Instant Movie Casting with Personality: Dive into the Movie System

Shigeo Morishima[1], Yasushi Yagi[2], and Satoshi Nakamura[3]

[1] Dept. of Advanced Science and Engineering, Waseda University,
3-4-1 Okubo Shinjuku-ku, Tokyo, Japan
[2] The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, Japan
[3] National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
shigeo@waseda.jp, yagi@am.sanken.osaka-u.ac.jp,
satoshi.nakamura@nict.go.jp

Abstract. "Dive into the Movie (DIM)" is a name of project to aim to realize a world innovative entertainment system which can provide an immersion experience into the story by giving a chance to audience to share an impression with his family or friends by watching a movie in which all audience can participate in the story as movie casts. To realize this system, we are trying to model and capture the personal characteristics instantly and precisely in face, body, gait, hair and voice. All of the modeling, character synthesis, rendering and compositing processes have to be performed on real-time without any manual operation. In this paper, a novel entertainment system, Future Cast System (FCS), is introduced as a prototype of DIM. The first experimental trial demonstration of FCS was performed at the World Exposition 2005 in which 1,630,000 people have experienced this event during 6 months. And finally up-to-date DIM system to realize more realistic sensation is introduced.

Keywords: Personality Modeling, Gait Motion, Entertainment, Face Capture.

1 Introduction

The purpose of DIM project is to aim to realize an innovative entertainment system in which especially all of the audience can become a movie cast instantly and can enjoy an immersive feeling and excited impression in the story. Future Cast System is one of a prototype of DIM which first was exhibited at the 2005 World Exposition in Aichi Japan. The Future Cast System offers two key features. First, it provides each member of the audience with the amazing experience of participating in a movie in real time. Second, the system can automatically perform all of the processes required to make this happen: from capturing the shape of each audience member's face, to generating a corresponding CG face in the movie. To create a CG character which closely resembles the visitor's face with FCS system, an original 3D range scanner first acquires the 3D geometry and texture of the visitor's face. Second, facial feature points, such as eyes, eyebrows, nose, mouth, and the outline of the face are extracted from the frontal face image. Next, a generic facial mesh is adjusted to smoothly fit

R. Shumaker (Ed.): Virtual and Mixed Reality, Part II, HCII 2011, LNCS 6774, pp. 187–196, 2011.

over the feature points of the visitor's face. Scanned depth of the visitor's face which corresponds to vertices on the deformed generic facial mesh is acquired and then the visitor's frontal face image is attached to the mesh. The entire process, from the face scanning stage to the CG character generation stage, is performed automatically as soon as possible before the show starts. Then, scanned data, including the face model, estimated age and gender information are transferred to a file server.

At the time of the show, we call up the visitor's character information, such as facial geometry, texture, and so on, from the file server and control the face using Cast/Environment Scenario Data which is defined for each individual cast member and scene of the movie. Then, the face, which is produced by facial expression synthesis and re-lighting, is rendered into the face region in the scene image. Therefore animation in which the CG character speaks and expresses his or her feelings as if it were the visitor itself can be generated.

2 Related Works

Face Geometry Acquisition. A variety of 3D range scanners [1] are already commercially available. However, it is difficult to capture the 3D geometry of areas which absorb lasers or sine-wave patterns such as the eyes, the inside of the mouth, and hair. In addition, these scanners are too expensive to use for our work. So we propose an original face scanning system which captures and reconstructs range data in a very short time maintaining high degree of accuracy.

Face Recognition. Wiskott et al developed elastic bunch graph matching using Gabor wavelets [2]. Kawade and Lee also developed techniques to improve this method [3],[4], and have been able to perform facial feature point extraction. This method can reliably extract facial features from frontal face images. Li et al have evaluated this technique using their facial database and, as a result, 77% of automatically extracted feature points were within 5 pixels of the ground truth. Baback and Lian have developed age and gender estimation techniques [5],[6] that are able to estimate age and gender with more than 80% accuracy. FCS system introduces 89 feature points extraction from frontal face image based on [2] with 94% accuracy of face modeling. In FCS, estimated age information is used for casting order and gender information is used to choose pre-scored voice track.

