3D Face Synthesis with KINECT

Soodamani Ramalingam and Nguyen Trong Viet
School of Engineering and Technology
University of Hertfordshire, AL10 9AB, UK
Email: s.ramalingam@herts.ac.uk, vietnguyen168@gmail.com

***Abstract*—This work describes the process of face synthesis by image morphing from less expensive 3D sensors such as KINECT that are prone to sensor noise. Its main aim is to create a useful face database for future face recognition studies.**

Keywords-3D Face Synthesis, avatars, morphology, disparity, point cloud, KINECT.

I. INTRODUCTION

3D face recognition systems have gained popularity in the last decade. Current security systems such as those used for Border Control however still continue to use 2D face recognition systems. Recent feasibility studies indicate the preference to 3D face recognition (3DFR) systems as well as the availability of large Government datasets makes 3DFR systems very popular.

Recent inventions in camera and imaging systems provide the means to measure and reconstruct 3D faces. Using three dimensional coordinates eliminates the effects of lighting condition and skin colours due to the fact that all the data is in form of coordinates, containing much more information than a flat image. The additional depth dimension allows researchers to analyse and study about the curvatures [1] and estimate samples' pose [2]. 3D capturing devices usually return data as point clouds, based on which a face's mesh is built and aligned to the desired angle [3] for further analyses.

One of the first applications of getting 3D information of a face, also known as face synthesis, is to generate avatars and personalize them so users can have their own distinguished and recognizable characters in online game. Lukasz Zalewski work *et al*. [4] proved that realistic 3D models can exhibit the human's facial expressions, and so are very helpful in transferring emotion between individuals in cyber world. Shopping guidance and robotics GUI also demand this new technique to become more friendly and convenient as stated in [5]. Microsoft Cooperation programmers produce a system where "flat" photographs combined with a base model provides morphable head meshes.

The proposed work is conducted with the main purpose to create a useful 3D face database developed with a commercial system, Microsoft KINECT for future facial recognition studies. Main challenge for any imaging system is processing the raw data from the capturing devices, which can contain loss of information due to the non-ideal conditions of lighting, distance and face angles, a constant problem for researchers to deal with. Other issues like the hair and the participant's clothes' shadow also interfere with face understanding leading to reduced performance of the system. The key stages of the system development focus on the following:

Firstly, to deal with the angle and illumination problem, 3D imaging sensors are considered to get the space coordinates of the subject that is independent of the brightness. At this stage, the sample usually has missing data in the face region resulting from the facial hair on the eyebrows and beard. It is necessary to fill in that absent information without inducing much error as a deviation from the ground truth. For example, if the eyebrow is missing, region filling algorithms must be applied to delete the peaks in depth data and that patch must have a smooth transition with respect to the surrounding pixels. More importantly, these operations must not affect other parts of the face and preserve as much information as possible. This is a prerequisite condition for any further processing towards model construction.

Secondly, since most of the practical imaging devices presently cannot guarantee total accuracy in their image pixels' values, smoothing will be an obvious requirement to clear out the noise caused by hardware. Once again, smoothed image has to maintain the highest similarity with the original. In differential geometry terms, the aim would be to keep the principle curvatures data while getting rid of discontinuous fractures on the sample face.

Thirdly, facial features are identified and marked in depth map due to the lack of information from the colour channels from KINECT which is normally one of the main methods for extracting the region of interest from the RGB image. Algorithms for land marks detection are introduced in an attempt to obtain key feature points on the face.

Finally, a visual interface is implemented for further analysis. This interface allows the process to be fully automatic for face synthesis as well as being able to populate the database in future.

Intelligent algorithms for carrying out these tasks that can be applied automatically for different faces are proposed in this work.

The rest of the paper is organised as follows: In Section II, a review of related work on 3D face recognition is carried out. In Section III, enrollment and key pre-processing steps on depth data from KINECT sensor is detailed out. Section IV outlines the process of morphing faces to enable model construction. Section details the process of image capture with KINECT and database creation. Section V concludes the paper with performance analysis and inferences, and suggestions for further work.

