Face Recognition using Computer-Generated Database

Won-Sook LEE
School of Information Technology and Engineering
University of Ottawa, 800 King Edward Ave
Ottawa, Ontario, K1N 6N5, Canada
Tel: (1-613) 562-5800 ext 6227
Fax: (1-613) 562-5664
wslee@uottawa.ca

Kyung-Ah SOHN
Human-Computer Interaction lab
Samsung Advanced Institute of Technology
San 14-1, Nongseo-ri, Giheung-eup
Yongin-si, Gyeonggi-do, Korea 449-712
Tel: (82-31) 280-1756
Fax: (82-31) 280-9257
kasohn@sait.samsung.co.kr

Abstract

Many face database and recognition systems have been constructed in specially designed studios with various illuminations, poses, and expressions. However, none of these databases yet satisfies a large variation of poses and illuminations that permit the study of systematic 3D human face information, which results in an unsatisfactory success rate. It is caused primarily by the difficulty of data collection of facial images to satisfy the large variation of poses and illuminations to fully represent the 3D characteristics of human faces. We present how computer-generated database can be used for face recognition experiment focused on multi-view face recognition/descriptor. We show multi-view face data collection using rendering of 3D models. We also illustrate our approach how to build a face descriptor containing 3D information of human face using multi-view concepts. This multi-view face recognition descriptor is a 3D descriptor using the concept how much powerful a view influences over nearby views, so called as "quasi-view" size.

1. Introduction

A variety of techniques borrowed from computer vision have been brought to bear on the computer graphics problems inherent in synthesis and animation. Nevertheless, these areas of computer graphics, though relatively young compared to computer vision, have made dramatic progress in recent decades in their central technology, that of rendering [4], and have also seen an entry into virtual human and virtual environment simulation [21][1][19], which results that computer graphics are now providing technology to other areas, such as computer vision and medical imaging.

The fusion and synergy of computer graphics (virtual) and computer vision (real) has given rise to intense activity. By the late 90's, face/body modeling and animation on one hand, and face/human detection and recognition in the other, were separate endeavors, but had started to merge their technologies. For instance, the quality of human face modeling has been enhanced using computer vision technology and the statistical approach to face recognition technology, whereas traditional computer graphics relied on conventional geometrical and deformation approaches. Supplementary views to provide complete 3D information have been critical, but the statistical approach based on a large database of conventional shapes has shown potential for determining individual face shapes with limited information (views). Conversely, computer vision research on face/human detection/recognition has benefited from computer animation technology for the extraction and description of moving shapes using 3D human anatomy. For the example of faces, see [21][2][18][3] and for bodies see [15].

When we closely examine the classic face recognition problem, one of the most significant extrinsic factors in variety of different people's faces is the pose, or the viewing direction at which the face is seen and illuminated. This is the basic reason that the typical face description algorithms based on a single view, especially frontal view, which has a fundamental limitation in its performance in a real situation. Some approaches such as novel view generation with given limited view [13] were proposed but reveal limitations in obtaining a high success rate. This is because the face is a 3D-object and it cannot be reconstructed from

2D information. There are other approaches to describe the 3D characteristics of human face for recognition purpose such as building a 3D model the subject. Some approaches using 3D morphable models [2] obtained using a laser scanner show quite good result but due to their heavy computational costs, it is still unavailable publicly. A 3D model reconstruction with stereoscopic camera is also used to retrieve the face identity [10], but the successful result has not come out yet.

1.1. Our Approach

Here we discuss how a computer generated 2D database can be used for face recognition. The 2D database is derived from the 3D database of shapes captured from the real world. The raw date from capture equipment is usually neither directly usable nor error-free; all the more reason for the great investment we plan in 3D data collection, especially the texture terms, based on realistic rendering. Our approach not only economizes data collection time and reduces the expense of a dedicated studio set up, but it also avoids "over-fitting" to any physically-fixed environment by varying the virtual environment.

Our approach for describing a face is by the use of a 2D-viewpoints mosaic [16] and we show which criteria are used to extend one-view to multi-views. Here, the multi-view face descriptor contains the 3D information, in a compact form, of all views of horizontal rotation [-90° … +90°], and vertical rotation [-30° … +30°]. The database for this multi-view face descriptor must cover a wide pose variation with a dense pose distribution. Many kinds of face databases already exist but none of them provides a satisfactory view range. Instead of building a complicated and high-cost studio for data collection, we used a simple method with 3D mesh models.

