Merging Live and pre-Captured Data to support Full 3D Head Reconstruction for Telepresence

Cédric Fleury[1,2,5] Tiberiu Popa[3,6] Tat Jen Cham[1,4] Henry Fuchs[1,2]

[1] BeingThere Centre
[2] University of North Carolina at Chapel Hill, USA
[3] ETH Zurich, Switzerland
[4] School of Computer Engineering, Nanyang Technological University, Singapore
[5] Université Paris-Sud & CNRS (LRI), Inria, France
[6] Concordia University, Canada

Abstract

This paper proposes a 3D head reconstruction method for low cost 3D telepresence systems that uses only a single consumer level hybrid sensor (color+depth) located in front of the users. Our method fuses the real-time, noisy and incomplete output of a hybrid sensor with a set of static, high-resolution textured models acquired in a calibration phase. A complete and fully textured 3D model of the users' head can thus be reconstructed in real-time, accurately preserving the facial expression of the user. The main features of our method are a mesh interpolation and a fusion of a static and a dynamic textures to combine respectively a better resolution and the dynamic features of the face.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: 3D Graphics and Realism—H.4.3 [Info. Syst. Appli.]: Communications Appli.—Computer conferencing, teleconf., and videoconf.

1. Introduction

Compared to standard 2D videoconferencing systems, 3D telepresence systems can improve social presence between remote users (feeling of "being together") [MF11]. The recent development of inexpensive hybrid sensors (color+depth) such as the Microsoft Kinect as well as consumer level 3D displays have the potential of bringing 3D telepresence systems to the main stream. Nevertheless, several important challenges remain. Simple set-ups consisting of one acquisition device and one screen are mandatory to make 3D telepresence suitable for the general public. However, this kind of set-up cannot, due to occlusions, reconstruct a complete model of the head. Moreover, the data captured by the commercial depth sensors are noisy and low quality. These data need to be improved to enable remote users to accurately perceive facial expressions.

In this paper, we propose a method for real-time 3D head reconstruction to be used in one to one 3D telepresence system suitable for the general public. This method uses a single hybrid sensor (color+depth) located in front of each user. Our method fuses the dynamic, but incomplete, low resolution and noisy real-time output from the hybrid sensor with a set of textured, complete and high-resolution, but static models of the user's head. The result is a complete and fully textured 3D model of the user's head animated in real-time that accurately preserves the current facial expression of the user. The static textured meshes are acquired in a simple preprocessing stage that takes only a few minutes. The user acquires a small set (around 8) of textured models of his/her head using different facial expressions by simply turning 360 degrees in front of the acquisition device. The reconstruction step uses a real-time face tracker [CGZZ10] to select and interpolate between the high-resolution models. The static textures acquired in the pre-processing stage is complemented by the texture from the live color camera in order to deal with dynamic face features (eyes, mouth, wrinkles, etc.).

This paper is composed of five sections. Section 2 presents the related work. Section 3 describes the method. Section 4 gives some results and discusses the method. Section 5 concludes and proposes some future works.

2. Related Work

3D head reconstruction is a wide area and the goal of this paper is not to do a exhaustive review of all existing techniques. We refer to [PL06] for a more detailed state-of-the-art.

Some existing techniques achieve very accurate 3D models of the head using high resolution cameras. These techniques use a set of 2D images taken from different points of view and do stereo matching. [BBB*10] has achieved impressive 3D models with high resolution details as skin pores. This technique has been extended to animated 3D head models in [BHB*11]. Structured lights can also be used. However, all these techniques require an expensive (high resolution cameras) and complex (calibration, space-consuming) material set-up. This kind of set-up is not suit-

