“Dive into the Movie” Audience-Driven Immersive Experience in the Story

Shigeo MORISHIMA^{†a)}, Member

SUMMARY “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, several techniques to model and capture the personal characteristics instantly in face, body, gesture, hair and voice by combining computer graphics, computer vision and speech signal processing technique. Anyway, all of the modeling, casting, character synthesis, rendering and compositing processes have to be performed on real-time without any operator. In this paper, first a novel entertainment system, *Future Cast System* (FCS), is introduced which can create *DIM movie* with audience’s participation by replacing the original roles’ face in a pre-created CG movie with audiences’ own highly realistic 3D CG faces. Then the effects of *DIM movie* on audience experience are evaluated subjectively. The result suggests that most of the participants are seeking for higher realism, impression and satisfaction by replacing not only face part but also body, hair and voice. The first experimental trial demonstration of FCS was performed at the *Mitsui-Toshiba pavilion* of the 2005 World Exposition in Aichi Japan. Then, 1,640,000 people have experienced this event during 6 months of exhibition and FCS became one of the most popular events at Expo.2005.

key words: innovative entertainment, instant casting, facial modeling and animation, immersive space, audience-driven movie

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, a 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 over the feature points of the visitor’s face.

Manuscript received February 20, 2008.

[†]The author is with Waseda University, Tokyo, 169-8555 Japan.

a) E-mail: shigeo@waseda.jp

DOI: 10.1093/ietisy/e91-d.6.1594

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 can be generated.

This paper is organized as follows: in Sect. 2, related works about key technologies of *DIM* are introduced. Section 3 shows summary of *Future Cast System*, how each visitor’s 3D face model is generated, and the process of the real-time direction of a visitor’s face. In Sect. 4, current and next stage of *DIM* project is reported. In Sect. 5, conclusion is presented.

2. Related Works

Face geometry acquisition

3D range scanners such as the Cyber-ware 3030/RGB, Minolta Vivid 910, NEC DANAE-R and Sanyo PIERIMO [1] are already commercially available. These scanners are able to capture face geometry with a high degree of accuracy. However, it is difficult for the Cyber-ware 3030/RGB, Vivid 910, and NEC DANAE-R 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. However, if a face is illuminated by rainbow stripes and the facial silhouette is complete, the PIERIMO can measure the facial geometry of these areas. In addition, this scanner is comparatively inexpensive. It was necessary to acquire the depth of pupils in *FCS* to estimate the center position of eyes in order to create eye model, a modified version of PIERIMO [1] is introduced.

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 with a high degree of accuracy. This method can reliably extract facial features such as ethnicity, gender, age, facial hair, and accessories 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, and only 2% of the points were more than 10 pixels in error. 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. Age information in FCS 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 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. A variety of basic shape patterns are prepared in generic face model to represent the target facial expression in advance. After generic model 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]. They developed a new skin reflectance model where parameters can be estimated from measurements. The model breaks measured skin data down into a spatially-varying BRDF, a diffuse albedo map, and a diffuse subsurface scattering.

However, it is difficult for us to measure BRDF, skin subsurface scattering, and a diffuse albedo using the Light Stage or the Face Scanning Dome [22], [25], because of the prohibitive costs of these large scale measuring devices. Furthermore, if reflectance characteristics of individual faces are measured with these devices, huge data storage areas would be required. It is also difficult to measure and reproduce light from the subsurface scattering of human skin. The custom-built device proposed by [25] is able to capture skin subsurface scattering, but because the device is a contact type, this device cannot be used due to hygiene factors.

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. Additionally, the reflectance characteristics of the face, for example subsurface scattering and the appearance of long hair, are intentionally represented by varying light intensity on the outline of the face. Thereby efficiently an image similar in quality to that of the pre-rendered background CG image can be generated.

In the next stage DIM system, to realize higher reality, it is inevitable to implement both BRDF and BSSRDF model with an instant capturing system of personal features.

Fig. 1 Creating DIM movies using FCS: The original role's face in a CG background movie is replaced with the highly realistic 3D CG faces of audience.