In Section 2, we analyze problems with existing face databases, with Section 3 devoted to 3D capturing technologies. Section 4 details the face recognition database construction using a rendered face image database with captured 3D facial models and includes a description of the multi-view face data collection procedure. Section 5 shows the experimental results using a 3D face descriptor, using systematic extension of a 2D face descriptor to multi-view one using quasi-view concepts. Section 6 concludes the paper, with a review of future work.

2. Face databases built in studio

2.2. Existing face databases

Many 2D face images have been obtained in a studio built entirely for that specific purpose. Some selected examples of face databases intended to cover poses and illumination are KISA [23], YALE [6], and CMU [22]. Cameras and lights are placed in different positions to simulate pose and light variances, this quickly produces satisfactory images. Nevertheless, certain constraints exist due to the studio setup: the total number of poses equals to the actual number of cameras, and hence limited number of pose variance such as 7 to 13 pose images in these examples are far too small to perform experiments for multi-view face recognition.

2.3. Ideal pose independent face recognition

Ideal posed face recognition means to cover the wide cases in terms of people and environment, in other words intrinsic and extrinsic parameters. We focus on extrinsic parameters in this paper. Two main extrinsic parameters are pose and illumination variations. For the pose, the training set must be complete in every view covering any view of horizontal rotation [$-90^{o}… 90^{o}$] and vertical rotation [$-30^{o}… 30^{o}$] ~ [$-60^{o}… 60^{o}$] as a view-mosaic in *Figure* 1 while registration can be made with only selected views. A homogenous background is more suitable, and permits the automatic segmentation using chroma-keying, so that the background can be replaced later. For the illumination, many lights, for example 64 lights, must be used to control various illumination environments.

Figure 1: Ideal face recognition region in terms of poses.

3. Capturing systems

We discuss about technologies to capture the 3D positional data of a human face. The three main methodologies for capturing shape are: laser scanners such as Cyberware Color Digitizer™ [8], an active light stripper [21], or a stereo camera [5][25]. They suffer from various problems and often reproduce the highly reflective surface of hairy parts (such as hair and eyebrows) and the shady regions (such as bottom

of nose and between two lips holes) as holes on the final model. Insufficient texture information, non-uniform material property or bad lighting conditions also result in a poor reconstruction quality. Distortion and occlusion also make the problem more difficult and computationally expensive.

3.1. Using 3D Mesh Data

No capture system is capable of producing error-free raw data; hence post-processing is always required. For example, 3D Studio Max [9] provides functionality to automatically fill gaps, but it can also remove certain characteristics from the mesh data. One of more human face shape oriented methodology is using an adaptive mesh methodology. The adaptive mesh method not only removes gaps, but it also provides a mesh structure pre-designed for the human face.

Lee et al [17] propose a mesh adaptation method that uses a two-step matching procedure. Initially the geometrical deformation is created using a rough match with the help of feature points, this is then refined using sophisticated projections on all surface points from the raw data. The feature information obtained in the rough matching step provides the partition of surface points on the raw data, so that the projection can be done either to the raw data surface or to the deformed surface with only feature points. This method provides satisfactory results even on faces where large errors were shown in raw captured data; *Figure* 2 exemplifies this procedure.

Figure 2: Raw data containing large error is processed to be error-free smooth surface as well as feature information on surface to able to facial expression.

4. Computer-generated face recognition database

In this section, we demonstrate how computer generated rendering images with 3D captured facial mesh models can be used to build face recognition systems. Firstly a multi-view face data collection procedure is described, followed by a detailed explanation of the 3D face descriptor construction; this

takes systematic extension of a 2D face descriptor to a multi-view using quasi-view concepts.

One way to collect large range of poses and also illumination is using rendering of 3D facial mesh models. We apply rotation of the desired degree to the mesh and then render it using 2D projection. The advantage of this method is that it ensures exact rotation degree of the face, various illumination conditions can be simulated virtually and hence any background can be used. The database contains 108 laser scans with texture images of size 480 by 400. The model is aligned to the front view using the information of a nose and two eyes found as facial features in the texture image; these have been directly calculated by their 3D locations on a mesh model, and then rendered with the desired view direction.

