Informal Technical Report
Current Status of Avatar Project and Potential Improvement Directions

Jingwen Dai, Greg Welch and Henry Fuchs
Department of Computer Science
University of North Carolina *at* Chapel Hill

14 December 2012

Abstract

This report summarizes the current status of mobile avatar project, mainly focusing on the disadvantages of current method and implementation. And then based on these disadvantages, several potential improvement directions are discussed, along with the related works being reviewed.

1 Current Statues

The mobile avatar tele-presence system can be divided into two parts, the *inhabiter station* and *mobile avatar*.

1.1 Inhabiter Station

In the inhabiter station, the user's head pose (especially the orientation) should be tracked continuously to control the avatar head on the mobile avatar side. Currently, the user has to wear a helmet, on which there are optical trackers for acquiring the position and orientation of the inhabiter and a camera to capture the frontal face imagery, as shown in Fig. 1.

For the user, the requirement of wearing a helmet or a hat is not very natural and convenient. And the user's 3D face model was generated by 3rd party software FaceWorx through moving control points in photographs showing the frontal and profile face [1, 2]. This procedure requires manual identification of distinctive facial features, so it is very time-consuming to model a user's face.

Using vision-related method (single-, multiple- camera or Kinect) to estimate the head pose and get the 3D face model for the avatar side projection is one of the approaches to solve these problems.

1.2 Avatar

Comparing with the previous avatar, the current avatar adopts rear projection instead of front projection, the projector is fixed rigidly with the face-shaped projection surface.

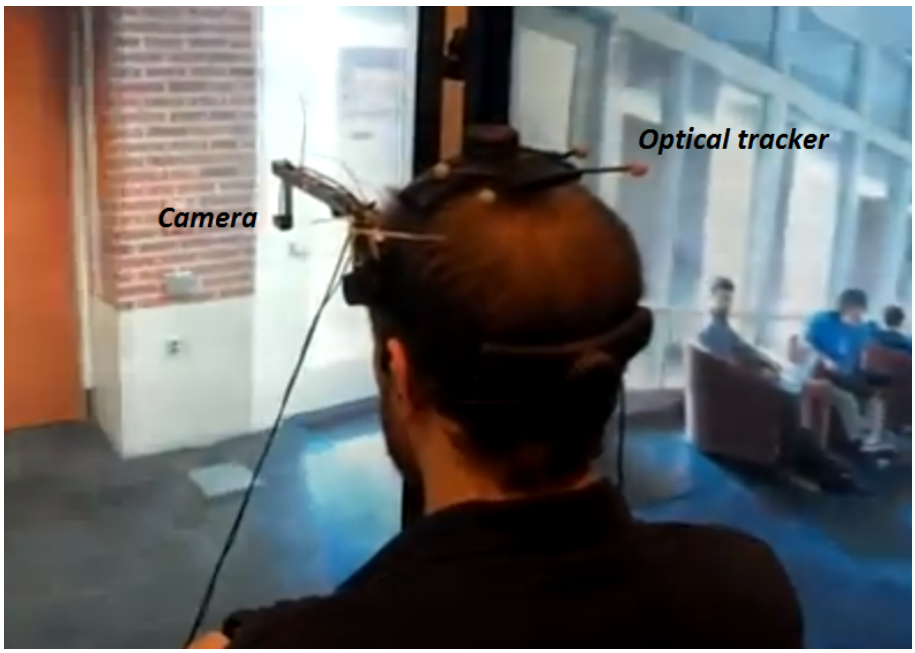

Figure 1: User in inhabiter station wearing a helmet with optical tackers and camera.

Currently, the alignment of projection image and the face-shaped surface is accomplished manually (the size, position and rotation of the projection image), this setup is not user-friendly. And there are still some misalignment between the projected image and the face-shaped surface, making the appearance of the avatar distorted, as shown in Fig. 2(a) and (b). When the inhabiter speaks or has some expression changes, this misalignment becomes more distinct. Moveover, due to the inter-reflection and specular reflection, the appearance of projected avatar face is not homogeneous, in Fig. 2(b) some errors are shown (especially the sparkling spot in the eye region).

