A QUANTITATIVE EVALUATION FOR 3D FACE RECONSTRUCTION ALGORITHMS

Vuong Le, Yuxiao Hu, Thomas S. Huang

Beckman Institute for Advanced Science and Technology
Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
405 North Mathews Avenue, Urbana, IL 61801, USA

ABSTRACT

In this work, we proposed to use quantitative method to evaluate the accuracy of 3D face reconstruction algorithms. The reconstructed 3D faces are first aligned to the ground truth by Iterative Closest Point (ICP) algorithm and then the shape difference between the two 3D faces is described by Signal to Noise Ratio (SNR). Finally, the error maps (EM) illustrated the reconstruction errors on corresponded vertices in different dimensions. Comparing with the subjective and indirect evaluation methods, the proposed method provides more precise and detailed evaluations for face shape reconstruction. Based on the SNR, different 3D face reconstruction algorithms can be compared directly and the EM also can suggest guidance for feature extraction.

Index Terms— quantitative evaluation, 3D face reconstruction, error map, iterative closest points

1. INTRODUCTION

3D face reconstruction from 2D images is an important problem in computer vision. After decades, many approaches have been proposed, including 3D from stereo, morphable model based methods, structure from motion and shape from shading techniques [4]. Among them, morphable model based 3D face reconstruction algorithms have attracted more and more attention in recent years. Vetter et al. proposed a 3D fitting algorithm [7] to recover the shape and texture parameters of the 3D morphable model. In their solution, the shape parameters are obtained from the texture error. Their algorithm uses linear equations to recover the shape and texture parameters regardless of pose and lighting conditions of the input face image. Along the same line, the work of Hu et al. [4] proposed a fully automatic linear algorithm to recover the shape information according to sparsely corresponded 2D facial points. The 3D face geometry was recovered from a frontal view input face and then the texture of face was extracted from the input image directly.

Techniques based on morphable model have been proved to be effective in achieving high recognition rate and robust against different PIE (Pose - Illumination - Expression) conditions [4, 7]. In order to demonstrate the effectiveness and evaluate the accuracy of these 3D fitting algorithms, different applications are conducted from computer graphics to face recognition, In computer graphics, the reconstructed 3D faces are rendered in different PIE and driven by MPEG Facial Animation Parameters (FAP)[8], which enriched the human computer interaction and improved the user experiences. In face-based biometrics, the reconstructed 3D faces are used to normalize or expand the probe/gallery data set under different PIE conditions, so that the test patterns are closer to the reference patterns before the matching step [2, 7]. However, all the above evaluation methods are either based on subjective experiments or indirect evaluation on face recognition, i.e., there is no information about the shape/texture error provided, which made it difficult to further analyze the reconstruction algorithm and improve the features they used. An effective quantitative evaluation is required to give us the clues for finding the strength of the algorithms, investigating their weakness and suggesting further guidance for feature selection and algorithm refinement. In this paper, we proposed to use Signal-Noise Ratio (SNR) and error maps (EM) to quantitatively evaluate the accuracy of a 3D face reconstruction algorithm and provide its detail performance on shape recovery. The framework of our quantitative evaluation algorithm is described in fig. 1. From the ground truth 3D face database [3], we obtain the input 2D face image by projecting a 3D face onto 2D plane. This 2D face image and the extracted features will then be fed to the evaluated reconstruction system to get the reconstructed 3D face. To compare the ground truth 3D face shape and the reconstructed 3D face, they are first aligned to each other by *iterative closest point* algorithm. The difference returned from the fitting process will then be used as the error term for calculating SNR. These measurements on all the vertices will congregate to form the error map, which provides final detail result of the evaluation process.

2. EVALUATION METHOD FOR 3D RECONSTRUCTION ALGORITHMS

2.1. Subjective evaluation methods

The most straightforward method to evaluate the performance of 3D face reconstruction is judging by human eyes. Typi-