II. REVIEW

In the last decade, the marriage of Image Processing and Computer Graphics techniques have led to successful face synthesis towards meeting face recognition tasks. Microsoft KINECT [12] based 3D data capture for face recognition has recently gained popularity. Typically, such systems use both colour and depth data for locating different fiducial points in 3D. In [6] gender classification of 101 participants constituting 39 males and 45 females in the age group 18-60 is carried out [6]. Variations in poses, illumination, expression and disguise have been attempted in [7] using 4784 samples of 52 subjects. The construction of avatars from average morphable models from depth data is experimented in [8] by energy fitting algorithms in 3D. The average shape from a set of samples lends itself well to building avatar models and tested on a small database of 20 subjects. In [9], a complex 3D transformation between the 3D model feature vertices and the corresponding points on 3D texture map is carried out. A high success rate of 92.3% of samples were synthesized successfully and achieved the Peak Signal to Noise Ratio (PSNR) of 38.7 dB compared with 35.4 dB of a previous study.

A video database captured with BU-3DFE scanner has been used for face synthesis built from texture requiring feature labeling and complex 3D transformation [11]. An Eigen based recognition technique from multiple images has been reported with high recognition rates. The FRGC v2 hybrid dataset combining 2D+3D from a Minolta Vivid camera has been used in [10] for face recognition. Again this technique requires a complex procedure for pose correction, spherical face representation and location of fiducial points before applying ICP. A very high recognition rate of over 99% for neutral poses and 95-98% for non-neutral poses at FAR=0.001 have been reported on 4950 images. Automatic face segmentation on depth maps is carried out in [11] where K-means clustering technique that segments the foreground from the background is adopted. Verification is carried out on FRGC v1 and v2 datasets with 99.9% accuracy of neutral faces and 96.5% overall.

The above review indicates the usefulness of depth maps in face recognition. However, all of these techniques suffer from being computationally intensive. With inexpensive cameras such as the KINECT, a lot of pre-processing is required to smooth sensor errors and deal with low resolution data. Even with the high resolution Minolta Vivid scanners, expensive algorithms are required to perform mapping between 3D texture and depth maps to locate the fiducial points in 3D. In this paper, we propose automatic *quick and dirty* techniques for dealing with sensor noise, determining a set of fiducial points in 3D and deriving morphable models that will be useful for future face recognition. 3D morphable models are constructed with the ability to cope with small changes in pose, expression and illumination. Initial results on model construction are reported.

III. IMAGE PRE-PROCESSING ON DEPTH MAPS

The proposed system consists of the following stages of pre-processing namely i) image capture setting ii) face segmentation from the background iii) extracting region of interest, and iv) noise removal. A KINECT camera is used for image capture. In addition to the depth data, KINECT provides infra-red (IR) data which is quite robust to illumination variation [

Figure 2].

A. Image Capture and Background Extraction

The depth data for a sample image derived in KINECT is as shown in Figure 3. Here the closest point to the camera has an intensity value zero and the furthest point 255 and therefore requires inverting the image. As part of its face tracking algorithm, KINECT has the ability to generate mesh connections superimposed on the RGB image; however these are not extracted as separate data. Hence the first step involved was to generate the mesh data for further stages of pre-processing [Fig.3].

Secondly, the horopter range is too far leading to background scene interference. The foreground object is therefore segmented from the background scene using thresholding, results of which are shown in Fig.4.

B. Image Capture and Region of Interest

The region of interest (ROI) is the face. To extract the face, it is important to maintain this horopter range which is approximately 1m as against the standard range [0.8m 4m] in which KINECT operates. To ensure that the ROI (face) is captured properly, it is required that subjects position their face to lie in a pre-determined area within the image highlighted by a rectangle. This makes the task of cropping the face region easy. Furthermore, in order to avoid remaining still during the image capture, a video sequence is first captured and from which a selection is made.

