3D HUMAN FACE MODELLING FROM UNCALIBRATED IMAGES USING SPLINE BASED DEFORMATION

Nikos Barbalios, Nikos Nikolaidis and Ioannis Pitas
Department of Informatics, Aristotle University of Thessaloniki,GR-54124, Thessaloniki, Greece, (+30) 2310 996361

Keywords: Uncalibrated 3D face reconstruction, face modelling, 3D reconstruction, mesh deformation, model based 3D reconstruction.

Abstract: Accurate and plausible 3D face reconstruction remains a difficult problem up to this day, despite the tremendous advances in computer technology and the continuous growth of the applications utilizing 3D face models (e.g. biometrics, movies, gaming). In this paper, a two-step technique for efficient 3D face reconstruction from a set of face images acquired using an uncalibrated camera is presented. Initially, a robust structure from motion (SfM) algorithm is applied over a set of manually selected salient image features to retrieve an estimate of their 3D coordinates. These estimates are further utilized to deform a generic 3D face model, using smoothing splines, and adapt it to the characteristics of a human face.

1 INTRODUCTION

Reconstructing an object's geometry from a set of images has been one of the most popular and challenging tasks in computer vision with applications such as virtual view synthesis, gaming, security, virtual reality, 3DTV etc. 3D face reconstruction is a special case of this problem. A popular solution to this problem are the so-called Structure from Motion (SfM) algorithms, that receive a number of points as input and return the 3D coordinates of those points

Applications that utilize face models, (i.e. advertising, gaming, movie industry) prescribe that the generated 3D face model should not only be accurate but plausible and realistic as well. Faces, however, are essentially non-rigid surfaces and only a few features can be reliably extracted and matched across facial images, thus point-based approaches proposed in the literature, (Mendonca and Cipolla(1999); Hartley(1994); Pollefeys and Gool(1997)), result to non-plausible face models restricted only to a small number of points. To compensate for the sparseness of the reconstructed 3D points, almost all algorithms use a prototype face model. Two types of prototype models are used: generic mesh models and 3D morphable models.

Generic mesh models comprise of a set of nodes corresponding to characteristic facial points of an average person. The generic model can be incorporated in the algorithm either before, during or after the 3D reconstruction procedure. In (J. Strom(1999)) the generic mesh model was used to initialize the 3D reconstruction procedure whereas in (Y. Shan(2001); Fua(2000)), it was used to constrain the so-called bundle adjustment optimization. Although these two methods yield good results, the algorithm converges close to the initial estimation resulting to a large bias towards the generic model. This problem was alleviated in (A.R. Chowdhury(2002); R. Hassanpour(2004)), by proposing the incorporation and deformation of the generic model after the optical-flow based estimation of the 3D structure coordinates.

3D morphable models, initially introduced in (Blanz and Vetter(1999)), are derived by transforming the shape and texture information of a database containing high resolution head laser scans into a vector space representation. By doing so, the 3D head is parameterized in a statistical way by a shape vector, containing the 3D coordinates of the model's nodes, and a texture vector, containing the corresponding texture value for each node. New faces can be obtained by forming linear combinations of these shape and texture vectors. Despite its realistic results, the main drawbacks of this approach are the lack of accuracy, its instability to illumination changes and the huge computational cost for the convergence of the algorithm, issues that where partially addressed in (M. Dimitrijevic(2004); M. Zhao(2006)).

Barbalios N., Nikolaidis N. and Pitas I. (2008).
3D HUMAN FACE MODELLING FROM UNCALIBRATED IMAGES USING SPLINE BASED DEFORMATION.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 455-459
DOI: 10.5220/0001081204550459

The proposed method is a combination of a robust SfM algorithm, introduced in (M. Pollefeys and Gool(1998)), with a 3D generic mesh face model. Salient features of the face are manually marked on the set of images and the SfM algorithm is used to produce a cloud of 3D points. This point cloud is subsequently utilized for the deformation of the generic face model. The proposed method is related to (R. Hassanpour(2004)) and (A.R. Chowdhury(2002)) since it also deforms the generic model after the SfM algorithm. However, in the proposed method the deformation of the generic mesh model is implemented using a different approach, namely thin plate smoothing spline functions which, as shown in the experimental results, accurately estimates the deformation of each node of the model. Moreover, our algorithm yields satisfactory results using fewer input images.