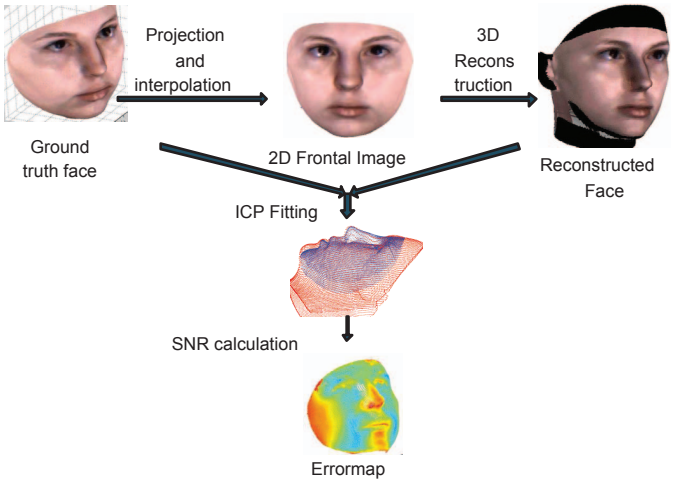

Fig. 1. Framework of quantitative evaluation algorithm

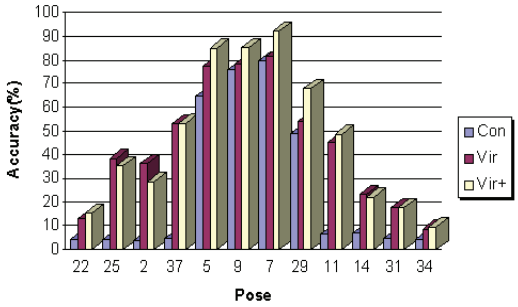

Fig. 2. Recognition accuracy comparison between face recognition using LDA. (Con: Conventional alg.; Vir: View based method with synthesized faces; Vir+: View based method, with the synthesized face images for individual pose)

cally, the reconstructed 3D faces are shown in different poses and expressions to confirm that they are realistic and with the same shape to the original face [5, 10]. Since people are more sensitive to familiar faces while not good at to tell the difference between strange faces [9], this kind of subjective experiments are heavily affected by the relation between the testing faces and the subjects and the face poses, which is not accurate and biased.

2.2. Indirect evaluation methods

Since one of the most important applications of 3D face reconstruction is biometrics, the reconstruction algorithms are often evaluated indirectly according to the face recognition accuracy based on the reconstructed faces. In face recognition vendor test in 2002 (FRVT2002) [6], Vetter et al.'s reconstruction algorithm is applied to non-frontal view face images. Then the reconstructed 3D faces are rotated to frontal view for face recognition, which substantially improved the face recognition accuracy on non-frontal view faces from about 50% to about 80% on a face database of 87 individuals. In Hu et al.'s work [4], a 3D face is first reconstructed from a frontal view face and then rendered in different poses, illuminations and expressions to enlarge the training data. Their face recognition performance based on CMU PIE database is shown in fig.2, from which a big improvement on near frontal view face recognition can be observed. Both of these works demonstrate the effectiveness of the 3D face reconstruction algorithms, because the better the 3D face reconstruction algorithm is, the higher recognition accuracy will be achieved. But since the performance of the 3D face reconstruction algorithm are evaluated indirectly based on different face database and testing strategies, it is difficult to compare different reconstruction algorithms. Moreover, the absolute accuracy of the face reconstruction is still unclear, so it is not easy to tell which parts of the face are more accurate than the other parts and how accurate they are.

2.3. Direct evaluation methods

The direct evaluation method based mostly on comparing the reconstruction results with the original ground truth face. These two may be represented by different number of vertices with (x,y,z) coordinates and (r,g,b) colors. So that, at the first step, we need to align the two faces by rotation and translation. *Iterative closest point algorithm* (ICP) is used for this fitting task. ICP algorithm was introduced in [1] as an accurate and efficient method for registration of 3D shapes which is independent on representation. The algorithm takes one shape considered as model and another as test data and output the translation and rotation for the data to fit with model. With respect to the mean square error objective function, ICP was proved to always converge monotonically to the local minimum.

After the faces are aligned, in order to calculate the error in terms of SNR at each vertex, we need to define the quantities for the roles of signal and noise. For noise, the obvious choice is the distance from considered point p on the original face O to the nearest point q in the reconstructed face after fitting R. This distance represents the misaligned quantity of the reconstruction at that point and is invariant to coordinate systems. The noise of each coordinate component can be expressed as the absolute differences

$$N_x^p = |p_x - q_x|, N_y^p = |p_y - q_y|, N_z^p = |p_z - q_z|$$

with $p \in O$ and q is the closest point to p in R. The combined noise at a position is defined to be the distance between two 3D points

$$N^p = \sqrt{N_x^{p2} + N_y^{p2} + N_z^{p2}} \qquad (1)$$

Unlike the case of noise, choosing a quantity to be in the role of signal amplitude is not straightforward. Naturally,

considering the mentioned definition of noise, we should use the coordinate amplitude of vertices as signals. However, this quantity depends on the frame of reference, i.e. when we change the origin or axes of coordinate system, these numbers will change. This fact would prevent the SNR from being well defined. We can address this issue by finding the coordinate system with which the amplitude of coordinate would have the smallest value. We can show that for a point $p(p_x, p_y, p_z)$ in the space, the summation of distances from it to all vertices in a set is minimized if it is the centroid $c(c_x, c_y, c_z)$ of the set.