The ROI typically consists of the face, neck, sometimes the shoulder and the background, each of which lies in specific depth map regions; thus there are 3 main

areas with distinguished values. Calculating mean intensities of these regions helps to decide the correct threshold value for getting binary mask [13].

However, sometimes unwanted parts still exist so a region labeling is carried out. The region with the largest area is the face. This crops the face and isolates redundant regions in the image. To determine such a blob, a face mask is derived from the original image and projected on the depth map. This convolution extracts the region of interest, results of which is shown in [Figs.5-8].

VII. Noise Removal in Depth Maps

Typically there are two types of noise in the depth data namely, holes and spikes. Holes occur as black spots within the ROI and indicate lack of information due to sensor noise. This requires a filling strategy that does not affect the rest of the image. A carefully constructed procedure is adopted as shown in Fig.9. A temporary image fill operation is performed by a *close* morphology operation with a structural element with radius = 5. This results in a blind fill image [10]. It then follows the following steps:

(i) Binary masks are derived for both the raw and morphed images and their difference derived. This results in holes and the boundaries due to the closing operation. The extraneous boundary regions are removed by mapping with the corresponding depth maps. The depth maps typically show very low gray level values at the edges and hence are easily filtered out [Figs.11-12].

(ii) The filled in region of the image is mapped back to the depth map resulting in a touch up of the holes without affecting the rest of the regions [Fig.13].

(iii) Two further operations are required namely image cropping to fit the square just around the image and illumination normalization. The latter is not performed uniformly; instead a contrast stretch is performed at only those points that are closer to the camera [Fig.14].

(iv) Face is a smooth image. However, the low resolution KINECT produces noisy image with discontinuities in the depth maps. Hence a cubic spline smoothing is opted with a parameter of 0.5 through trial and error [Figs.15-16].

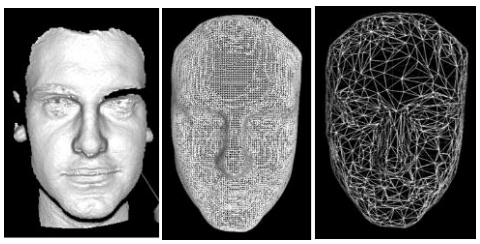


Figure 1 A 3D model, corresponding wireframe and mesh [3]

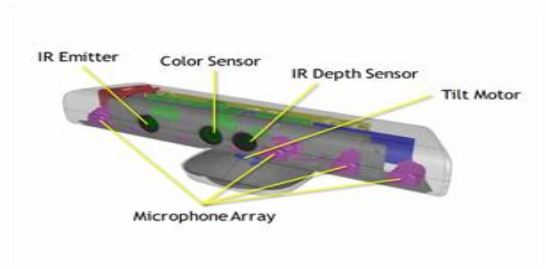


Figure 2. KINECT camera [12]