2 3D RECONSTRUCTION

Our method is based on the iterative SfM algorithm proposed in (M. Pollefeys and Gool(1998)). This algorithm is a combination of a projective SfM algorithm with an auto-calibration algorithm to restrict the ambiguity of the reconstruction from projective to metric. The basic steps of the algorithm will be shortly described below for the sake of completeness of this paper.

The method is based on the ideal pinhole camera model. In such a camera, the projection $\mathbf{m} = [x,y,0]^T$, on the image plane, of an object point $\mathbf{M} = [X,Y,Z,0]^T$ is described, in homogeneous coordinates, by

$$\mathbf{m} \sim \mathbf{P} \cdot \mathbf{M} \qquad (1)$$

where $\mathbf{P}$ is the 3×4 projection matrix of the camera, encoding the camera's intrinsic and extrinsic parameters. The symbol $\sim$ denotes that equality is valid up to an arbitrary scale factor, since we use homogeneous coordinates.

Initially, a set of N characteristic points (features) are extracted from the image set and their correspondences across the images are identified. Using those feature points, the 3×3 fundamental matrix $\mathbf{F_{ij}}$, which is an algebraic representation of the epipolar geometry between two views i and j, can be evaluated using the following homogeneous system of equations:

$$\mathbf{q_i^T} \cdot \mathbf{F_{ij}} \cdot \mathbf{q_j^T} = \mathbf{0} \qquad (2)$$

where $\mathbf{q_i} = [\mathbf{m_{i,1}}, \mathbf{m_{i,2}}, \mathbf{m_{i,3}}..., , \mathbf{m_{i,N}}]^\mathbf{T}$ and $\mathbf{m_{i,k}}$ denotes the k-th feature in image i. Note that since $\mathbf{F_{ij}}$ has 7 degrees of freedom (M. Pollefeys and Gool(1998)), at least 7 features should be present in all the images of the set, i.e. $N \geq 7$.

After evaluating, using (2), the fundamental matrix $\mathbf{F_{12}}$ between the views 1 and 2, the projection matrices $\mathbf{P_1}$ and $\mathbf{P_2}$ for those two views can be estimated. If we consider that the first view is aligned with the world coordinate system, its projection matrix is $\mathbf{P_1} = [\mathbf{I_{3\times3}}|\mathbf{0_3}]$ where $\mathbf{I}, \mathbf{0}$ denote the identity matrix and the all-zero vector respectively. Subsequently, using the epipolar constraint the projection matrix $\mathbf{P_2}$ of the second view is constrained to $\mathbf{P_2} = [[\mathbf{e_2}] \times \mathbf{F_{12}} + \mathbf{e_2}\tau^\mathbf{T}|\sigma\mathbf{e_2}]$,where $\tau^\mathbf{T}$ is a 3×1 vector determining the reference frame of the reconstruction, σ is the scale factor of the reconstruction and $\mathbf{e_2}$ the position of the epipole in the second image. After the evaluation of the two projection matrices, an estimate $\mathbf{M_k}$ of the position of the k-th feature point ($k = 1,...,N$) in the projective space can be obtained using (1).

In the next step, this projective reconstruction $\mathbf{M_k}$, is further used to iteratively retrieve, through (1), the projection matrices for all the additional views of the image set. Up to this step, the evaluation of all the projection matrices is based on the projective reconstruction $\mathbf{M_k}$ retrieved only from the first two views. The next step, which is called bundle adjustment, comprises the overall optimization of the reconstruction which refines both the reconstructed structure and the projection matrices for all the views. The optimization criterion is the overall minimization for all the views of the re-projection error, i.e. the distance between the 2D feature points and the estimates produced by back-projecting the produced 3D model on the images :

$$(\mathbf{P_i}, \mathbf{M_k}) = \arg\min_{P_i, M_k} \sum_i \sum_k (\mathbf{m_{i,k}} - \mathbf{P_i} \cdot \mathbf{M_k})^2 \qquad (3)$$

where $\mathbf{P_i}$ is the projection matrix for view i, $\mathbf{M_k}$ is the 3D coordinate vector of the k-th feature and $\mathbf{m_{i,k}}$ denotes the 2D coordinates vector of the k-th feature in image i.