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 researches have focused on developing research prototypes for interactive drama [26]–[28]. In *Bond Experience* [25], the player's image is inserted into a virtual world where the player interacts with James Bond in vignettes taken from a longer story. *Façade* [28] attempts to create a real-time 3D animated experience akin to being on stage with two live actors, but in a virtual world, who are motivated to make a dramatic situation happen. These examples of interactive drama have shown great potential in facilitating interactive and immersive dramatic experiences, but they do not yet exhibit convincing artistic or entertainment values [29]; in particular, the gap between academic research and practical application in the entertainment industry is still wide. 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. The extra requirement of the special devices may not only inhibit these systems' popularity in the entertainment industry, but also limit story richness and the number of participants who are able to appear on the stage simultaneously.

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 movie called *Background Movie* with audience's own high-realism 3D CG faces (see Fig. 1). The origin of this system is "Fifteen seconds of fame" in 1998 I proposed as user attainable real famous movie title. 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. Specifically, for the DIM movie, we define a successful experience as one in which the audience experiences greater dramatic presence and engagement, and more intensive emotional response than in

a traditional movie. Dramatic presence, originally presented in Kelso et al.'s study of staged actors [30], refers to an audience's sense of "being in dramatic situation." Engagement refers to a person's involvement or interest in the content or activity of an experience, regardless of the medium. Emotional response includes two dimensions: emotional arousal and valence. Emotional arousal indicates the level of activation, ranging from very excited or energized at one extreme to very calm or sleepy at the other, and emotional valence describes the degree to which an affective experience is negative or positive [31].

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; 1,640,000 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 within 15 seconds. So in case modeling failure is detected by attendant, it is possible to capture face again and again.

Figure 2 shows the processing flow of bringing audience into the pre-created movie.

The processing flow of *FCS* is as follows.

1. Under the theater attendant's guidance, participants put their faces on a 3D range-scanner window for capturing their facial information.
2. The captured facial images are transmitted to Face Modeling PCs through a Gigabit Ethernet.
3. The frontal face images of participants are transferred from Face Modeling PCs to Data Storage Server.
4. Here, 2D thumbnail images of participants are created and sent to Attendant PC from Data Storage Server. The theater attendant checks to see if the captured im-

Fig. 2 The processing flow of FCS.

ages are suitable to undergo modeling. If it is true, then the decision on face modeling is sent back to Face Modeling PCs.

5. In Face Modeling PCs, the participants' 3D facial geometry is first generated from the images projected with a rainbow stripe pattern and the images with a blue back for silhouette extraction. Then, Face Modeling PCs generate the personal CG face model from the 3D geometry and the frontal image. At this point, the personal key shapes for facial expression synthesis are also generated.
6. The personal face model, age and gender information used for assigning movie roles and the personal key shapes are transmitted to Data Storage Server.
7. Thumbnail images of participants' face model and information about age and gender is transmitted to Attendant PC from Data Storage Server. The attendant judges if the participant's face model is suitable to be appeared on the movie. If it is true, 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 or face model was not good enough, one of the default face model was replaced at Aichi Expo.2005 or it goes back to the process No.1 to capture the guest face again with careful instruction at HUIS TEN BOSCH from 2007.
8. The decisions on assigning roles to participants are transmitted to Data Storage Server. At this time, based on gender information, a pre-recorded male or female voice track is selected for these participants' CG characters.
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 Data 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.

In the following sub-sections, we present several key issues in the pipeline of bringing audience into the movie.

Fig. 3 3D scanner. Seven digital cameras (blue arrows) and two slide projectors (green arrows) are placed in a half-circle.

Fig. 4 The Captured Images. The top line is the images with a rainbow stripe pattern and the bottom line is the images taken under fluorescent lighting conditions. 3D geometry is reconstructed from these images.

Fig. 5 Face modeling result. (a) The frontal face image, (b) and (c) The 3D geometry.