Figure 3 KINECT Depth Map, Background Removal and Capture Box

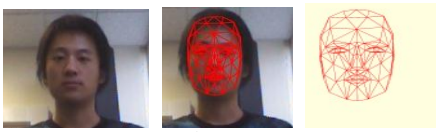


Figure 4 Mesh Image from KINECT

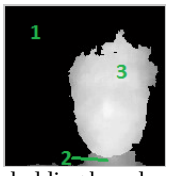


Figure 5 Thresholding based on mean values

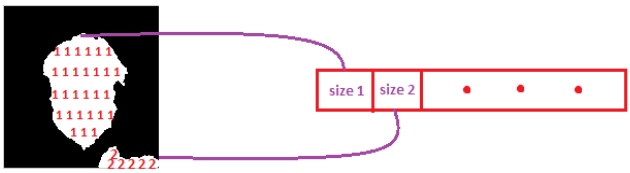


Figure 6 Labeling process

Figure 7 Masked Gray Scale Image and Redundant Parts

Figure 8 Face area after getting largest labeled area

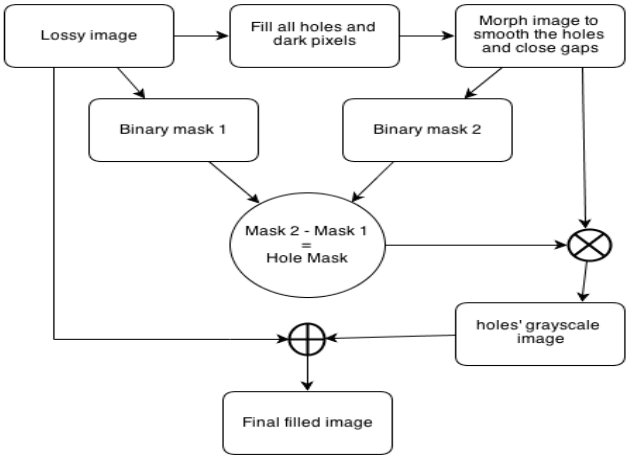


Figure 9 Image Filling Algorithm

Figure 10 Image Fill Operation

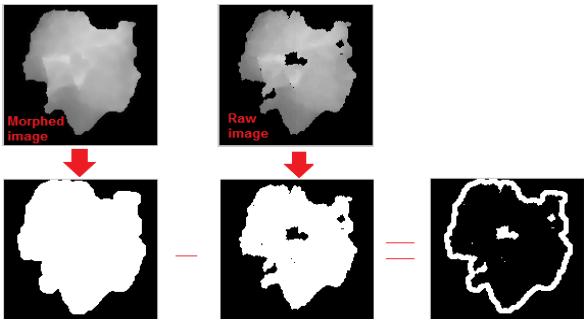


Figure 11 Binary Masks and Their Difference

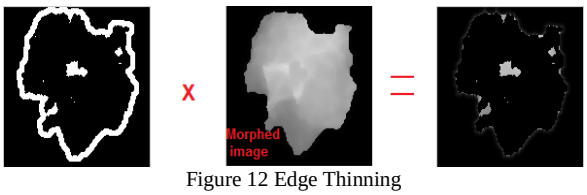


Figure 12 Edge Thinning

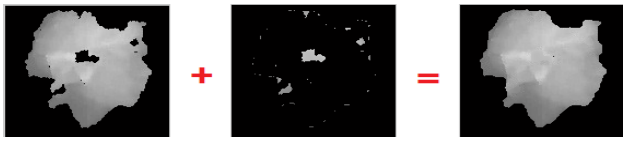


Figure 13 Image Fill - Final Result

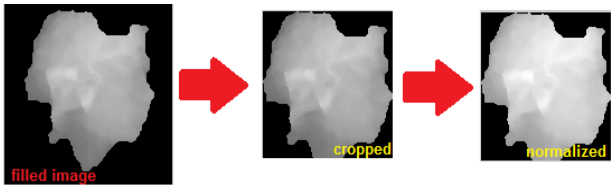


Figure 14 Image Cropping and Illumination Normalisation

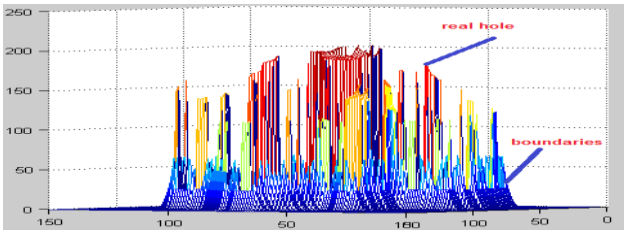


Figure 15 Intensities at the Boundaries Appears <50

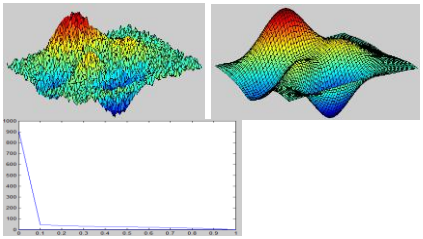


Figure 16 Image Smoothing and RMSE

IV. Face Synthesis by Image Morphing

In this section, we consider the process of image morphing towards building a face model. This requires determining regions around the fiducial points on the face namely temple, eyes, jaw, and the nose.