Facial Animation. Since the earliest work in facial modeling and animation [7], generating realistic faces has been the central goal. A technique to automatically generate realistic avatars has been developed using facial images [8]. The CG face model is generated by adjusting a generic face model to target facial feature points, which are then extracted from the image. There are many researches which can construct facial geometry directly from images using morphable facial models [9]-[11]. In FCS system, by fitting generic model onto one frontal face image and then get a model depth from range scanning data at the same time to simplify total face modeling process within 15 seconds now.

Current major facial synthesis techniques include: image-based, photo-realistic speech animation synthesis using morphed visemes [12], a muscle-based face model

has been developed [13]-[15], as has a blend shape technique where facial expressions are synthesized using linear combinations of several basic sets of pre-created faces for blending [11],[16]-[19]. Additionally, a facial expression cloning technique has also been developed [20].

The blend shape-based approach applicable to individuals is introduced in FCS to reduce the calculation cost of real-time process and also the labor cost of production process. Basic shape patterns are prepared in generic face model to represent the target facial expression in advance. After fitting process, personalized basic shapes are automatically created and prepared for blend shape process in the story.

Face Relighting. To achieve photorealistic face rendering, human skin's Bi-directional Reflectance Distribution Function (BRDF) and subsurface scattering of light by human skin need to be measured for a personal skin reflectance model. A uniform analytic BRDF is measured from photographs[10]. A Bi-directional Sub-surface Scattering Reflectance Distribution Function (BSSRDF) is developed with parameters based on biophysically-based spectral models of human skin[21]. The method which generates a photorealistic re-lightable 3D facial model is proposed by mapping image-based skin reflectance features onto the scanned 3D geometry of a face[22]. An image-based model, an analytic surface BRDF, and an image-space approximation for subsurface scattering are combined to create highly-realistic facial models for the movie industry[23]. A variant based on this method for real-time rendering on recent graphics hardware is proposed[24]. 3D facial geometry, skin reflectance characteristics and subsurface scattering are measured using custom-built devices[25].

To render faces in FCS, only a face texture (diffuse texture) captured by a range scanner can be used. Furthermore, in "Grand Odyssey" the movie title specially arranged for FCS, there are scenes where a maximum of five characters appear simultaneously. Thus, the rendering time for each character's face is restricted. An important role of our face renderer is to generate images that reduce the differences between visitors' faces and the pre-rendered story images. For skin rendering, first the texture of the scanned face together is used with uniformly specular lighting as a diffuse albedo. Then a hand-generated specular map is used on which speculars on the forehead, cheeks, and tip of the nose become more non-uniformly pronounced than on other parts of the face. Thereby efficiently an image similar in quality to that of the pre-rendered background CG image can be generated.

Immersive Entertainment. The last decade has witnessed a growing interest in employing immersive interfaces, as well as technologies such as augmented/virtual reality (AR/VR) to achieve a more immersive and interactive theater experience. For example, several significant academic research have focused on developing research prototypes for interactive drama [26]-[28]. One technical feature of the current immersive dramatic experience systems is characterized by a reliance on VR/AR devices and environments that create the immersive experience. Typically, players need to wear video-see-through head-mounted displays (HMD) to enter the virtual world.

DIM system is world first an immersive entertainment system commercially implemented as FCS system in Aichi Expo.2005. In DIM movie is generated by

replacing the faces of the original roles in the pre-created background movie with audience's own high-realism 3D CG faces. DIM movie is in some sense a hybrid entertainment form, somewhere between a game and storytelling, and its goal is to enable audience easily to participate in a movie as its roles and enjoy an immersive, first-person theater experience.

3 Summary of FCS

Future Cast System (FCS) has two key features: first, it can automatically create a CG character in a few minutes from capturing the facial information of a participant and generating her/his corresponding CG face, to inserting the CG face into the movie; second, FCS makes it possible for multiple people easily to take part in a movie at the same time in different roles, such as a family and a circle of friends. This study is not limited to academic research; 1,630,000 people enjoyed the DIM Movie experience at the Mitsui-Toshiba pavilion at the 2005 World Exposition in Aichi, Japan.

And now it is extended to new entertainment event at HUIS TEN BOSCH, Nagasaki, Japan from March 2007 and also becoming to the most popular attraction there. In this HUIS TEN BOSCH version, the face capturing process is performed only less than 15 seconds. So in case modeling failure is detected by attendant, it is possible to capture face again and again. Fig. 1 shows the processing flow of bringing audience into the pre-created movie.

Processing flow of FCS is as follows.