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 seven 800×600 pixels CCD digital cameras and two slide projectors which are able to project rainbow stripe patterns (see Fig. 3). 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. 5 (a)). And the other six cameras on both sides capture facial images twice (i.e., with rainbow stripe patterns and under fluorescent lighting conditions) to reconstruct 3D geometry (see Fig. 4). The hybrid method with an Active-Stereo technique and a Shape-from-Silhouette technique [1] are used to reconstruct 3D geometry. Figure 5 shows a frontal face image and its corresponding 3D face geometry constructed by the scanner system.

Fig. 6 Personal face model. (a) 89 feature points, (b) the resampled 128 feature points and (c) personal face model.

Face Modeling

Face modeling is completed in approximately 2 minutes in Aichi Expo. version and 15 seconds in HUIS TEN BOSCH version of FCS system. Overall this process includes four aspects: facial feature point extraction, face fitting, face normalization and gender and age estimation.

Facial Feature Point Extraction

To generate participant's personal CG face using captured face texture and 3D facial geometry, it is necessary to extract the feature points of his/her face. First 89 facial feature points are extracted, including the outlines of the eyes, eyebrows, nose, lips, and face using the Graph Matching technique with Gabor Wavelet features [3], [4]. Figure 6 (a) shows the 89 facial feature points which are defined to create a Facial Bunch Graph. Then Gabor Wavelet features extracted from any facial images (including gender, ethnic characteristics, facial hair, etc) are registered into each node of the Facial Bunch Graph. To enable our feature point detector to correctly extract facial feature points, a total of 1500 facial images are collected as training data.

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 using a cubic spline curve and then resample 128 points on an approximated curve to achieve correspondence between the feature points. This process is performed for

eyes, eyebrows, and the outline of the lip. Figure 6 (b) shows the resampled 128 feature points.

Normalizing Face

There are individual differences in the face size and the face position and angle placing on the scanner window for each participant, whereas the size of the region where the participant's face will be embedded in the background movie is determined by the generic face model. Thus we must normalize all personal face models based on the generic one. We normalize personal faces from the three aspects (i.e., position, rotation and facial size) by Hirose [35]. To normalize position, we relocate the rotation center of each personal face model to the barycenter of all vertices from the outline of the eyes because our feature point detector can reliably extract these points. To normalize rotation, we estimate the regression plane from vertices on the personal face mesh and the normal vector of this regression plane can be assumed to be current direction of a face; then we estimate the rotational matrix that makes the angle between the normal vector and the z-directional vector approximate zero. To normalize face size, we adjust the size of personal face model, expanding or reducing in size in order to make the ratio between inner corners of the eyes equal to that of the generic face model. This eye ratio is also used to ensure that facial width remains proportionate. Figure 6 (c) shows an example of personal face model.

Gender and age estimation

The automation estimation of gender and age is performed by not-linear Support Vector Machine classifiers which classify Gabor Wavelet Features based on Retina Sampling [6]. The age information from the classifier's output is categorized into five groups: under 10, 10–20, 20–40, 40–60, and over 60. The estimation is used for assigning roles.

Controlling Real-time Character

This process of controlling CG characters includes synthesizing facial expressions based on blend shape techniques and relighting faces based on the light environment in each movie scene.

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 shape 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 ($\mathbf{D} = d_1, \dots, d_{34}$) between the neutral mesh (generic face

Fig. 7 Some examples of the personal key shapes.

mesh f) and the specific expressions' mesh (one of 34 key shapes) such as mouth opening. Some examples of key shapes are shown in Fig. 7.

Due to the large individual differences in the width and height of faces for each participant, we must customize the displacement vector. Thus, we introduce a weight w to automatically adjust magnitudes for each "generic displacement vector" by calculating the ratios of the face width and height between the individual face mesh and the generic one. The facial expression is synthesized using the formula (1).

$$b = f + \sum_{i=1}^{34} \beta_i w d_i, \quad (1)$$

Where b refers to a synthesized face; $w d_i$ refers to customized displacement vector; β_i refers to the pre-defined blending weight of each cast members' displacement vectors.

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 reflection. 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 laungo hair lighting effects by varying intensity on the outlines of the characters' faces.