A. Temple region

With a frontal view, the nose is the closest point to the camera and the brightest in the depth map, which is first determined. The nose ridge is determined by the first gradient change from negative to positive when moving downwards from the forehead. A polynomial fit of the line between the nose tip and the temple appears as shown in [Fig.17].

Gradient change and facial geometry help determine the regions around the eyes and the jaws. These regions also show low intensities in the depth maps. Downward and upward gradients determine the low intensities of the eyes and jaws respectively [Fig.18]. Combining the upper half of the downward gradient and the lower half of the upward gradient determines the eyes mask. A closing morphological operation follows to fill in regions around the eyes.

B. Nose region

In 3D, the nose region can be treated as a protrusion that can be determined by geometric principal curvatures. These curvatures pick up the boundaries surrounding the nose namely, subnasale, upper end of the nose ridge, left and right ala (wings) [Figs.19-20]. A dilation with a disk element size=1, enhances the curvature image.

The nose ridge and temple points have been previously determined shown by red points. The nose wings in green colour are determined by the intersection of a horizontal line passing through the nose ridge and intersecting with the curvatures. The nose region is segmented by a polynomial fit connected to the landmarks.

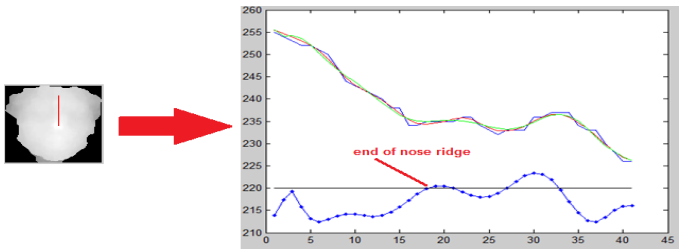


Figure 17 Nose Bridge Profiling

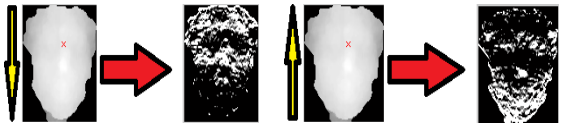


Figure 18 Downward and Upward Gradients

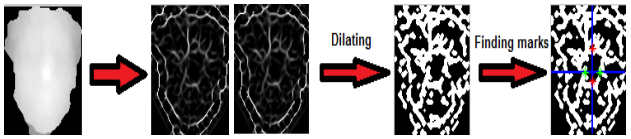


Figure 19 Principal Curvatures and Landmarks

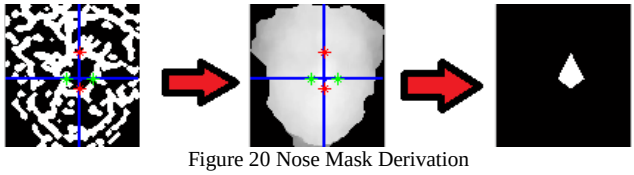


Figure 20 Nose Mask Derivation

The process described in this section allows detection of face in the scene, and touches up various types of noise in the image, extracts just the face region, reconstructs a face model with reference to the depth map. This involves a series of image processing procedures that are carried out automatically.

V. FACE DATABASE CONSTRUCTION

This section describes the process of building a face database and applying face synthesis to the database which then forms the templates.

In order to prove the robustness of the system, video streams from KINECT were captured under varying illumination and face appearance conditions. The only requirement is that the subjects need to sit before the camera and move the head slowly from left to right enabling 2D but not 3D rotation [Fig.21]. Canonical views are generated from this video sequence as shown in Fig.22. An initial database of 14 students has been constructed consisting of mixed race and gender.

