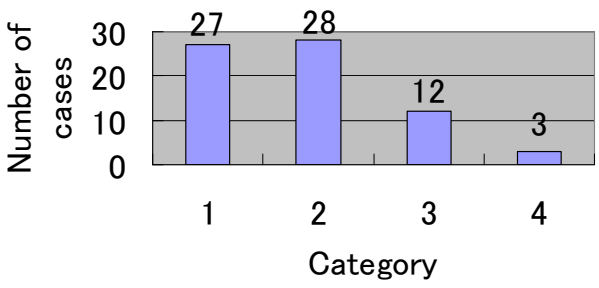

Fig. 5. Number of cases based on the category in Table VI.

Fig. 5 indicates that "unskilled" and "instinctive" comprise 57% of all combinations. This indicates that more than 50% of shots were taken without understanding the proper shooting method.

C. Relationship between Shot Suitability and Confidence

The ratio of appropriately answered questions and examinee's self-evaluation were compared. Fig. 6 shows the overall score from the experiment, which is the average of shot content and answered content, and examinee's average self-evaluation. Extraction of their self-evaluation was carried out in form of written questionnaires after the experiment.

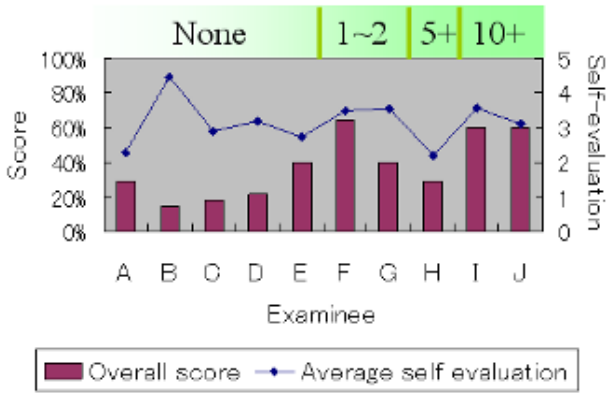

Fig. 6. Comparison between subjective and objective evaluation.

They evaluated themselves from "fair (scored 5)" to "poor (scored 1)" on each shot. The line chart in Fig. 6 shows their average self-evaluation. It indicates that people of less shooting experience are not capable of evaluating themselves appropriately.

There are two findings through the experiment. One is that if the users have some shooting experience, they are capable of recognizing if the shots are appropriately taken or not. The other is that even if the users think the shot was shot well, it is not always shot appropriately.

IV. A MODEL FOR INTERACTIVE VIDEO PRODUCTION

Results from the experiment suggest that improper shooting experiences do not always help acquire shooting skills which enable the expression of intended emotive impressions. In order to enable appropriate shooting by a user with little knowledge or experience, the system has to support the user with shooting directions at the time of shooting.

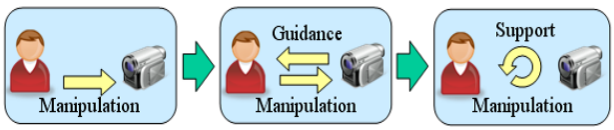

Fig. 7. Evolution of interaction models

Fig. 7 indicates the difference in interaction models between a user and a video camera. Initially, it was unidirectional interaction; the user manipulated the system through functions such as zoom. At the second stage, which is the present bidirectional interaction model, the system technically supports the user via functions like autofocus, auto white balance adjustment or auto exposure. In addition to this technical support, the system is equipped with functions such as a guideline for the rule of two-thirds. It gives the user a useful guideline; however it does not suggest

better shooting methods.

In the proposed model which we call *incremental interaction model*, the system supports the user's shooting activity and aims to match shooting method to the user's intention properly. Fig. 8 is the system design between the system and the user.

1) The user enters impression or atmosphere of his/her choice for the shooting.
2) The system determines which shooting method is appropriate based on the selected impression.
3) The system analyzes the obtained image.
4) The system instructs the user how to shoot the scene according to the analyzed result and the selected impression.
5) The user shoots by following the instructions.

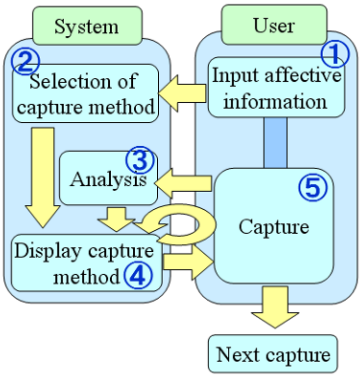

Fig. 8. System design.

V. Change of Production Style and Open Issues

A. Evolution of Video Production Style

Previously, the user set shooting parameters on the video camera even though the user does not have sufficient knowledge. In the proposed interaction model, the system suggests the shooting method according to the camera manipulation made by the user. This enables the user to gain more understanding of the relationship between the shooting method and the impression which the camerawork expresses, and it helps the user to acquire basic knowledge of the appropriate filmic expressions. The early interaction model only utilized user's knowledge and experience, and the video camera was positioned as a medium to express them, but the proposed model changes the interaction style to a collaborative creation between the user and the system.

B. Open Issues on System Design

When the system assists shooting, it is significant to consider how to structure sensible interaction between the user and the system. Below are issues to consider:

1) The way of expressing affective information and consideration of respective user's shooting skills

It is necessary to analyze the expressive shots the user tends to wish to shoot, types of input method to reflect user's intention, and feasibility of suggested method for the user; e.g. if the user tends to be shaky when shooting, the system should recommend shooting with a tripod. The suggested shots should be determined according to the level of skill of each user.