DOI: 10.2312/egsh.20141002

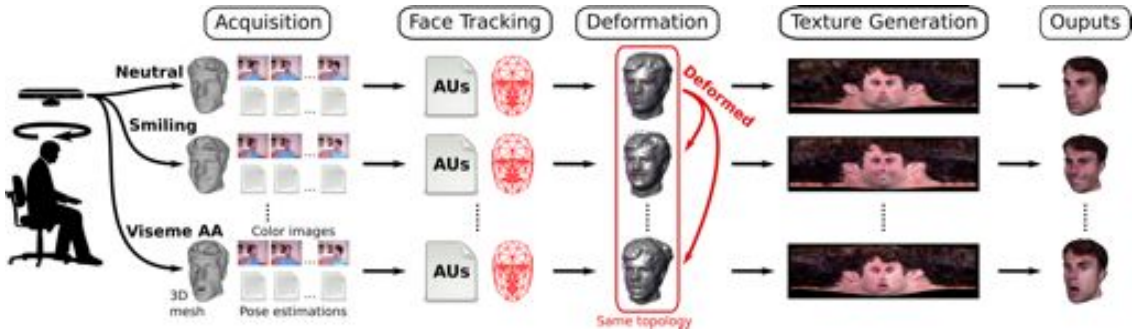

Figure 1: *Acquisition step: data are processed to create a complete and fully textured 3D head model for each facial expression.*

able for general public use. Moreover, these techniques do not perform in real-time and require up to several seconds to process one frame which is not suitable for telepresence.

Some other techniques achieve real-time reconstruction by fitting a deformable face model to the depth data of the face captured by a depth sensor. This deformable face model is usually created from a large database of human face scans. [WBLP11] and [LYYB13] use this solution to do real-time facial animation by re-targeting the facial expression of the deformable face model to a virtual character. Even if these techniques achieve really nice results for facial animation, it is not suitable for telepresence at this state. Indeed, deformable face models do not fit the general appearance of the user's head because hairs, eyes, interior of the mouth are missing. A texture of the user's face can be generated in [WBLP11]. However, this texture is static and follows the deformation of the mesh, so some inconsistencies can appear for the small face features (eyes, teeth, tongue, wrinkles, etc.). For 3D telepresence, remote users might look like video game characters with these techniques. This can destroy social presence because of the "Uncanny Valley".

Our goal is to achieve 3D reconstruction of users' head in 3D telepresence systems affordable for the general public. It involves using a cheap and simple material set-up as a single camera with no calibration. The proposed technique must reconstruct a 3D model as close as possible of the real head with all the face features (hairs, eyes, ears, interior of mouth, etc.) according to the quality of data from the material set-ups. The texture of the head model has to show all the small face features even if the 3D mesh is not precise enough to model these small features. So we think that a dynamic texture (linked to the video stream) is better for telepresence.

3. Full 3D Head Reconstruction

Our method proposes to use pre-captured data (3D meshes and textures) to improve the live output of an hybrid sensor to reconstruct in real-time a complete and fully textured 3D model of the users' head. Our method can be decomposed in two separated steps: an acquisition step where data are captured and pre-processed (3.1) and, then, a reconstruction step where the head model is reconstructed in real-time (3.2).

3.1. Acquisition Step

To capture data for several facial expressions, each user spins on a chair in front of a hybrid sensor with different facial expressions. The user must keep the same facial expression during one entire turn. We ask him to do several facial expressions (neutral, opened mouth, smiling, eye brown up and down) and visemes (/aa/, /p/, /t/, /uh/). For each facial expression (i.e. for each turn), data are accumulated using the KinectFusion [IKH*11] algorithm. Outputs are a 3D mesh of the head, color images taken all around the head, and pose estimations of the camera for each color image (see Fig. 1). If we integrate the KinectFusion algorithm into an automatic process, this step would take little time (only 2-3 minutes) despite having to capture several facial expressions.