1. Under the attendant's guidance, participants put their faces to a 3D range-scanner.
2. The captured facial images are transmitted to Face Modeling PCs.
3. The frontal face images are transferred from Face Modeling PCs to Storage Server.
4. 2D thumbnail images are created and sent to Attendant PC to check to see if the captured images are suitable to undergo modeling. If OK, the decision on face modeling is sent back to Face Modeling PCs.
5. In Face Modeling PCs, 3D facial geometry is first generated which make the personal CG face with texture. Personal key shapes are generated for blend shape.
6. The personal face model, age and gender information used for assigning movie roles and the personal key shapes are transmitted to Storage Server.
7. Thumbnail images of participants' face model and information about age and gender is transmitted to Attendant PC from Storage Server. The attendant judges if the participant's face model is suitable to be appeared on the movie. If Yes, Attendant PC automatically assigns a movie role for each participant based on the gender and age estimation. When an attendant found the modeling error happened, it goes back to the process No.1 to capture the guest face again with careful instruction.
8. The decisions on assigning roles to participants are transmitted to Storage Server. A voice track is selected based on the gender information.
9. The information required for rendering movie, including casting information, participant's CG face model, frontal face texture, and customized key shapes for facial expression synthesis, are transmitted to the Rendering Cluster System from Storage Server.

10. Finally, Rendering Cluster System embeds participant's face into the pre-created background movie images using the received information and the completed movie images are projected onto a screen.

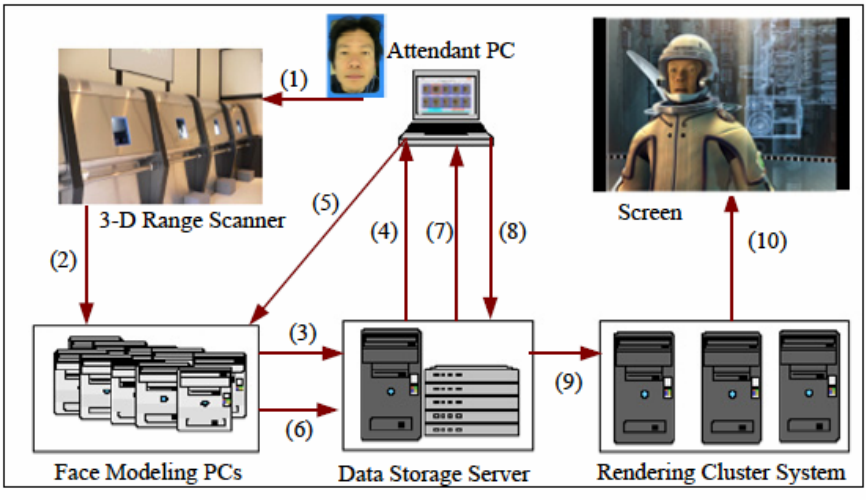
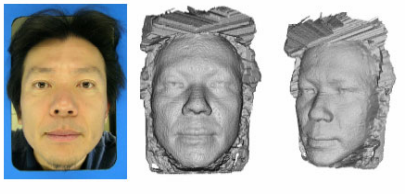
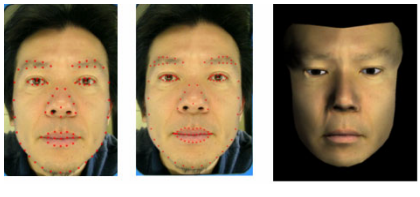


Fig. 1. Processing Flow of FCS



(a) (b) (c)

Fig. 2. Face Modeling Result. (a) frontal image. (b) (c) constructed geometry.



(a) (b) (c)

Fig. 3. Personal Face Model. (a) 89 feature points. (b) 128 feature points. (c) face model with personality.

Face Capturing System. A non-contact, robust and inexpensive 3D range-scanner is developed to generate participant facial geometry and textural information. This scanner consists of a few CCD digital cameras and two slide projectors which are able to project stripe patterns. They are placed in a half circle and can instantly capture face images in few seconds. The central camera captures the frontal face image to generate texture mapping (Fig. 2.a). The hybrid method with an Active-Stereo technique and a Shape-from-Silhouette technique [1] are used to reconstruct 3D geometry. Fig. 2 shows a frontal face image and its corresponding 3D face geometry constructed by a scanner system.