The final step of the algorithm is the self-calibration procedure which retrieves, for each view i, the camera intrinsic parameters matrix $\mathbf{K_i}$ and upgrades the structure to the metric space (M. Pollefeys and Gool(1998)). Once the intrinsic parameters are retrieved, the reconstructed 3D points are transformed from the projective space into the metric space, leading to a 3D point cloud.

In our implementation, we used 15 characteristic points of the human face namely the corners of the mouth, the midpoints of the lips, the tip and base of the nose, the corners of the eyes, the point between

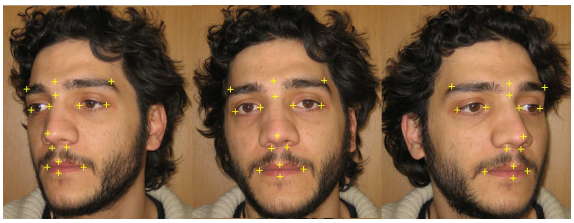

Figure 1: The selected characteristic points depicted on a sequence of facial images.

the eyebrows and two characteristic points of the eyebrows. These points, shown in Figure 1, are selected not only because they are easily identifiable, but also because they have a representation in the nodes of the generic model used. It is also worth mentioning that these points and their correspondences are manually defined and thus, the absence of large position errors and false correspondences between them is assumed.

3 GENERIC MODEL DEFORMATION

The second part of our algorithm deals with the incorporation of a generic face model into the reconstruction procedure. Our algorithm uses a generic face model derived from the Candide face model, developed by the Linkoping University (Ahlberg(2001)). Since Candide is a rather limited resolution model, comprised of 103 3D nodes and 184 triangles, it has been enhanced in three steps: Initially, extra nodes have been added in areas where the original model has very few nodes (i.e. nose and eyes areas), to increase the deformation flexibility of those areas. Then the PN triangles mesh subdivision algorithm (A. Vlachos(2001)) has been applied leading to an enhanced generic model composed by 1729 nodes and 3392 triangles. Finally, the enhanced model was smoothed to eliminate the sharp-edges and its faceted appearance leading to the final generic model used in our algorithm. It should be noted that the nodes of the original model are included in the nodes of the enhanced model.

During the first step of the deformation, the generic model undergoes global rotation, translation and scaling, so that it is roughly aligned with the cloud of 3D points. To determine the scale factor, the mean distance between the two corners of the eyes and the two corners of the mouth is evaluated both in the generic model and the point cloud and their ratio is used as the scale factor. After scaling, the generic model is rotated so that its orientation is roughly the same with that of the point cloud. Finally, the generic model is translated so as to minimize the sum of Euclidean distances between the eyes and mouth centers of the generic model and the point cloud.

Having aligned the generic model and the point cloud, the next step is to deform all the nodes of the generic model according to the reconstructed 3D point cloud. This problem can be formulated as follows. Given the 3D points $\mathbf{G_i}$ of the generic model ($i = 1,...,L$ where $L = 1729$), the 3D points $\mathbf{M_k}$ of the point cloud ($k = 1,..,N$ where $N = 15$) and the distances $[dx_k, dy_k, dz_k]$, $k = 1,...,N$ between the N 3D points of the cloud and their corresponding nodes of the generic model, our objective is to estimate the new coordinates $\mathbf{G'_i}$ of the $L-N$ nodes of the generic model that do not have a corresponding point in the point cloud.

We introduce three error functions

$$\begin{aligned} f_x(x) &= dx = x - x' \\ f_y(y) &= dy = y - y' \\ f_z(z) &= dz = z - z' \end{aligned} \tag{4}$$

where $[x,y,z]^T$ are the coordinates of a node of the generic model and $[x',y',z']$ the coordinates of the corresponding point of a smooth surface that passes through the points $\mathbf{M_k}$ of the point cloud. In other words, these error functions provide the distance $\mathbf{D_i} = [dx_i, dy_i, dz_i]^T$ along the three axes between a point $\mathbf{G_i}$ of the generic model and its corresponding point in the target surface. Since these error functions can be evaluated only for the N model nodes that have a corresponding point $\mathbf{M_k}$ in the point cloud, the values $\mathbf{D_i}$ for the rest of the generic model nodes should be estimated through extrapolation and interpolation of those N known values.