Automatic Casting and Voice selection

FCS can automatically select a role from the background movie for each participant according to age and gender estimation, or a role can be assigned manually. In addition, we

Fig. 8 Rendering cluster system.

Fig. 9 An event that makes audience feel tension.

pre-recorded a male and female voice track for each character, and voices were automatically assigned to characters according to the gender estimation results.

Real-time Movie Rendering

To ensure high performance, reliability, and availability of the FCS, we implemented a cluster system to implement the 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) (see Fig. 8).

The IGPC and PPC are interconnected by Gigabit networks. The PPC first receives the control commands, time codes, and participants' CG character information (face model, texture and parameters for facial expressions) 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. Finally, the PPC receives these frame images and sends them to the projector in synchronization with the time codes.

Evaluation

FCS has been successfully exhibited from March 25th to September 25th at the 2005 World Exposition in Aichi, Japan, and showed a high success rate (93.5%) of bringing audience into the movie. However, this rate does not mean how many audiences can identify themselves on the screen. And in current FCS system, game factor to search for himself, his family or friends is main reason of popularity because it is not easy to identify him with unified size body without hair and pre-scored actor's voice. So immersive feeling to the story has not completely come true yet. However, the story "Grand Odyssey" itself sometimes makes audience feel high tension (Fig. 9).

4. Next Stage of DIM Movie

In the current 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 is 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. Now, the progress in GPU performance is so high that real-time rendering and motion control of full body and hair become possible recently.

Head modeling

To emphasize personality feature on character facial animation, a head and hair modeling are introduced. After scanning 3D range data of personal head with hair, a wisp model of hair is generated automatically from a personally adjusted skull 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.

Also motion capturing method of actor's hair and direction algorithm of hair dynamics by combining simulation and data driven approach are proposed [36], so high realism of hair motion can be generated on real-time process and it makes easier for audience to identify him in the story. Figure 10 shows current FCS character (a) and new character with hair (b). Figure 11 shows a result of hair simulation in a different wind force. This makes each character more attractive and impressive.

Sometimes, in case of special casting, arrangement or modification of hair style has to be covered. So after choosing a favorite hair style or appropriate hair model is selected, all of face model, head, jaw and ears model and adjusted hair model have to be combined automatically to casting model. Figure 12 shows an example of casting model with favorite hair model different from real personal hair style.

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.

Figure 13 shows GUI of facial muscle director plug-in: “Phy-Ace”. Creator can easily generate and direct a target faical expression in a very short time. In this system, after

Fig. 10 Comparison of character model.

Fig. 11 Hair simulation in a different wind force.

Fig. 12 An arranged casting model.

(c)2003松本富士/プラネット・銀河鉄道管理局
(c)2006 早稲田大学

Fig. 13 GUI of facial muscle director.

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 characteristic can be generated in a same face model. How to optimize the location and physical character of each muscle based on a captured face image without any landmarks on face is next our research target. By considering transient

Fig. 14 Character model with body variation.

feature of physics equation of face muscle, a delicate nuance of dynamic expression and lip-sync can be automatically generated.

About a skin reflectance model, an instant and multi wave length BRDF and BSSRDF capturing system is under construction. To make a precise model of furrow and bump in face motion, updating texture from high speed video has to be introduced without any landmarks on the face.

Body and Gait modeling

Character model with a variety of body size is pre-constructed and optimum size is chosen according to the captured silhouette image. Figure 14 shows examples of body variation in size and height. By fitting the size of neck, head, face and hair to the body size, a full body character model is assembled. Also personality in gait is modeled by frequency feature patterns of joint angles in a generic body skeleton model. Figure 15 shows a gait motion capturing scene on treadmill and the graph shows time series of joint angle in shoulder (upper line) and hip (lower line). Gait personality is modeled by PCA of captured joint angle motion vectors in generic full body skeleton model. Personal gait feature is estimated by the silhouette captured by video cameras from a variety of angles (Fig. 16).

Voice modeling

In current FCS system, a prerecorded voice of either an actor or actress is used as a substitute for that of each participant. The substitute voice is selected depending on 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

Fig. 15 Gait motion capturing and modeling.