Face Modeling. Face modeling is completed in approximately 2 minutes in Aichi Expo. version, 15 seconds in HUIS TEN BOSCH version of FCS system and now

up-to-date scanner can complete it less than 10 seconds with the highest accuracy. Overall this process includes four aspects: facial feature point extraction, face fitting, face normalization and gender and age estimation. Currently, we proposed "Active Snap Shot" which can generate a texture mapped personal head model in about 2 seconds only with 1 single 2D snap shot and 2mm vertex accuracy.

Face Fitting. In the face fitting phase, a generic face model is deformed using the scattered data interpolation method (Radial Basis Functions transformation, [32]-[34]) which employs extracted feature points as targets. This technique enables us to deform the model smoothly. However, we found that deformed vertices on the model do not always correspond to the outlines of facial areas other than the target points. According to our own experience, it is important that the vertex of the model conform closely to the outline of the eyes and the lip-contact line on the facial image because these locations tend to move a lot when the model performs an action such as speaking or blinking. We therefore must accurately adjust the model for the feature points corresponding to the outline of the eyes, lips, and especially, the lip-contact line. Consequently, we interpolate these areas among the 89 feature points (Fig.3.a) using a cubic spline curve and then resample 128 points (Fig.3.b) on an approximated curve to achieve correspondence between the feature points. This process is performed for eyes, eyebrows, and the outline of the lip.

Real-time Facial Expression Synthesis. We selected the blend shape approach to synthesize facial expression, since it can operate in real-time and it's easy to prepare suitable basic expression patterns with which animators can produce and edit specific facial expressions in each movie scene. Also, the parameters for facial expression are very simple because they are only expressed with the blending ratio of each basic face. We first create universal 34 facial expression key-shapes based on generic face mesh, and then calculate the displacement vectors between the neutral mesh (generic face mesh) and the specific expressions' mesh (one of 34 key shapes) such as mouth opening. In the current DIM system, basic face is constructed by physics based facial muscle control adjusted to each audience's estimated facial muscle structure.

Relighting Face. The goal of relighting faces is to generate a matching image between the participant's face and the pre-created movie frames. Considering the requirement for real-time rendering, we implemented a pseudo-specular map for creating facial texture. Specifically, we chose to use only the reflectional characteristic of participants' skin which is acquired from the scanned frontal face images; however, the images still contain a combination of ambient, diffuse and specular refection. We therefore devised a lighting method to ensure a uniform specular reflection for the whole face by installing an alignment of florescent lights fitted with white diffusion filters. To represent skin gloss, we developed a hand-generated specular map on which speculars on the forehead, cheeks, and tip of the nose become more non-uniformly pronounced than other parts of the face. The intensity and sharpness of the speculars were generated by white Gaussian noise. The map can be applied to all participants' faces. Additionally, we approximated subsurface scattering and hair lighting effects by varying intensity on the outlines of the characters' faces.

Currently, personal specular feature is customized to each individual by the captured data from custom made skin sensor.

Automatic Casting and Voice Selection. FCS can automatically select a role from the background movie for each participant according to age and gender estimation. In addition, we pre-recorded a male and female voice track for each character, and voices were automatically assigned to characters according to the gender estimation results.

In current DIM, an individualized voice is generated by huge voice actor database.

Real-time Movie Rendering. The rendering cluster system consists of 8 PCs tasked for image generation of movie frames (Image Generator PC, IGPC) and a PC for sending the image to a projector (Projection PC, PPC). The PPC first receives the control commands, time codes, and participants' CG character information from Data Storage Server, and then sends them to each IGPC. The IGPCs render the movie frame images using a time division method and send them back to the PPC.

The performance of graphics hardware is advanced drastically, current DIM system can generate and render multiple characters with whole body motion and rendered hair as well as facial expression on real-time process.

Evaluation. FCS has been successfully exhibited from 3/25 to 9/25 at the 2005 World Expo. in Aichi, Japan, and showed a high success rate (94%) of bringing audience into the movie. However, evaluation result shows self-awareness rate is only 65%. Now up-to-date DIM movie can make self-awareness rate to 86% by introducing new technology described next.

4 Up-to-Date DIM Movie

In the former FCS system, 3D geometry and texture of audience face without glasses can be reproduced precisely and instantly in a character model. However, facial expression is controlled only by a blend shape animation, so personality of smile is not reflected in animation at all. Also the head of character is covered by a helmet in the story "Grand Odyssey" because of the huge calculation cost of hair animation and body motion, which are treated as a back ground image with fixed size and fixed costume assigned to the casting. So sometimes a few of audience cannot identify himself on the screen even if his family or friends can identify him. The progress in GPU performance is so high that real-time rendering and motion control of full body and hair become possible now.