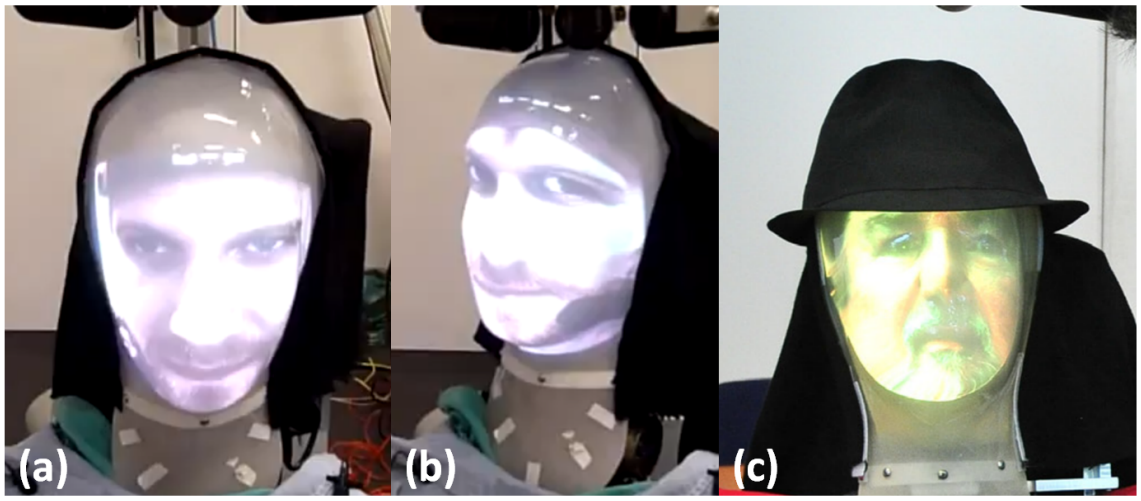

Figure 2: Some imperfections of projected avatar face: (a) & (b) misalignment, (c)inhomogeneous appearance due to inter-reflection or specular reflection.

So the potential improvement on the avatar side may include

- Automatic alignment of projected face image and face-shaped surface

- Compensation of the error caused by inter-reflection and specular reflection to get more realistic face projection, even adding the feedback of a camera.
- Automatic adjustment for the misalignment caused by expression variations of the inhabiter

The rest of this report will discuss some potential research directions with the review of related works.

2 Potential Improvement Directions

2.1 Vision-based Head Pose Estimation

2.1.1 Related Works

- **2D methods**

  1. *Single View*

     Head pose estimation using 2D images captured by single camera was studies extensively in computer vision society. A recent literature review [3] summarized many typical approaches on this topic. These approaches included appearance template methods, detector array methods, nonlinear regression methods, manifold embedding methods, methods based on flexible models etc. In general, these approaches had some limitations: some of them could not estimate the head pose continuously, and some of them could not handle large pose variations, since they mainly depend on the detection of facial features points. So many researcher try to find solutions in multiple view case.

  2. *Multiple Views*

     In [4], a head-pose tracker was presented by rigidly mounting multiple cameras on the subject's head, mapping head pose estimation problem to that of ego-motion estimation. In [5], a head pose tracking approach was proposed using multiple calibrated (both internal and external parameters) cameras. This approach relied on tracking face landmarks at each camera and triangulating their features, having limitations on robustness due to the modeling of the face as a plane. In [6], a robust approach to real-time 3D head pose tracking using multiple cameras with unknown camera placements was proposed, in which a generic face model was employed to overcome the difficulties due to the lack of prior knowledge of camera placement. The feature points that this method tracked was Harris feature, which was more robust to the pose variation and occlusion.

- **3D methods**

  Estimating head pose, i.e. orientation and translation in 3D, from 2D images is intrinsically problematic. In general, the 2D approaches are sensitive to the variations of 2D face image, due to the change of ambient illumination, facial expression and facial self occlusion. Since 3D sensing devices have become available, especially the recent introduction of Microsoft Kinect in consumer market, making it is very convenient to get 3D information in very low price. In computer vision society,

many researchers have started to leverage the additional depth information for solving some of the inherent limitations of image-based methods. Some of the recent works use depth as primary cue [7, 8, 9] or in addition to 2D images [10, 11, 12].