Fig. 16 Gait silhouette image in difference angles.

know. Therefore we decided to focus on selecting the similar speaker from speech database to reduce this discomfort feeling.

A method to measure the perceptual similarity of speech is proposed, because it is impossible to record all speech of participants in advance, and to convert voice quality sufficiently with the present conversion technology.

Speaker recognition which deals with the similarity of speech has been researched and applied to several fields such as security field. In the security field, the similarity of speakers is generally determined based on likelihood of Gaussian Mixture Model (GMM) between speakers. However, the perceptual similarity is focused rather than the similarity of speaker models. Additionally, the aim of speaker recognition is to perceive oneself, not to search a similar speaker. Meanwhile, as for the relation between the perceptual similarity and the acoustic distance, Amino et al. [37] proved the strong correlation between the cepstral distance and the perceptual similarity. Nagashima et al. proved the strong correlation between spectrum distances at 2–10 kHz and the perceptual similarity of speech in which utterance speed and intonation were controlled by speakers [38]. Because the personality of speech appears not only in voice quality but also in utterance speed or intonation, the acoustic feature is examined for the estimation of perceptual similarity. As an evaluation result, combination of GMM and DTW is almost good for the perceptual speaker similarity criterion. Now a huge database is constructed for choice of similar voice samples depending on wide variety of hometown, age and gender.

5. Conclusions and Future Work

In this paper, an innovative visual entertainment system *DIM* is presented. *DIM* has two key features. First, it

provides audiences with an amazing experience: the ability to take part in a movie as a 3DCG member of the cast co-starring alongside famous movie stars, singers and other celebrities in film, anime and CG movies. Additionally, they are able to sit down and watch a movie in which they themselves are the stars. Second, *DIM* can automatically perform all of the processes required to make this possible; from capturing visitor's facial images and recreating them as 3DCG characters, to actually running the movie. The face modeling process, the core technology of our system, can automatically generate visitors' CG faces with an accuracy of 93.5%. The rendering system is robust enough to show movies without dropping any frames. The current version of the Future Cast System provides audiences with a movie in which only their faces are embedded. As the next DIM system, how to improve the audience's sense of personal identification while watching the movie has to be considered. The key lies in generating CG characters which share the visitors' physical characteristics, such as hairstyle, hair motion, body shape, body motion, skin characteristics, and even accessories such as glasses.

Moreover, physiological evaluation is now considered. It could help to identify perceptually meaningful parameters of new entertainment algorithms, which in turn can lead to more efficient and effective character generation technologies. The detail of DIM project is as follows.

<http://www.diveintothemovie.net/en/index.php>

Acknowledgments

This study is supported by the Special Coordination Funds for Promoting Science and Technology of Ministry of Education, Sports, Science and Technology of Japan. I give special thanks to all members and supporters of *Dive into the Movie* project.