$$\sum_{q \in O} norm(c-q) \geq \sum_{q \in O} norm(p-q) \qquad for\ all \qquad p \in \mathbf{R}^3$$

. From this observation, we will choose the centroid of the face to be the origin of the coordinate system and the maximum distance between the origin and the vertices will be considered as the amplitude of the signal.

$$S_x = \max_{p \in O}(|p_x - c_x|), S_y = \max_{p \in O}(|p_y - c_y|), S_z = \max_{p \in O}(|p_z - c_z|)$$

signal amplitude for combined evaluation will be

$$S = \max_{p \in O} \sqrt{(c_x - p_x)^2 + (c_y - p_y)^2 (c_z - p_z)^2} \qquad (2)$$

Once the signal S and noise N are ready, SNR can be calculated at each point as

$$SNR_x^p = 20 \times log_{10}(S_x / N_x^p)$$

$$SNR_y^p = 20 \times log_{10}(S_y / N_y^p)$$

$$SNR_z^p = 20 \times log_{10}(S_z / N_z^p)$$

and combined SNR:

$$SNR^p = 20 \times log_{10}(S / N^p) \qquad (3)$$

3. EVALUATION EXPERIMENTS AND RESULTS

3.1. Experiment configurations

We applied our evaluation method on the algorithm introduced in [4]. It first conducted the face detection and 2D face alignment on the input frontal image. After that, the allocated key facial points are used to compute the 3D shape coefficients of the shape morphable model. The result of this step is the 3D face shape represented by a combination of principal components. Hence, it has the same format of those components which includes 8955 3D points, covering the whole face region as the green shapes in fig. 3.

The ground truth used to evaluate above algorithm is IFP 3D face database [3], which consists of 675 3D faces obtained by laser scanner. Each face in ground truth is presented by 33420 vertices with dense correspondences, covering the frontal part of the face as depicted as red shapes in fig. 3.

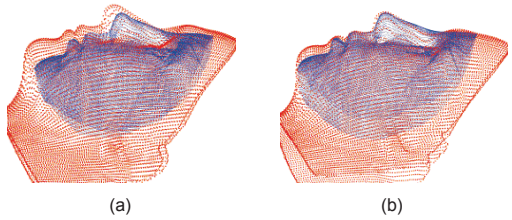

Fig. 3. Fitting result, original face is in green, reconstructed shape is in red (a) initial position; (b) fitted shapes.

In the first step of our process, for each face in the database, the frontal image is created by projecting the 3D face to the frontal plane. The resulted 2D face image is fed into the reconstruction system as an input. The output of this process is a 3D shape in the same scale but with different number of vertices. The vertices in the output 3D face are not corresponded to the ones in the original ground truth data. So we need to find the best fit for them and then calculate the difference between corresponding vertices, where ICP will be applied for the solution.

In our case, the reconstructed face covers the larger area of face than the original face. Thus, we will use it as model and the original face will act as fitted data. Since ICP can not guarantee global minimum, we first manually find the transformation parameters which give every pair of faces a relatively near-to-optimal position. The ICP iteration is then employed to get the fitting result (rotation and translation parameters) together with the square error on each vertex of the original face. A sample of starting position and fitting result is depicted in fig. 3a and 3b. respectively.

3.2. Experiment result

The experiment procedure is applied on 50 faces in IFP database. After fitting, the error maps are calculated as in (3). The mean of combined SNR is 28.96, for x component is 31.62, for y component is 38.64 and for z component is 30.00. The error maps are shown in fig. 4a-d and the metric used is SNR.

We can observe on the map that the most significant inaccuracy was made on nose tip, chin and cheek area of the face. Moreover, the error of z component is the most considerable. This can be explained by the nature of the evaluated reconstruction algorithm, which uses 2D facial points on the frontal face so that the z component is determined purely by the statistical information embedded the 3D morphable model.