4.1. Detection of Features

Texture images of our mesh models are carefully aligned in the process of scanning, that is, the face center is located in the middle of the image and its orientation is upright as in the first image of Figure 3. A heuristic approach using gray values is used to find the features. First, the minimal rectangular search region for nose location is set and then summates the gray pixel values of each row in the region. The sum around the nose-end row will be the smallest, or the darkest, due to nostrils and shadows. The darkest pixel block along that row corresponds to the real nose-end position. With similar methods, the left and right eye positions are captured in separate search regions. For the eyes, search regions are defined automatically from the acquired nose position. And column-wise summation is initially performed to reduce false detection caused by the eyebrow. In post-processing step for correction, if the height of two detected eyes differs too much, the higher detection is from eyebrow regions, hence, moving the search space to lower region. If the detected eyes are too close, they are from nostrils, not real eyes, therefore moving the search region higher. By reexamining the updated search region, we were able to locate feature positions successfully for each image in the database.

4.2. Alignment of the 3D Mesh Model

Due to the fact we know the 3D coordinates of eyes and nose, say, $p_{left\_eye}, p_{right\_eye}, p_{nose}$, from their locations in the image, the alignment process is quite straightforward: if we define $v_1 = p_{left\_eye} - p_{right\_eye}$, $v_2 =$ the vector from p_{nose} and orthogonal to v_1 ,

which is similar to $(p_{right\_eye} + p_{left\_eye})/2 - p_{nose}$, and $v_3 = v_1 \times v_2$, the rotation given by $\left[\dfrac{v_1}{\|v_1\|} \dfrac{v_2}{\|v_2\|} \dfrac{v_3}{\|v_3\|}\right]$ will transform the model to near-frontal view. The resulting model, however, is slightly tipped forward for the up-vector approximation and is not parallel to the real up-direction in the face model; therefore we rotate the model to upper view using a fixed angle. With the face being aligned and we are able to render it using an arbitrary view direction. We align the face location on the image by fixing two eye locations for each view as shown in the last column of Figure 3.
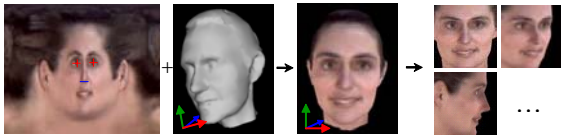

Figure 3: Detected feature points in the texture image are mapped to the 3D mesh model, which is aligned to the frontal view and then rendered with various view directions

4.3. Definition of the Face Localisation

The images have size 56x56 and the positions of two eyes in the front view are on (0.3, 0.32) and (0.7, 0.32) when width and height are considered as 1.0. For example, (0.3, 0.32) means coordinates (17, 18), which is 17th pixel in horizontal coordinate from left and 18nd pixel in vertical coordinate from top for this image size case. The left eye position of the positive horizontal rotation is retained at (0.3, 0.32) while the right eye position of the negative rotation does (0.7, 0.32). Vertical rotation has the same eye positions as the ones on zero vertical rotation images. Figure 4 shows one example of face localization (face alignment) definition.
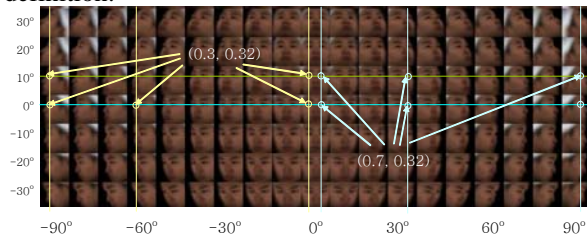

Figure 4: eye-positions on view-mosaic of the center face of rendered images of 108 3D-facial mesh models

(a)   *(b)*

Figure 5: (a) Rendered images of a virtual face in unlimited illumination conditions and poses. (b) Face images taken in pre-built studio with 576 viewing conditions

A big advantage using 3D mesh data is the freedom to easily set the environment; some examples of various rendered backgrounds are shown in Figure 6 without the need of chroma-keying.

Figure 6: different background options: Model is oriented with horizontal ration 70º and the light source is set with horizontal rotation 40º with directional light.