References

- [1] K. Fujimura, Y. Matsumoto, and T. Emi, "Multi-camera 3d modeling system to digitize human head and body," Proc. SPIE, pp.40-47, 2001.
- [2] L. Wiskott, J. Fellous, N. Kruger, and C. der Malsburg, "Face recognition and gender determination," Proc. International Workshop Automatic Face and Gesture Recognition, pp.92-97, 1999.
- [3] M. Kawade, "Vision-based face understanding technologies and applications," Proc. 2002 International Symposium on Micromechanics and Human Science, pp.27-32, 2002.
- [4] L. Zhang, H. Ai, S. Tsukiji, and S. Lao, "A fast and robust automatic face alignment system," Tenth IEEE International Conference on Computer Vision 2005 Demo Program, 2005.
- [5] B. Moghaddam and M. Yang, "Gender classification with support vector machines," Proc. Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp.306-311, 2000.
- [6] H. Lian, B. Lu, E. Takikawa, and S. Hosoi, "Gender recognition using a min-max modular support vector machine," International Conference on Natural Computation 2005, vol.2, pp.438-441, 2005.
- [7] F.I. Parke, "Computer generated animation of faces," ACM'72: Proc. ACM Annual Conference, 1972.
- [8] M. Lyons, A. Plante, S. Jehan, S. Inoue, and S. Akamatsu, "Avatar creation using automatic face processing," Proc. Sixth ACM International Conference on Multimedia, pp.427-434, 1998.
- [9] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D.H. Salesin, "Synthesizing realistic facial expressions from photographs," SIGGRAPH '98: Proc. 25th Annual Conference on Computer Graphics and Interactive Techniques, 1998.
- [10] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," International Conference on Computer Graphics and Interactive Techniques, Proc. 26th Annual Conference on Computer Graphics and Interactive Techniques, pp.187-194, 1999.
- [11] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," Comput. Graph. Forum, vol.22, no.3, pp.641-650, 2003.
- [12] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," SIGGRAPH '02: Proc. 29th Annual Conference on Computer Graphics and Interactive Techniques, pp.388-398, 2002.
- [13] Y. Lee, D. Terzopulus, and K. Waters, "Realistic modeling for facial animation," Proc. SIGGRAPH'95, pp.55-62, 1995.
- [14] B. Choe, H. Lee, and H. Ko, "Performance-driven muscle-based facial animation," Proc. Computer Animation, pp.67-79, 2001.
- [15] E. Sifakis, I. Neverov, and R. Fedkiw, "Automatic determination of facial muscle activations from sparse motion capture marker data," ACM Trans. Graphics, vol.24, no.3, pp.417-425, 2005.
- [16] D. DeCarlo and D. Metaxas, "The integration of optical flow and deformable models with applications to human face shape and motion estimation," CVPR '96: Proc. 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96), pp.231-238, 1996.
- [17] P. Joshi, W.C. Tien, M. Desbrun, and F. Pighin, "Learning controls for blend shape based realistic facial animation," 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp.187-192, 2003.
- [18] J.X. Chai, J. Xiao, and J. Hodgins, "Vision-based control of 3d facial animation," SCA '03: Proc. 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp.193-206, 2003.
- [19] D. Vlasic, M. Brand, H. Pfister, and J. Popovi, "Face transfer with multilinear models," ACM Trans. Graphics, vol.24, no.3, pp.426-433, 2005.
- [20] J.Y. Noh and U. Neumann, "Expression cloning," SIGGRAPH '01: Proc. 28th Annual Conference on Computer Graphics and Interactive Techniques, pp.277-288, 2001.
- [21] A. Krishnaswamy and G.V.G. Baranowski, "A biophysically-based spectral model of light interaction with human skin," Comput. Graph. Forum, pp.331-340, 2004.
- [22] P. Debevec, T. Hawkins, C. Tchou, H.P. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," SIGGRAPH '00: Proc. 27th Annual Conference on Computer Graphics and Interactive Techniques, pp.145-156, 2000.
- [23] G. Borshukov and J.P. Lewis, "Realistic human face rendering for the matrix reloaded," ACM SIGGRAPH 2003 Technical Sketch, 2003.
- [24] P.V. Sander, D. Gosselin, and J.L. Mitchell, "Real-time skin rendering on graphics hardware," SIGGRAPH '04: ACM SIGGRAPH 2004 Sketches, p.148, 2004.
- [25] T. Weyrich, W. Matusik, H. Pfister, B. Bickel, C. Donner, C. Tu, J. McAndless, J. Lee, A. Ngan, H.W. Jensen, and M. Gross, "Analysis of human faces using a measurement-based skin reflectance model," SIGGRAPH '06: ACM SIGGRAPH 2006 Papers, pp.1013-1024, 2006.
- [26] M. Cavazza, F. Charles, and S.J. Mead, "Interacting with virtual characters in interactive storytelling," Proc. 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems: Part 1, pp.318-325, 2002.
- [27] A.D. Cheok, W. Weihua, X. Yang, S. Prince, F.S. Wan, I.M. Billinghurst, and H. Kato, "Interactive theatre experience in embodied + wearable mixed reality space," Proc. International Symposium on Mixed and Augmented Reality (ISMAR'02), 59-317, 2002.
- [28] M. Mateas and A. Stern, "Facade: An experiment in building a fully-realized interactive drama," Proc. Game Developers' Conference: Game Design Track, 2003.