3.1.1. Face Tracking and Face Feature Detection

The next step is to track the head and to characterize the facial expression in each set of data (see Fig. 1). We use the face tracker from the Microsoft Kinect SDK [CGZZ10] which gives as outputs the head pose estimation, a set of values (AUs) which characterize the face features and a 3D coarse mesh which fits the face. AUs are deltas from a neutral shape to the current shape for several features (lip, jaw, eye brown positions, etc.). For each data set, we store AUs to describe the corresponding facial expression.

3.1.2. 3D Mesh Deformation

After data acquisition, we have a 3D mesh for each facial expression. These meshes will be used to reconstruct a 3D head model which fits the user's current facial expression. We do not have the exact mesh that fits his facial expression, so we will use a weighted interpolation of several meshes (more details in 3.2.2). To do this interpolation, all the meshes need to have the same topology (which is not the case). So the mesh of the neutral face is deformed to fit the other meshes one by one (see Fig. 1). Each time, the deformed mesh is saved. A set of deformed meshes with the same topology can thus be created (one for each facial expression).

For the deformation, the method proposed in [KS04] is used. This method is based on a cross-parametrization of both the source and the target meshes. It requires to have

Figure 2: *Cylindrical texture: (a) without alignment, (b) with alignment, and (c) with the Poisson interpolation.*

some common features to establish correspondences between the two meshes, and a coarse mesh linking these features together. Each triangle of the coarse mesh is used to parametrized an particular area of the two meshes. This method is particularly suitable for our system because we have a coarse mesh from the face tracker for each facial expression. The features of these coarse meshes correspond to some face features (nose tip, eye and mouth corners, etc.).

3.1.3. Cylindrical Texture Generation

For the texture, a cylindrical texture is generated by combining the color images of each data set using a cylindrical projection. For a particular facial expression, the set of color images is projected on the 3D mesh using the camera pose estimations. Then, each pixel is projected back on a cylinder around the user's head. If several images contribute to the same pixel of the cylindrical texture, the dot products of each camera direction (when the image was taken) and the normal of the mesh triangle which contains the pixel are computed and the image with the biggest dot product is chosen.

Camera pose estimations are not accurate enough and some misalignments appear in the cylindrical texture. We propose to align the camera positions two by two, by minimizing the Euclidean distance (in the RGB space) between the common contributions of two successive images. Moreover, color balance and lighting between the color images are not exactly the same and some seam lines appear in the cylindrical texture. The Poisson interpolation proposed in [PGB03] is used to remove these seam lines (see Fig. 2).

3.2. Real-time Reconstruction Step

The acquisition step (3.1) generates a 3D mesh (with the same topology than the other meshes), a cylindrical texture and AUs descriptors for each facial expression. This set of head models (mesh/texture/AUs) are used to improve the reconstruction of the user's head during the 3D telepresence session. In this step, each user is sitting in front of a hybrid sensor as if he is doing videoconferencing with a webcam.

3.2.1. Head Model Selection

We need first to select the head models used for the reconstruction. The same face tracker than for the acquisition (see 3.1.1) is run on the current output of the hybrid sensor to determine the head pose and the face features (let AU_{C_i} be the i-th AUs detected in the output). The AUs have also been stored for each head model (let M be the set of head models and AU_{m_i} be the i-th AUs of the model $m \in M$). The current face features are compared with the ones of each head model $m \in M$ by computing the Euclidean distance as follows:

$$Dist(m) = \sum_{i=1}^{n} AU_{C_i} - AU_{m_i} \quad (1)$$

where n is the number of AUs ($n = 6$ with the face tracker from the Microsoft Kinect SDK).

3.2.2. 3D Mesh Reconstruction

To reconstruct a 3D mesh which fit the current facial expression of the user, we choose a subset of head models $X \subset M$ with the lowest Euclidean distance $Dist(x)$ for $x \in X$ (i.e. which are the closest to the current facial expression). These meshes are interpolated to create a new mesh a. The 3D coordinate $Coord_a$ of the vertice v in the mesh a is the weighted interpolation of $Coord_x$ of v in each mesh $x \in X$ (all the meshes have the same topology, so each v in a mesh has a equivalent in the others). Weights are computed as follows:

$$Coord_a(v) = \sum_{x}^{X} (1 - \frac{Dist(x)}{\sum_t^X Dist(t)}) \, Coord_x(v) \quad (2)$$

The new mesh a is added into the depth data from the hybrid sensor instead of the depth data of the head using the current head pose estimation from the face tracker (see Fig. 3(b)).