A comparison in illumination variation between rendered images from 3D models in a virtual environment and photographs taken using 64 lights [6] is also shown in Figure 5. We attempted to simulate similar results with the Yale Face Database [6] taken in a pre-built studio, but with larger horizontal rotation to show the freedom of any poses and any illumination possible as shown in Figure 5. We have used a non-commercial rendering software POV-Ray [11] to create shadow. As a reference, the Yale Database [6] used 64 light cones in a sphere shape studio and a total 9 poses of 10 people were captured; five poses at 12° and three poses at 24° from camera's axis.

5. Experimental Results

Using a statistical approach is one of the most popular methods for face recognition methods, and many approaches are based on Eigenfaces [20], which uses Principle Component Analysis as a feature extraction method. Conventional statistical face recognition systems have a training step and test step. The training step is used to build a feature extraction system with given images that maximize the difference between various identities while minimizing the (intrinsic and extrinsic) variation of one identity. The test step is used to perform classification in given images using the feature extraction obtained in the training step. Better yet, it classifies unknown face image to the proper identity, with a higher accuracy of face recognition. Most systems, based on training-test, often suffer from lack of variety and over-fitting in given in circumstances. With the help of freedom in virtual environment, we are able to create a large range of various posed facial images. Figure 7 shows rendered images from 3D facial model in various view points.
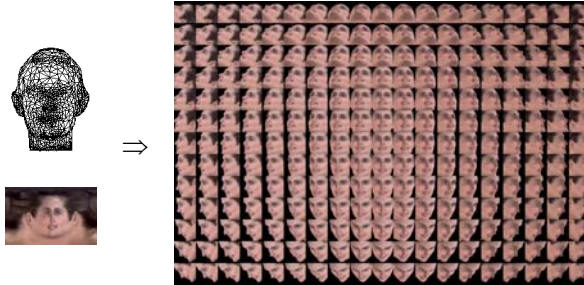

Figure 7: Rendered images from a 3D facial model in various view points.

With many 3D models, we can generate numerous images in any virtual environment. However there is another problem we have to consider to build a face descriptor which is able to recognize the identity. There have been various approaches focused on front view face recognition and we need to take systematic extension to build all-view-face recognition. With large-pose-variation database, we create a multi-view face descriptor for any-view face recognition. If we register every 10º apart, we have to register 19 x 7 = 133 views to cover horizontal rotation [-90° … +90°] and vertical rotation [-30° … +30°], which results $133 \times$ single view descriptor size. Then the descriptor becomes too large. In this paper, we focus on feature extraction and descriptor optimization among many issues in multi-view face recognition.

As we learn from frontal-view face descriptors, a registered view is used to retrieve nearby views (quasi-frontal) with high retrieval rate and we extend the concept of quasi-frontal to quasi-view, from frontal view to general view.

Definition: Quasi-view with error rate K - Quasi-view V^q of a given (registered) view V with error K means faces on V^q are retrieved with V with error retrieval rate less than or equal to K.

5.1. Feature Extraction

We use a modified version of the best algorithm in the MPEG-7 advanced face descriptor XM which is selected in competition among various algorithms to retrieve faces. More details are found in MPEG document [24].
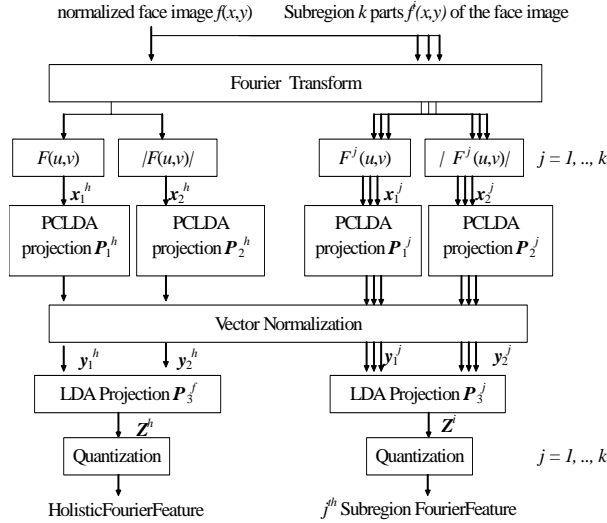
Figure 8: Feature extraction used for Multi-view 3D Face Descriptor