In [10], a neural network-based system that fusing skin color histograms and depth information was presented. It could track at 10 fps but required the face to be detected in a frontal pose in the first frame. In [11], intensity and depth image were employed to build a prior model of the face using 3D view-based eigenspaces to compute the absolute difference in pose for each new frame. However, the pose range is limited and manual cropping is necessary. In [12], a regularized maximum likelihood deformable model fitting (DMF) algorithm was developed to align a 3D face model to an RGB-D input stream for tracking features across frames, with special emphasis on handling the noisy input depth data.

In [7], the nose position was estimated in range image by directional maxima. In [8], a shape signature was proposed to identify noses in range images, then candidates for their positions were generated, and then many pose hypotheses were generated and evaluated in parallel using GPU. This system could handle large pose variations, facial expressions and partial occlusions, as long as the nose remains visible, but the real-time performance is only guaranteed by parallel GPU computations.

Fanelli et al. [9] proposed a random forest-based framework for real time head pose estimation from high resolution depth map [13] and low resolution depth map [14] and extend it to localize a set of facial features [15] in 3D. The algorithm takes a voting approach, where each patch extracted from the depth image can directly cast a vote for the head pose. Moreover, the algorithm could work on each frame independently and achieve real time performance without GPU. The author provided the source code for head pose estimation by Kinect data [1]. And the mean error and standard deviation are around 4° and 6°, which is a little greater than that of the approaches using 2D images.

2.1.2 Potential Research Direction

2D image based head/face pose estimation has a long history more than 20 years, researchers tried to solve the problem from different perspectives, and the reported performances were good under certain constraints (such as all facial landmarks were visible, no occlusions or the head pose was controlled in a special range etc.). Hence, the space in this area is relatively small.

The emergence of low-cost 3D sensors (Kinect or ToF camera) makes 3D capture simply and convenient, so one potential research direction is to using the data from Kinect for real-time head pose estimation, the work in [9] is an good example, which would run in real-time without GPU acceleration. The estimated head pose could not only be used in avatar head control but also be used in face modeling through multiple point clouds alignment and registration.

[1] `http://www.vision.ee.ethz.ch/~gfanelli/head_pose/head_forest.html`

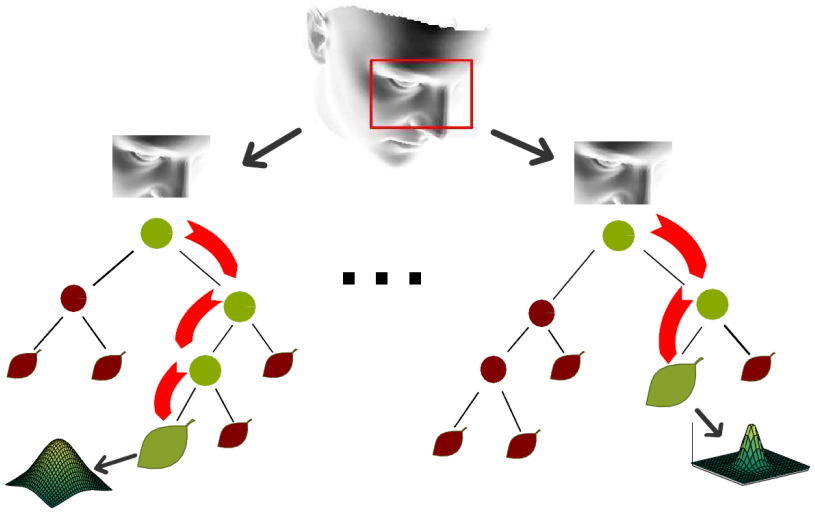

Figure 3: Example of regression forest for head pose estimation. For each tree, the tests at the non-leaf nodes direct an input sample towards a leaf, where a real-valued, multivariate distribution of the output parameters is stored. The forest combines the results of all leaves to produce a probabilistic prediction in the real-valued output space [9].

2.2 3D Face Modeling

2.2.1 Related Works

Image-based 3D face modeling has been widely studied in literature. In general, they employed single image, multiple images (frontal face image and profile face image, or face image sequence, or stereo image pair), to fit a general model or to reconstruct the face directly by shape-from-X(motion, shading, etc.). With the appearance of low-cost 3D sensor (ToF camera and Kinect) in consumer market, many researcher start thinking about how to use the depth information to model human face or other objects.