3.2.3. Texture Creation

The texture tex_a of the mesh a is created in two steps. First, a static texture $sTex_a$ is created by interpolating for each pixel p the cylindrical textures $cylTex_x$ of each model $x \in X$ with some weights similar to Equation (2):

$$sTex_a(p) = \sum_{x}^{X} (1 - \frac{Dist(x)}{\sum_t^X Dist(t)}) \, cylTex_x(p) \quad (3)$$

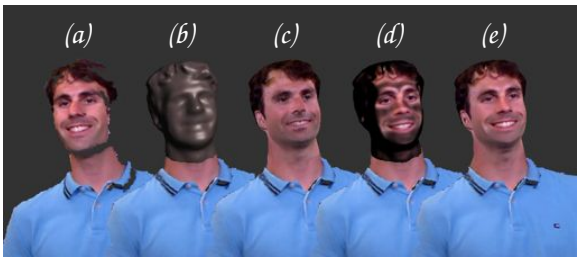

Figure 3: *(a) 3D reconstruction from sensor output, (b) interpolated 3D mesh from the pre-captured data textured with (c) the static texture, (d) the dynamic texture and (e) both.*

The static texture allows to have a 360° texture of the head with a better resolution than in the sensor color images (see Fig. 3(c)). However, the dynamic face features as the eyes, mouth, wrinkles, etc. are not consistent with the current user's face. So we propose to use the static texture for the background, but to use a dynamic texture from the current video stream of the hybrid sensor for the strong face features. Getting only the strong features from the dynamic texture and only the background from the static texture avoids to have visible seams when merging the two textures together. The gradient $\nabla_{cci}(p)$ of the current color images (cci) from the sensor is used to extract the strong face features (see Fig. 3(d)). The final texture tex_a is generated by merging together the static texture $sTex_a$ and the dynamical texture (current color image) $dTex_{cci}$ as follows (see Fig. 3(e)):

$$tex_a(p) = \alpha(p)\, dTex_{cci}(p) + (1-\alpha(p))\, sTex_a(p) \quad (4)$$

with $\alpha(p) = dot_{cam}(p) \,.\, \nabla_{cci}(p)$ where $dot_{cam}(p)$ is the dot product between the hybrid sensor current direction and the normal of the triangle containing the pixel p in the mesh a. In this way, the dynamic texture contributes only to the part of the mesh which is in front of the hybrid sensor.

4. Preliminary Results and Discussion

8 models with different facial expressions have been captured on European and Asian users ($\|M\| = 8$) and the 2 closest have been selected for reconstruction ($\|X\| = 2$). Complete and fully textured head models have been reconstructed in real-time while the output of the sensor gives only the front part of the face with holes (see Fig. 4). Our method can be used for 3D telepresence on standard network because only color images and depth maps need to be sent. The reconstruction can be done remotely to enable users to see each other in 3D. Our method can also be used to modify the viewpoint to conserve eye contact as proposed in [KPB*12].

Although in our experiments the face tracker provided accurate and reliable results for a wide range of subjects and lighting conditions, it can occasionally fail. When this happens, our method cannot choose the correct head models for the reconstruction and the head model with a neutral face is rendered, which can lead to rendering artifacts especially for extreme facial expressions (large smile, wide opened mouth, etc.). Fortunately, this happens mostly when the user is looking away and not at the camera, therefore the choice of facial expression is not relevant in these cases. Our method does not depend of this particular face tracker, so the results can be improved by using a more accurate face tracker.