3.3. Upper bound of reconstruction

The error we got from fitting the reconstructed and ground truth 3D faces comes from both the reconstruction inaccuracy and the noise of signal sampling and shape fitting. To assess

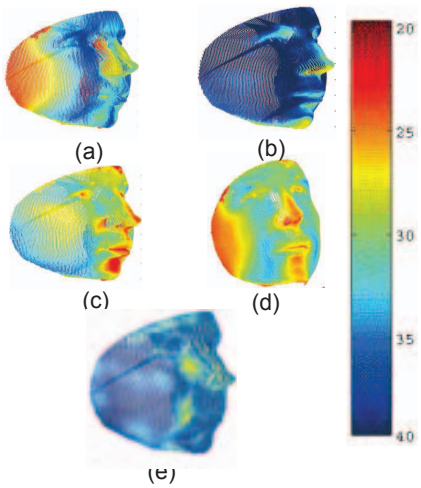

Fig. 4. Error map of (a) x components; (b) y components; (c) z components; (d) combined signal (e)ideal reconstruction

the significance of those two components, we directly converted the ground truth faces to the format of the output to simulate an ideal reconstruction. Comparing this ideal reconstruction result with the ground truth gives us the error caused only by sampling and fitting but not by reconstruction. The error map from this comparison is shown in fig. 4e and the combined SNR is 34.21.

The error map shows that the noise made by other causes is relatively uniform over the face. This fact proves that the variety of error at different parts of face as in fig. 4a-d comes from the reconstruction. This variety will provide us precious information about the strong and weak point of considered algorithm.

The ideal reconstruction error rate not only shows the underlying noise, but also provides a very important landmark, which can be used as the upper bound of the reconstruction performance that any reconstruction algorithms can reach.

4. CONCLUSIONS AND DISCUSSION

In this paper, the algorithm for quantitative performance evaluation of a face reconstruction systems is proposed. The error maps with SNR metric are built to provide both local and global assessment of the inaccuracy introduced by the reconstruction algorithm. The ideal reconstruction case was also given to evaluate the underlying error made by format conversion for the upper bound of the algorithms performance, so that the SNR can be evaluated by relative comparison.

From the evaluation result on the selected reconstruction algorithm [2], the error concentrates at some particular parts of faces such as nose tip, chin and cheek. The reason for the inaccurate nose and chin height is that, in the reconstruction algorithm we evaluated, the depth information are deduced from the width/height information in the 2D face according to the statistical model, whose performance is limited by the size of the training set. The reason for the inaccurate cheek is that, the 2D facial contour actually is formed by projection, which are not well defined in the 3D face. So their corresponded points can not be precisely located. In summary, the inaccuracy comes from the limited training data and the lack of control points at salient and complicated areas. These observations suggests new strategies of 2D control points configuration and build larger database or specific training set for different ethnicity/gender groups.

In the future works, the performance comparison among various algorithms based on the proposed direct quantitative evaluation will be made. The above shape error evaluation algorithm can also be applied for texture error evaluation. This will provide us more concrete perspective of the strength and weakness of each algorithms and suggest the combination strategy to achieve faster and more accurate reconstruction results.

5. REFERENCES

[1] P. Besl and N. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14:239–256, 1992.

[2] Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. Zhang. Automatic 3d reconstruction for face rec. *Intl. Conf. on Automatic Face and Gesture Recog. (FGR2004)*.

[3] Y. Hu, Z. Zhang, X. Xu, Y. Fu, and T. Huang. Building large scale 3d face database for face analysis. *MCAM2007*.

[4] D. Jiang, Y. Hu, and S. Yan. Efficient 3d recons. for face rec. *Journal of Patern Recog., Spec. Iss. on Img. Unders. for Dig. Photographs (PR2005)*, 38:787–798, 2005.

[5] G. Kalberer, P. Muller, and L. Gool. Modeling and synthesis of realistic visual speech in 3d. *Idea Group Publishing Inc., Hershey*, 2003.

[6] P. Phillips, P. Grother, R. Micheals, and et al. Face rec. vendor test 2002: Evaluation report. *FRVT, 2002*.

[7] S. Romdhani, V. Blanz, and T. Vetter. Face ident. by fitting a 3d morphable model using lin. shape and tex. error functions. *CVPR 2005*.

[8] H. Tang, Y. Hu, Y. Fu, M. Hasegawa-Johnson, and T. Huang. Real-time conversion from a single 2d face image to a 3d text-driven emotive a/v avatar. *ICME2008*.

[9] C. Wallraven, A. Schwaninger, and H. Bthoff. Learning from humans: Comp. modeling of face rec. *Network 16(4)401-418 (12 2005)*.

[10] Z. Wen and T. S. Huang. 3d face proc.: Modeling, anal. and synthesis. *The Intl. Series in Video Comp.*, 8, 2004.