- [29] N. Szilas, "The future of interactive drama," Proc. 2th Australasian Conference on Interactive Entertainment, pp.193–199, Sydney, Australia, 2005.
- [30] M. Kelso, P. Weyhrauch, and J. Bates, "Dramatic presence," *Presence: Teleoperators and Virtual Environments*, vol.2, no.1, 1992.
- [31] P.J. Lang, "The emotion probe," *American Psychologist*, vol.50, pp.372–385, 1995.
- [32] N. Arad and D. Reisfeld, "Image warping using few anchor points and radial functions," *Comput. Graph. Forum*, pp.35–46, 1995.
- [33] J.C. Carr, R.K. Beatson, J.B. Cherrie, T.J. Mitchell, W.R. Fright, B.C. McCallum, and T.R. Evans, "Reconstruction and representation of 3D objects with radial basis functions," *SIGGRAPH '01: Proc. 28th Annual Conference on Computer Graphics and Interactive Techniques*, pp.67–76, 2001.
- [34] S. Lee, G. Wolerg, and S.Y. Shin, "Scattered data interpolation with multilevel b-spline," *IEEE Trans. Vis. Comput. Graphics*, vol.3, no.3, pp.228–244, 1997.
- [35] M. Hirose, *Application for Robotic Vision of Range Image Based on Real-time 3D Measurement System*, Ph.D. Thesis, Chukyo University, 2002.
- [36] Y. Kazama, E. Sugisaki, and S. Morishima, "Directable and stylized hair simulation," *Proc. International Conference on Computer Graphics Theory and Applications 2008*, no.72, 2008.
- [37] K. Amino, T. Sugawara, and T. Arai, "Speaker similarities in human perception and their spectral properties," *Proc. WESPAC IX*, 2006.
- [38] I. Nagashima, M. Takagiwa, Y. Saito, Y. Nagao, H. Murakami, M. Fukushima, and H. Yanagawa, "An investigation of speech similarity for speaker discrimination," *Acoustical Society of Japan 2003 Spring Meeting*, pp.737–738, 2003.

Shigeo Morishima was born in Japan on August 20, 1959. He received the B.S., M.S. and Ph.D. degrees, all in Electrical Engineering from the University of Tokyo, Tokyo, Japan, in 1982, 1984, and 1987, respectively. From 1987 to 2001, he was an associate professor and from 2001 to 2004, a professor of Seikei University, Tokyo. Currently, he is a professor of School of Science and Engineering, Waseda University. His research interests include Computer Graphics, Computer Vision, Multimodal Signal Processing and Human Computer Interaction. Dr. Morishima is a member of the IEEE, and ACM SIGGRAPH. He is a leader of Human Communication Science Special Interest Group in IEICE-J and a trustee of Japanese Academy of Facial Studies. He received the IEICE-J achievement award in May, 1992 and the Interaction 2001 best paper award from the Information Processing Society of Japan in Feb. 2001. Dr. Morishima was having a sabbatical staying at Visual Modeling Group, Department of Computer Science, University of Toronto from 1994 to 1995 as a visiting professor. He is now a temporary lecturer of Meiji University, Japan. Also he is a visiting researcher of ATR Spoken Language Communication Laboratory from 2001. He is a leader of Digital Animation Laboratories from 2004.

processing and Human Computer Interaction. Dr. Morishima is a member of the IEEE, and ACM SIGGRAPH. He is a leader of Human Communication Science Special Interest Group in IEICE-J and a trustee of Japanese Academy of Facial Studies. He received the IEICE-J achievement award in May, 1992 and the Interaction 2001 best paper award from the Information Processing Society of Japan in Feb. 2001. Dr. Morishima was having a sabbatical staying at Visual Modeling Group, Department of Computer Science, University of Toronto from 1994 to 1995 as a visiting professor. He is now a temporary lecturer of Meiji University, Japan. Also he is a visiting researcher of ATR Spoken Language Communication Laboratory from 2001. He is a leader of Digital Animation Laboratories from 2004.