Our Subregion-based Linear Discrimination analysis (LDA) on Fourier space as shown in Figure 8 is designed for multi-view purpose. The biggest differences between the MPEG-7 XM [12] and our model are (i) feature extraction in luminance space is removed, (ii) the subregion decomposition, which was in luminance space, is now in Fourier space, and (iii) the number and positions of subregions are varied depending on a given view. The first two modifications give more efficient feature extraction with smaller descriptor size. The third modification is caused by the non-frontal-view extension. The new view-dependent subregion is shown in Figure 9.
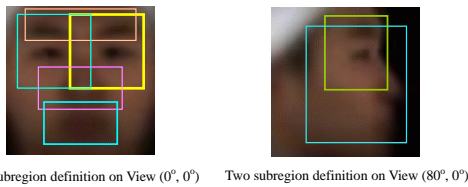

Five subregion definition on View (0º, 0º)   Two subregion definition on View (80º, 0º)

Figure 9: Subregion definition depending on views is shown as superimposed on the center face of our rendering face database

5.2. Quasi-view

Graham and Allinson [7] have calculated the distance between faces over pose to predict the pose dependency of a recognition system. The faces are further apart they will be easier to recognize using distance measures in the Eigenspace and consequently, the best pose samples to use for an analysis should be concentrated around this range. Note that they have checked only horizontal rotation of human heads.

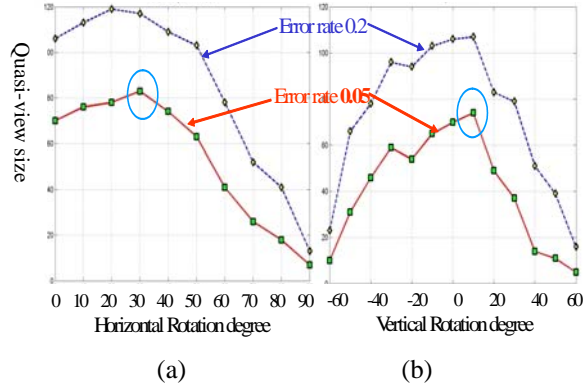
(a)   (b)

Figure 10: Quasi-view sizes depending on a registered view.

Quasi-view size also depends on the view. We experiment quasi-view inspection with error rate 0.05. *Figure 10* (a), which has very similar pattern to the ones in [7], shows how the quasi-view size varies with horizontal and vertical rotations of a head. To make a fair comparison between different views, we extracted 24 holistic features (without using subregion features) for each view. The views (20°, 0°) ~ (30°, 0°) have both the largest quasi-view size and the largest Euclidean distance between the people in Eigenspace among views (0°, 0°), (10°, 0°), ... and (90°, 0°). *Figure 10* (b) shows the views (0°, 0°) ~ (0°, 10°) have the biggest quasi-view size among views (0°, -60°), (0°, -50°), and (0 °, 60°). One point to note is that the head-down views have bigger quasi-view size than ones of the head-up and it demonstrates it is easier to recognize people when they look downward more than they look upward.

5.3. Descriptor optimization

Optimization criteria depends on where the emphasis is placed, such as the smallest number of registration poses, the smallest descriptor size, or the biggest coverage of poses by selecting views efficiently describing the 3D of the facial features.

Our first experiment was made using 50/50 ratios with rendered images from 3D mesh models; this means that half the images were used for training and the other half for test. We concentrate on the selection by quasi-view size, the number of subregions, and the number of features in each subregion. So for some views, 5 holistic features and 5 features for 5 subregions are extracted and for other views 5 holistic features and 2 features for 5 subregions are extracted. If a view is close to profile, we use 5 holistic features

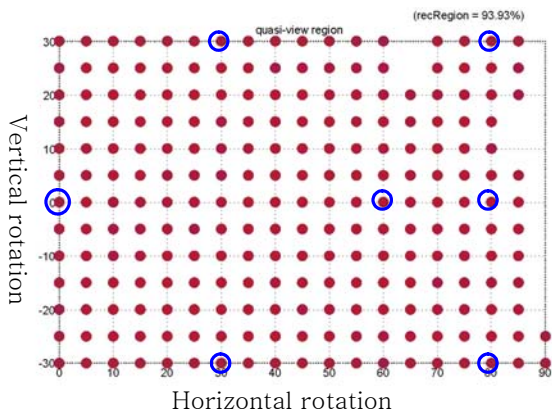and 5 features for 2. For one view, one image is selected.