A. Observations and Issues

Although the canonical views are captured, face synthesis is carried out only on the frontal view as the nose tip is a key reference point which is assumed to be positioned closest to the camera. The rest of the canonical views require different criteria that are currently under investigation.

Clothes interfere with the texture data that requires further pre-processing. Further, 3D rotation causes less discrimination between the neck and jaw regions in the depth maps.

B. Point Cloud Generation

The results of face synthesis are represented in RGB for visualization purpose. In Fig.23, (a) represents an acceptable case; however (b) has interference due to hair affecting major part of the eye region. This generated several minor valleys which get further amplified due to the closing operations. In case of (c), the mouth is open; however this has been an acceptable case. The rest of the images show reasonably acceptable results. The corresponding 3D point cloud generated for one of the images is shown in Fig.24. This is then represented as a mesh plot in 3D as shown in Fig.25.

Figure 21 Participant's Head Movement **[14]**

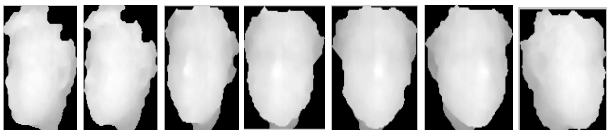
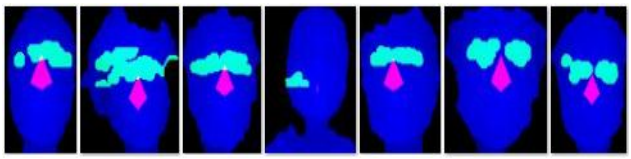
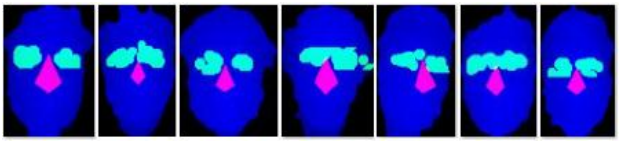


Figure 22 Canonical Views Generated



(a) (b) (c) (d) (e) (f) (g)



(h) (i) (j) (k) (l) (m) (n)

Figure 23 Face Models

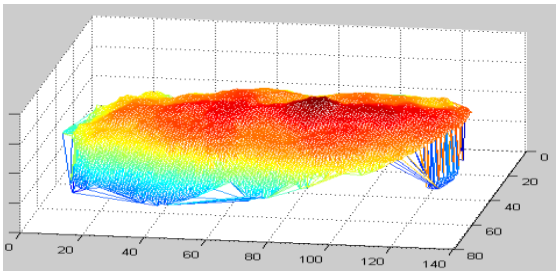


Figure **24** Point Cloud Generation

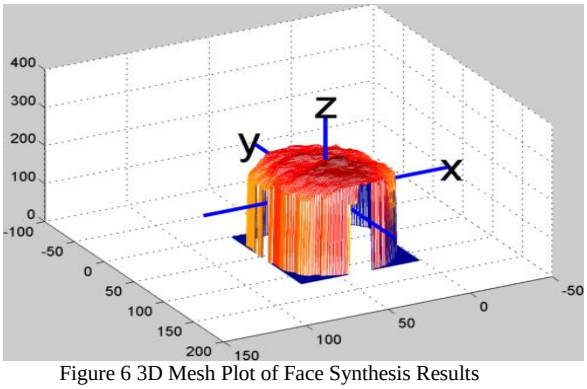


Figure 6 3D Mesh Plot of Face Synthesis Results

VI. Conclusion and Future Recommendations

The main aim of this project is to set up and initialize a system for useful 3D image data capture from a low cost camera such as the KINECT. In addition, effective pre-processing procedures are defined to fill in holes, segment the face from the background, and detect landmark regions to build a 3D face model. With this pre-processing in place, a point cloud is generated and represented as a 3D mesh plot.