Head Modeling. To emphasize personality feature on character facial animation, a head and hair modeling are inevitable. After scanning face, we tried to fit generic skin head model to captured face geometry and then a pre-designed wisp model of hair is generated automatically on this head model. To reduce the calculation cost, each wisp of hair is modeled by a flat transparent sheet with hair texture and an extended cloth simulation algorithm is applied to make hair animation.

Figure 4 shows the generated hair variation from an automatically constructed head model. The technical detail is described in [29]. To change hair style makes each character more attractive and impressive.

Fig. 4. Head Model with Hair Style Variation

Face Modeling. Blend shape facial animation is so simple that it is very convenient for creator to direct a specific expression and it takes very low CPU cost to synthesize on real-time process. However, to reflect personal characteristics on facial animation, physics based face model is effective because it is easy to give variation in motion by changing the location and stiffness of each facial muscle.

In the most recent DIM system, after fitting a face wireframe to scanned face range data, the location of each standard muscle can be decided automatically. By modifying the location of muscles, a variety of personal individuality can be generated in a same face model. This optimization of the location and physical character of each muscle is performed based on a captured face video appearing smile without any landmarks on face. The detail is expressed in [29].

Individual skin feature is captured by custom made skin sensor and reflected to the rendering of personalized characters [32].

Body and Gait Modeling. Character model with a variety of body size is pre-constructed and optimum size is decided according to the captured silhouette image. By fitting the size of neck, head, face and hair to the body size, a full body character model is assembled by morphing of these basic body shapes. Also personality in gait is modeled automatically by combination of key motions according to the silhouette estimation captured by 2 cameras. This technique is expressed in [30].

Voice Modeling. In former FCS system, a prerecorded voice of an actor or actress is used as a substitute for that of each participant. The substitute voice is selected by only each participant's gender information which is estimated based on the scanned face shape without consideration of other information such as age and voice quality. This caused some sense of discomfort for those who perceive the voice of the character to be different from their own or the people they know. Therefore we tried to choose or synthesize voice with personality based on the similarity measurement [31].

Immersive Experience with First Person Perspective. DIM project also produces both visual and audible immersive environment with first person perspectives. Prototype for high fidelity 3D audio recording and play-back [34] and omnidirectional panoramic video recording and screen system [33] have been constructed.

5 Conclusions

In this paper, an innovative visual entertainment system DIM is presented. The face modeling process, the core technology of our system, can automatically generate visitors' CG faces with an accuracy of 93.5%. So the former Future Cast System provides audiences with a movie in which only their faces are embedded.

Fig. 5. Personalized Character In "GrandOdyssey"

In current DIM system, how to improve the audience's sense of personal identification while watching the movie is considered. The key lies in generating CG characters which share the visitors' physical characteristics, such as hairstyle, body shape, gait motion, and voice personality. By introducing these factors, self-awareness rate is improved to 86% from 65% of former FCS system.

The detail of DIM project is as follows.

http://www.mlab.phys.waseda.ac.jp/DIM/home_e.html

References

1. Fujimura, K., et al.: Multi-camera 3d modeling system to digitize human head and body. In: Proceedings of SPIE, pp. 40–47 (2001)
2. Wiskott, L., et al.: Face recognition and gender determination. In: FG 1999, pp. 92–97 (1999)
3. Kawade, M.: Vision-based face understanding technologies and applications. In: Proc. of 2002 International Symposium on Micromechatronics and Human Science, pp. 27–32 (2002)
4. Zhang, L., et al.: A fast and robust automatic face alignment system. In: ICCV 2005 (2005)
5. Moghaddam, B., Yang, M.: Gender classification with support vector machines. In: Proceedings of the fourth IEEE FG 2000, pp. 306–311 (2000)
6. Lian, H., et al.: Gender recognition using a min–max modular support vector machine. In: International Conference on Natural Computation, pp. 438–441 (2005)
7. Parke, F.I.: Computer generated animation of faces. In: ACM 1972: Proceedings of the ACM annual Conference (1972)
8. Lyons, M., et al.: Avatar creation using automatic face processing. In: Proceedings of the sixth ACM international conference on Multimedia, pp. 427–434 (1998)
9. Pighin, F., et al.: Synthesizing realistic facial expressions from photographs. In: SIGGRAPH 1998 (1998)
10. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of SIGGRAPH 1999, pp. 187–194 (1999)
11. Blanz, V., et al.: Reanimating faces in images and video. In: Computer Graphics Forum 2003 (2003)