In [16], a method for 3D object scanning by aligning depth scans that were taken from around an object with a time-of-flight camera was described. The challenge was that the sensor's random noise is substantial and there was a no-trivial systematic bias in ToF camera. To solve the problem, this method was based on a combination of a 3D super-resolution method with a probabilistic scan alignment approach that explicitly took into account the sensor's noise characteristics. Some results of this method are shown in Fig. 4. This method has potential to be extended to Kinect data.

Figure 4: Left: raw scan. Middle: proposed method. Right: laser scan as ground-truth [16].

Weise et al. [17] proposed a system for performance-based character animation that enabled any user to control the facial expression of a digital avatar in realtime. The user was recorded in a natural environment using Kinect. To map low-quality 2D images and 3D depth maps to realistic facial expressions, they introduced a face tracking method combining geometry and texture registration with pre-recorded animation priors in a single optimization. This paper just mapped expression to virtual avatar other than the personalized 3D face model.

In [18], a algorithm for computing a high-quality personalized avatar from a single color image and the corresponding depth map captured by Kinect was proposed, which combined the advantage of robust non-rigid registration and fitting a morphable model. Some reconstruction results are demonstrated in Fig. 5. Due to the iterative optimization of morphable model fitting, the reconstruction of facial avatar took average 18 seconds on a high performance desktop.

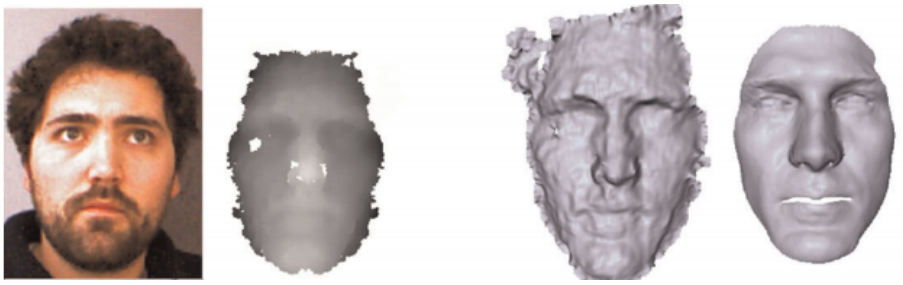

Figure 5: Face Reconstruction results of [18]: (from left to right) input RGB and depth image, processed depth scan, fitted face model.

Hernandez et al. [19] proposed a method to produce laser scan quality 3-D face models from a freely moving user using Kinect. This method did not rely on any prior face model and could produce faithful geometric models of star-shaped objects. The object was presented in cylindrical coordinates, enabling filtering operations to perform very efficiently. They initialized the model with the first depth image, and then registered each subsequent cloud of 3-D points to the reference using a GPU implementation of the ICP algorithm. Both temporal and spatial smoothing of the successively incremented model were performed. Some results are shown in Fig. 6. However some regions in reconstruction face were over filtered, especially the regions around eyes and mouth. Even using GPU acceleration, it needed about 10 seconds to get a complete face model.

Figure 6: Noise removal using filtering [19]: (from left to right) accumulated raw input mean filtering only, bilateral filtering only, Both filters.

In [20], KinectFusion[2] system took live depth data from a moving Kinect camera and created a high quality 3-D model for a static scene object. Aligning all point clouds with the complete scene model from large environment provided very accurate tracking of the camera pose and mapping. In [20], the author presented a result of human face modeling, as shown in Fig. 7. However, this approach is only suitable to rigid and static object, it could not handle expression variation during capture in face modeling.

Figure 7: Left: noisy raw data from a single frame. Middle: normal map. Right: Phong-shaded renderings [20].

2.2.2 Potential Research Direction

Getting an accurate 3-D face model from the depth sensor is a challenging problem. In general, the quality of single frame is not sufficient to generate reasonable 3-D face models. First, the resolution of the depth map captured from Kinect is low, i.e. 640×480 pixels with 11-bit depth. And due to the constraint of Kinects working distance, the size of the face in depth map is smaller than 300×300. Second, depth data near boundaries can be very noisy and simple averaging on time is not sufficient. One idea to compensate for the noisy depth data is to use several poses, accumulated and refine noisy information through time. The objective is to find a coarse-to-fine approach to model a morphable and animatable 3D face in real-time. The general model of avatar head may be employed as prior knowledge to simplify the problem.