Further work involves the following:

- Coping with poses – presently model construction is achieved only for the frontal views and this needs to be extended for other views.
- Fusion of RGB and depth maps to manage hair and clothes interference in the image.
- Removal of spikes and smoothing the point clouds generated.
- True avatar generation for working in Computer Graphics domain.
- Extending the database to a larger set and performing face recognition.

VIII. References

[1] Nilesh U. Powar, Jacob D. Foytik, Himanshu Vajaria and Vijayan K. Asari, "Facial Expression Analysis using 2D and 3D Features," in *Proceedings of the 2011 IEEE National Aerospace and Electronics Conference (NAECON)* , Fairborn, Ohio, 2011.

[2] Yun Sheng, Abdul H. Sadka and Ahmet M. Kondoz, "Automatic Single View-Based 3-D Face Synthesis for Unsupervised Multimedia Applications," *IEEE Transactions on circuits and system for video technology*, vol. 18, no. 7, pp. 961 - 974, 2008.

[3] V. Bevilacqua, F. Andriani and G. Mastronardi, "3D Head Normalization with face geometry analysis, genetic algorithms and PCA," *Journal of Circuits, Systems, and Computers*, vol. 18, no. 8, p. 1425–1439, 2009.

[4] Lukasz Zalewski and Shaogang Gong, "2D statistical models of facial expressions for realistic 3D avatar animation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 2005.

[5] Hyo-seok Kang, Sung-hoon Yu and Mignon Park, "3D Matching applied ICP Algorithm for 3D Facial Avatar Modeling Using Stereo Camera," in *Aging Friendly Technology for Health and Independence, 8th International Conference on Smart Homes and Health Telematics*, Seoul, 2010.

[6] Y. v. d. Hurk, "Gender classification with visual and depth images," Tilburg University, 2012.

[7] Li, B.Y.L. ; Mian, A.S. ; Wanquan Liu ; Krishna, A., "Using Kinect for face recognition under varying poses, expressions, illumination and disguise," in *IEEE Workshop on Applications of Computer Vision (WACV), 2013*, 2013.

[8] Michael Zollhöfer, Michael Martinek, Günther Greiner, Marc Stamminger, Jochen Süßmuth, "Automatic reconstruction of personalized avatars from 3D face scans," *Journal of Visualization and Computer Animation*, vol. 22, pp. 195-202, 2011.

[9] Y. Sheng, A.H. Sadka and A.M. Kondoz, "Automatic 3D face synthesis using single 2D video frame," *Electronics Letters* , vol. 40, no. 19, pp. 1173 - 1175, 2004.

[10] Ajmal S. Mian, Mohammed Bennamoun, and Robyn Owens, "An effecient multimodal 2D-3D hybrid approach to automatic face recognition," in *IEEE Transactions on pattern analysis and machine intelligence*, 2007.

[11] Maurício Pamplona Segundo, Luciano Silva, Olga Regina Pereira Bellon and Chauã C. Queirolo, "Automatic Face Segmentation and Facial Landmark Detection in Range Images," in *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2010.

[12] "Kinect for Windows," [Online]. Available: http://www.microsoft.com/en-us/kinectforwindows/develop/tutorials.aspx. [Accessed 19 April 2013].

[13] J. V. Manjon-Herrera, 28 August 2005. [Online]. Available: http://www.mathworks.co.uk/matlabcentral/fileexchange/8379-kmeans-image-segmentation. [Accessed 18 April 2013].

[14] "HE ergonomic brand," 06 07 2006. [Online]. Available: http://www.hemouse.com/health/487.html. [Accessed 25 April 2013].

[15] Simon J.D.Prince, James Elder, Y.Hou, M.Sizintev and E.Olevskiy, "Toward face recognition at a distance," in *The Institution of Engineering and Technology Conference on Crime and Security*, London, 2006.