12. Ezzat, T., et al.: Trainable videorealistic speech animation. In: SIGGRAPH 2002, pp. 388–398 (2002)
13. Lee, Y., Terzopulus, D., Waters, K.: Realistic modeling for facial animation. In: Proceedings of SIGGRAPH 1995, pp. 55–62 (1995)
14. Choe, B., Lee, H., Ko, H.: Performance-driven muscle-based facial animation. In: Proceedings of Computer Animation, pp. 67–79 (2001)
15. Sifakis, E., Neverov, I., Fedkiw, R.: Automatic determination of facial muscle activations from sparse motion capture marker data, pp. 417–425 (2005)
16. DeCarlo, D., Metaxas, D.: The integration of optical flow and deformable models with applications to human face shape and motion estimation. In: CVPR 1996, pp. 231–238 (1996)
17. Joshi, P., Tien, W.C., Desbrun, M., Pighin, F.: Learning controls for blend shape based realistic facial animation. In: SCA 2003, pp. 187–192 (2003)
18. Xiang, J., Chai, Xiao, J., Hodgins, J.: Vision-based control of 3d facial animation. In: SCA 2003, pp. 193–206 (2003)
19. Vlasic, D., Brand, M., Pfister, H., Popovi, J.: Face transfer with multilinear models. ACM Transactions on Graphics 24(3), 426–433 (2005)
20. Noh, J.Y., Neumann, U.: Expression cloning. In: SIGGRAPH 2001, pp. 277–288 (2001)
21. Krishnaswamy, A., Baranoski, G.V.G.: A biophysically-based spectral model of light interaction with human skin. Comput. Graph. Forum, 331–340 (2004)
22. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: SIGGRAPH 2000, pp. 145–156 (2000)
23. Borshukov, G., Lewis, J.P.: Realistic human face rendering for ”the matrix reloaded. In: ACM SIGGRAPH 2003 Technical Sketch (2003)
24. Sander, P.V., Gosselin, D., Mitchell, J.L.: Real-time skin rendering on graphics hardware. In: SIGGRAPH 2004, Sketches, p. 148 (2004)
25. Weyrich, T., et al.: Analysis of human faces using a measurement-based skin reflectance model. In: SIGGRAPH 2006, pp. 1013–1024 (2006)
26. Cavazza, M., Charles, F., Mead, S.J.: Interacting with virtual characters in interactive storytelling. In: Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems: part 1, pp. 318–325 (2002)
27. Cheok, A.D., Weihua, W., Yang, X., Prince, S., Wan, F.S., Billinghurst, I.M., Kato, H.: Interactive theatre experience in embodied + wearable mixed reality space. In: Proc. of ISMAR 2002, pp. 59–317 (2002)
28. Mateas, M., Stern, A.: Facade: An experiment in building a fully-realized interactive drama. In: Proceedings of Game Developers' Conference: Game Design Track (2003)
29. Maejima, A., et al.: Realistic Facial Animation by Automatic Individual Head Modeling and Facial Muscle Adjustment. In: HCI International 2011, ID:1669 (to appera, 2011)
30. Makihara, Y., et al.: The Online Gait Measurement for Characteristic Gait Animation Synthesis. In: Proc.of HCI International 2011, ID:1768 (to appear, 2011)
31. Shin-ichi, K., et al.: Personalized Voice Assignment Techniques for Synchronized Scenario Speech. In: HCI International 2011, ID:2218 (to appear, 2011)
32. Mashita, T., et al.: 'Measuring and Modeling of Multi-layered Subsurface Scattering for Human Skin. In: Proc.of HCI International 2011, ID:2219 (to appear, 2011)
33. Kondo, K., et al.: Providing Immersive Virtual Experience with First-person Perspective Omnidirectional Movies and Three Dimensional Sound Field. In: HCI International 2011, ID:1733 (to appear, 2011)
34. Enomoto, S., et al.: 3-D sound reproduction system for immersive environments based on the boundary surface control principle. In: Proc. HCI International 2011, ID: 1705 (to appear, 2011)