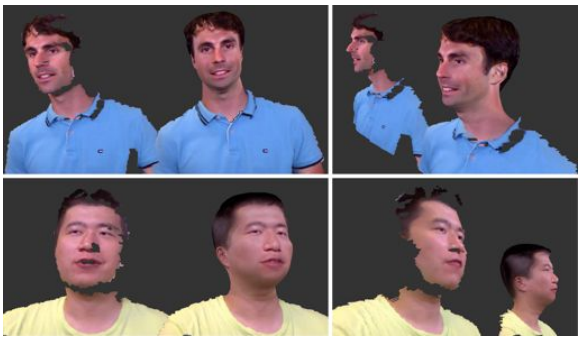

Figure 4: *For each image, (left) 3D reconstruction from sensor output, (right) 3D head reconstruction using our method.*

5. Conclusion and Future Work

We propose a method for reconstructing in 3D the head of remote users for consumer level 3D telepresence systems. This method consists in merging the live output of an hybrid sensor and data captured with the same sensor in a short preprocessing stage. In the future, we would like to engage in a large scale study to evaluate our system and better understand the features which are the most important to the user.

Acknowledgments

This research, which is carried out at BeingThere Centre, is supported by Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. This research and the Centre are also supported by ETH Zurich, NTU Singapore, and UNC Chapel Hill.

References

[BBB*10] BEELER T., BICKEL B., BEARDSLEY P., SUMNER B., GROSS M.: High-Quality Single-Shot Capture of Facial Geometry. *ACM ToG (siggraph'10) 29*, 3 (2010), 40:1–40:9. 1

[BHB*11] BEELER T., HAHN F., BRADLEY D., BICKEL B., BEARDSLEY P., GOTSMAN C., SUMNER R. W., GROSS M.: High-quality passive facial performance capture using anchor frames. *ACM ToG (siggraph'11) 30*, 4 (2011), 75:1–75:10. 1

[CGZZ10] CAI Q., GALLUP D., ZHANG C., ZHANG Z.: 3D Deformable Face Tracking with a Commodity Depth Camera. In *Proc. of ECCV* (2010), pp. 229–242. 1, 2

[IKH*11] IZADI S., KIM D., HILLIGES O., MOLYNEAUX D., NEWCOMBE R., KOHLI P., SHOTTON J., HODGES S., FREEMAN D., DAVISON A., FITZGIBBON A.: KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera. In *Proc. of UIST* (2011), pp. 559–568. 2

[KPB*12] KUSTER C., POPA T., BAZIN J.-C., GOTSMAN C., GROSS M.: Gaze Correction for Home Video Conferencing. *ACM ToG (siggraph asia'12) 31*, 6 (2012), 174:1–174:6. 4

[KS04] KRAEVOY V., SHEFFER A.: Cross-parameterization and Compatible Remeshing of 3D Models. *ACM ToG (siggraph'04) 23*, 3 (2004), 861–869. 2

[LYYB13] LI H., YU J., YE Y., BREGLER C.: Realtime Facial Animation with On-the-fly Correctives. *ACM ToG (siggraph'13) 32*, 4 (2013). 2

[MF11] MAIMONE A., FUCHS H.: A First Look at a Telepresence System with Room-Sized Real-Time 3D Capture and Large Tracked Display. In *Proc. of ICAT* (2011). 1

[PGB03] PÉREZ P., GANGNET M., BLAKE A.: Poisson Image Editing. *ACM ToG (siggraph'03) 22*, 3 (2003), 313–318. 3

[PL06] PIGHIN F., LEWIS J. P.: Performance-Driven Facial Animation. In *ACM Siggraph Courses* (2006). 1

[WBLP11] WEISE T., BOUAZIZ S., LI H., PAULY M.: Realtime Performance-Based Facial Animation. *ACM ToG (siggraph'11) 30*, 4 (2011). 2