2) Display method when supporting shooting

It is necessary to determine the most effective and understandable way of instructional presentation when images shot are analyzed.

3) Respectful shooting support

Shooting is an important creative process. If unidirectional instructions are given and the shots are taken in accordance with them, the user would feel suppressed. In order to avoid this problem, the system must be seen as a "shooting support" and not a "shooting template".

4) Response time

When the user is shooting, the system has to analyze the image being shot and display the suggested shooting method at the same time. Appropriate response time is necessary in order not to give unnecessary stress on the user.

5) When suggested shooting method is unfeasible

Depending on the shooting location, suggested shot sizes and camera motion are not always physically feasible. If the system is able to give an optimal support, it will cover more shooting conditions.

VI. Conclusion

In this paper, we discussed the issues caused by user's lack of knowledge or skills concerning shooting. An experiment was carried out to investigate the way ordinary users shoot scenes according to the actual shooting method.

The result shows that if the users have much shooting experience, it is likely that they can recognize if the shots were taken appropriately. However, much experience does not always lead to appropriate shooting, and even if the users believe that the shots were taken properly, they are not always taken appropriately.

In order to shoot appropriately, the system is required to suggest the appropriate shooting method, and support the user according to his/her the shooting situation. With incremental shooting support, users can acquire appropriate shooting experience.

The user interacts with the system as follows: the user enters impression desired for shooting, and the system analyzes the shot being captured and suggests to the user the camerawork which fulfills the impression selected. By using the system based on this model, the interaction between the user and the system transforms from user's unidirectional manipulation of the system to an incremental interaction style between them. The user receives support at the appropriate timing, therefore can shoot more appropriately and effectively without specific knowledge, thereby acquiring basic shooting skills more smoothly and effectively during video production.

Acknowledgment

We would like to thank Professor Mary-Ann Mooradian for her English proofreading.

References

[1] D. Kirk, A. Sellen, R. Harper, and K. Wood, "Understanding Videowork," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 61-70, 2007.

[2] B. Adams, S. Venkatesh, and R. Jain, "IMCE: Integrated media creation environment," ACM Transactions on Multimedia Computing, *Communications, and Applications*, vol. 1, pp. 211-247, 2005.
[3] B. Mamer, Film Production Technique: creating the accomplished image, 2nd edition ed.: Wadsworth Pub Co, 2000.
[4] D. Arijon, Grammar of the Film Language: Silman-James Press, 1976.
[5] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.
[6] M. Davis, "Active capture: integrating human-computer interaction and computer vision/audition to automate media capture," in *Proceedings of the 2003 International Conference on Multimedia and Expo*, vol. 1, pp. 185-188, 2003.
[7] T. Mei, X. S. Hua, H. Q. Zhou, and S. Li, "Modeling and mining of users' capture intention for home videos," *IEEE Transactions on Multimedia*, vol. 9, pp. 66-77, 2007.
[8] B. Adams and S. Venkatesh, "Situated event bootstrapping and capture guidance for automated home movie authoring," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pp. 754-763, 2005.
[9] B. Barry and G. Davenport, "Documenting life: Videography and common sense," *International Conference on Multimedia and Expo*, vol. 2, pp. 197-200, 2003.
[10] W. Q. Yan and M. S. Kankanhalli, "Detection and removal of lighting & shaking artifacts in home videos," in *Proceedings of the Tenth ACM International Conference on Multimedia*, pp. 107-116, 2002.
[11] M. Kumano, K. Uehara, and Y. Ariki, "Online Training-Oriented Video Shooting Navigation System Based on Real-Time Camerawork Evaluation," *IEEE International Conference on Multi media and Expo*, pp. 1281-1284, 2006.
[12] X. Hua, L. Lu, and H. Zhang, "AVE: automated home video editing," in *Proceedings of the Eleventh ACM International Conference on Multimedia*, pp. 490-497, 2003.
[13] A. Girgensohn, J. Boreczky, P. Chiu, J. Doherty, J. Foote, G. Golovchinsky, S. Uchihashi, and L. Wilcox, "A semi-automatic approach to home video editing," in *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology*, pp. 81-89, 2000.
[14] T. Chiueh, T. Mitra, A. Neogi, and C. K. Yang, "Zodiac: a history-based interactive video authoring system," *Multimedia Systems*, vol. 8, pp. 201-211, 2000.
[15] A. Yoshitaka and Y. Deguchi, "Rendition-Based Video Editing for Public Contents Authoring," I*EEE International Conference on Image Processing*, pp. 1825-1828, 2009.

Hiroko Mitarai graduated from Keio University in 2009 and received M.S. degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in 2011. She is currently a PhD candidate at Japan Advanced Institute of Science and Technology. Her interests include human- computer interface, video processing and production. She is a student member of the Institute of Electronics, Information and Communication Engineers and has received Hokuriku Region Excellent Student Award from the Institute of Image Information and Television Engineers in 2011.

Atsuo Yoshitaka graduated from Hiroshima University in 1989, and received his degrees of Master and Doctor of Engineering in 1991 and 1997, respectively. He is currently an Associate Professor at School of Information Science, JAIST. His research interests include multimedia data analysis, image/video based human interfaces, and affective information processing. He is a member of IEEE Computer Society, the Institute of Image Information and Television Engineers, and the Information Processing Society of Japan.