Figure 11: Rendered image experiment: The region (93.93%) covered by 7 quasi-views in the view-mosaic of positive horizontal rotation with error rate 0.05. Registration with 13 views are enough to retrieve over 90% of views of horizontal rotation [-90° ... 90°] and vertical rotation [-30° ... 30°] with error rate 0.05.

Through experiments with various combinations of quasi-views, a set of views is selected to create multi-view 3D descriptor. An example of several possible descriptors shown in *Figure* 11 has 240 dimensions with rendered images. This descriptor is able to retrieve the rendered images in the test database with error rate 0.05 covering 93.93 % views of total region of view-mosaic of horizontal rotation [-90°… 90°] and vertical rotation [-30° … 30°]. For a reference, it covers 95.36% for error rate 0.1, 97.57% for error rate 0.15 and 97.98% for error rate 0.2.
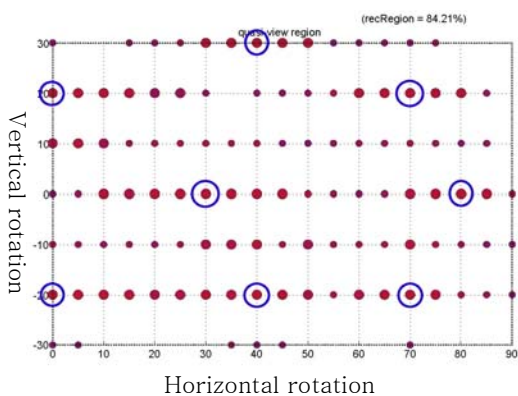

Figure 12: Video image experiment: The region (84.21%) covered by 14 quasi-views in the view-mosaic of horizontal rotation [-90°...90°] and vertical rotation [-30°...30°] with error rate 0.2.

For comparison, video images taken in real situation are also experimented in the same processes. Here the DB size was 114 with 50/50 ratio, which means 57 identities used for training and 57 identities for test. We position a video camera at the front of a human subject and video streams are taken by rotating the human subject's chair, so we were able to obtain images of horizontal at interval 1°. We have taken 7 video streams of each human subject where each time the human's neck is fixed in different vertical rotation. Due to the difficulty to collect large variation of poses in real environment, only 10° apart vertical rotation images are taken while the horizontal rotation is as dense as the rendered images. Experiment with images extracted from video images has a lower success rate. The training data does not have pose information as accurate as the rendered images from the 3D-facial mesh models, especially vertical axis of views. In addition the human subject has been sat for a while and rotated to take video streams, so facial, hair and body movement and deformation exist. So the quasi-view size of video images is smaller than the one of rendered images. An example of many possible descriptors is shown in Figure 12. Other examples are 300 dimensions with 87.22% coverage and 270 dimensions with 85.71% coverage with the same error rate.

As a reference for dimension, the current Advanced Face Descriptor [24][12] has 48 dimensions with error rate, called as ANMRR, 0.3013 and 128 dimensions with ANMRR 0.2491 for 50/50 ratio for photograph images. Here ANMRR is a MPEG definition for error rate. AFD is verified only with limited pose variation such as about (±30°, 0°) and (0°, ±30°), which means the coverage of view is very small compared to our 180° horizontal rotation and 60° vertical rotation.

6. Conclusion

We have clearly shown the usage of computer-generated rendered images with captured 3D mesh facial data for face recognition purpose and multi-view face descriptor methodology which is a systematic extension of conventional one-view face recognition descriptor. The database construction for very dense poses using a rendering of a 3D facial mesh obtained from a laser scanner is explained and also a comparison with video images are also given. Our multi-view approach with quasi-view dedicated to human face structure shows acceptable size of descriptor and satisfactory recognition rate for recognizing identity of human face in any view.

Virtual environments are far easier to manipulate than physically built-in environments to produce variety in extrinsic parameters like poses and illumination, which has been bottle-neck in database construction for face recognition. Our experiment for multi-view face recognition shows the potential of CG models in various applications. As the computer-generated 3D mesh technologies progress to produce more and more realistic human face modeling and animation, the intrinsic parameters such as facial expression, hair, and aging can be simulated in virtual environment for face recognition purpose as well as extrinsic parameters we have discussed in this paper.