2.3 Projection Correction

2.3.1 Related Works

Need more time to review this part of work in literature.

2.3.2 Potential Research Direction

- Automatic alignment of projected face image and face-shaped surface
- Compensation of the error caused by inter-reflection and specular reflection to get more realistic face projection, even adding the feedback of a camera.

[2] An open source implementation of KinectFusion would be find in PCL (`http://pointclouds.org/news/kinectfusion-open-source.html`) and Microsoft is planning to integrate KinectFusion module into Kinect SDK.

- Automatic adjustment for the misalignment caused by expression variations of the inhabiter

3 The Work of Next Week

- Find some more papers on projection compensation, especially for the surface with specular reflection and inter-reflection.
- Try the open source code of KinectFusion to test its performance on human face modeling.

References

[1] Peter Lincoln, Greg Welch, Andrew Nashel, Andrei State, Adrian Ilie, and Henry Fuchs. Animatronic shader lamps avatars. *Virtual Reality*, 15(2-3):225–238, 2011.

[2] Peter Lincoln, Greg Welch, and Henry Fuchs. Continual surface-based multi-projector blending for moving objects. In *Proceedings of IEEE Virtual Reality*, pages 115–118, 2011.

[3] E. Murphy-Chutorian and M.M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[4] Sarah Tariq and Frank Dellaert. A multi-camera 6-dof pose tracker. In *Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 296–297, 2004.

[5] R. Ruddarraju and I. A. Essa. Fast multiple camera head pose tracking. In *Proceedings of Vision Interface*, 2003.

[6] Qin Cai, A. Sankaranarayanan, Q. Zhang, Zhengyou Zhang, and Zicheng Liu. Real time head pose tracking from multiple cameras with a generic model. In *Proceeding of IEEE Workshop on Analysis and Modeling of Faces and Gestures*, 2010.

[7] Xiaoguang Lu and A.K. Jain. Automatic feature extraction for multiview 3d face recognition. In *Proceedings of 7th International Conference on Automatic Face and Gesture Recognition*, pages 585–590, 2006.

[8] M.D. Breitenstein, D. Kuettel, T. Weise, L. van Gool, and H. Pfister. Real-time face pose estimation from single range images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[9] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, pages 1–22, 2012.

[10] Edgar Seemann, Kai Nickel, and Rainer Stiefelhagen. Head pose estimation using stereo vision for human-robot interaction. In *Proceedings of the Sixth IEEE international conference on Automatic face and gesture recognition*, pages 626–631, 2004.

[11] Louis-Philippe Morency, Patrik Sundberg, and Trevor Darrell. Pose estimation using 3d view-based eigenspaces. In *Proceedings of IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 45–52, 2003.

[12] Qin Cai, David Gallup, Cha Zhang, and Zhengyou Zhang. 3d deformable face tracking with a commodity depth camera. In *Proceedings of the 11th European conference on computer vision*, pages 229–242, 2010.

[13] G. Fanelli, J. Gall, and L. Van Gool. Real time head pose estimation with random regression forests. In *Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 617–624, 2011.

[14] Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc Van Gool. Real time head pose estimation from consumer depth cameras. In *Proceedings of the 33rd international conference on Pattern recognition*, pages 101–110, 2011.

[15] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2578–2585, 2012.

[16] Yan Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt. 3d shape scanning with a time-of-flight camera. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1173–1180, 2010.

[17] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Realtime performance-based facial animation. In *ACM SIGGRAPH*, pages 77:1–77:10, 2011.

[18] Michael Zollhofer, Michael Martinek, Gnther Greiner, Marc Stamminger, and Jochen Submuth. Automatic reconstruction of personalized avatars from 3d face scans. *Computer Animation and Virtual Worlds*, 22(2-3):195–202.

[19] M. Hernandez, Jongmoo Choi, and G. Medioni. Laser scan quality 3-d face modeling using a low-cost depth camera. In *Proceedings of the 20th Europea Signal Processing Conference*, pages 1995–1999, 2012.

[20] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011.