7. References

[1] Badler, N., Allbeck, J. and Bindiganavale, R., "Describing human and group movements for interactive use by reactive virtual characters," Army Science Conference, 2002

[2] Blanz, V. and Vetter, T. "A Morphable Model for the Synthesis of 3D Faces", SIGGRAPH'99 Conference Proceedings, pp. 187-194

[3] Blanz, V. and Vetter, T., "Face Recognition Based on Fitting a 3D Morphable Model", IEEE Transactions on Pattern Analysis and Machine Intelligence archive, Volume 25, Issue 9, September 2003

[4] Buehler, C., Bosse, M., Gortler, S., Cohen, M. and McMillan, L., "Unstructured Lumigraph Rendering," Proc. ACM SIGGRAPH '01, pp. 425-432, 2001.

[5] Chen, Q. and Medioni, G., "Building Human Face Models from Two Images", IEEE 2nd Workshop Multimedia Signal Processing, pp. 117-122, Dec. 1998.

[6] Georghiades, A.S., Belhumeur, P.N. and Kriegman, D.J., "From few to many: Generative models for recognition under variable pose and illumination", In Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, 2000.

[7] Graham, D. B. and Allinson, N. M., "Characterising Virtual Eigensignatures for General Purpose Face Recognition" in Face Recognition: From Theory to Applications (H. Wechsler, PJ Phillips, V. Bruce, FF Soulie and TS Huang, eds.), Berlin: Springer-Verlag, pp. 446-456, 1998

[8] http://www.cyberware.com/wb-vrml/

[9] http://www.discreet.com/3dsmax/

[10] http://www.geometrix.com

[11] http://www.povray.org/

[12] Kamei, T., Yamada, A., Kim, H., Hwang, W., Kim, T.-K. and Kee, S. C., "CE report on Advanced Face Recognition Descriptor", ISO/IEC JTC1/SC29/WG11 M9178, Awaji, JP, December 2002

[13] Kim, W.-Y., Lee, J.-H. Park, H.-S. and Lee, H.-J.,"PCA/LDA Face Recognition Descriptor with Pose", MPEG-7 video, Oct. 2002.

[14] Lee Y., Terzopoulos D. and Waters K., "Realistic Modeling for Facial Animation", In Computer Graphics (Proc. SIGGRAPH), ACM Press, pp. 55-62, 1995.

[15] Lee, M. W., Cohen, I., Jung, S. K., "Particle Filter with Analytical Inference for Human Body Tracking," IEEE Workshop on Motion and Video Computing, 2002.

[16] Lee, W.-S. and Kee, S. C., "3D-Face Descriptor: proposal for CE," ISO/IEC JTC1/SC29/WG11 M9422, Pattaya, Thailand, March 2003

[17] Lee, W.-S. and Magnenat-Thalmann, N., "Fast Head Modeling for Animation", Journal Image and Vision Computing, Volume 18, Number 4, pp.355-364, Elsevier Sceince, 1 March, 2000.

[18] Lee, W.-S. and Sohn, K., "Multi-view 3D Face Descriptor: proposal for CE", ISO/IEC JTC1/SC29/WG11 M9801, Trondheim, Norway, July, 2003

[19] O'Brien, J. F. and Hodgins, J. K., Graphical Modeling and Animation of Brittle Fracture. The proceedings of ACM SIGGRAPH 99, Los Angeles, California, August 8-13, pp 137-146, 1999.

[20] Pentland, A., Moghaddam, and B., Starner, T., "View-Based and Modular Eigenspaces for Face Recognition", Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'94), 1994.

[21] Proesmans, M. and Van Gool, L., "Reading between the lines - a method for extracting dynamic 3D with texture". In Proc. of VRST'97, pp. 95-102, 1997.

[22] Sim, T., Baker, S. and Bsat, M., "The CMU Pose, Illumination and Expression Database", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003.

[23] Virtual Media, "Face database constriction for research", KISA, 2002. 12

[24] Yamada, A. and Cieplinski, L., "MPEG-7 Visual part of eXperimentation Model Version 17.1", ISO/IEC JTC1/SC29/WG11 M9502, Pattaya, Thailand, March 2003

[25] Zhang, L., Curless, B. and Seitz, S., "Spacetime Stereo: Shape recovery for dynamic scenes", CVPR 2003