For this purpose we used thin plate smoothing spline functions, which interpolate a curve over a fixed set of nodes in a way that minimizes the bending energy of the curve (Bookstein(1989)). The interpolation of the aforementioned error functions results to a vector field defined in the 3-dimensional space that expresses the difference between the generic model and a smooth surface passing through the 15 points of the point cloud. Using these error functions, the distances $\mathbf{D_i} = [dx_i, dy_i, dz_i]^T$, $i = 1,...,L$, for all the nodes of the generic model can be evaluated. Thus the new coordinates $\mathbf{G'_i}$ of each node of the deformed 3D face model are derived through $\mathbf{G'_i} = \mathbf{G_i} + \mathbf{D_i}$.

Initially, the functions $f_x(x)$ and $f_y(y)$ are interpolated, estimating the displacements of the generic model's nodes along the x and y axis respectively. Interpolation of function $f_z(z)$, which estimates the depth of the face, is done separately and requires different handling. This can be explained intuitively by

realizing that if one's eyes are wide apart, the mouth, nose and the whole face are likely to be wider as well. For the depth of a face however this is not true since the fact that somebody has a long nose has no effect in the overall distribution of the face nodes along the z axis. Additionally, the depth of the various parts of the face, i.e. face, nose, eyes, may vary dramatically which prescribes that they should be dealt separately. For all those reasons, the various parts of the face are interpolated seperately along the z axis, and the final model emerges by combining those partial $3D$ patches into the final deformed 3D face model.

Up to this point, all the error functions have been interpolated and the deformed generic model can be obtained. The final stage of the deformation step aims to improve the model's visual appeal by adding texture to the model. This is done by re-projecting the deformed 3D face model on a 2D image and generating texture for each triangle of the face mesh. Since no corresponding color is available for some triangles, due to occlusion, the blank areas are filled by means of color interpolation from their neighbor areas.

4 EXPERIMENTAL RESULTS

For our experiments we used a variety of sets of high resolution (2048×1536) facial images, acquired with an off-the-shelf uncalibrated camera, and user supplied features and correspondences. We used 3 images per set, which is the minimum number of images for a proper reconstruction, and obtained several face models which were later used to evaluate our algorithm. Experiments conducted with sets containing more than 3 images lead to similar results.

The deformed textured model, derived from the 3 facial images depicted in Figure 1, is presented in Figure 2. It can be seen that the model encodes efficiently well the basic facial characteristics of the subject. Despite the fact that the camera positions used to acquire the facial images are quite close to each other, the proposed method estimates adequately well the depth of the face and the produced 3D face model resembles the subject considerably well.

Table 1: Mean, min and max normalized approximation error of 8 models.

Mean Error	Min Error	Max Error
0.006	0.000	0.044

Due to the lack of real 3D face data (i.e. laser scans), we used the face models produced by our algorithm as ground truth to assess the effectiveness of our method. More specifically, synthetic images, each corresponding to a different view of a certain 3D face model, were obtained and used as input image sets for our algorithm. The produced 3D face model was subsequently compared with the corresponding original 3D model. The objective criteria to measure the similarity between the two face models was the $3D$ Euclidian distances between the 116 nodes of the original face model and their corresponding nodes in the produced face model. These nodes represent the basic nodes of the Candide model along with the nodes we added to enhance the resolution of the model in the nose area. Results are presented in Table 1 where the mean, minimum and maximum approximation errors averaged over 8 such reconstructions is depicted. The approximation error is calculated for all the 116 nodes of the 3D face. It should be noted that the approximation errors in this table have been normalized with the height of the produced 3D face and thus an error value of 0.01 corresponds to a approximation error equal to $1/100$-th of the head height.

Figure 2: Different views of the final reconstructed 3D model for the image set depicted in Figure 1.

It is obvious that despite the fact that the generic model was deformed using information only from 15 3D points, the use of thin-plate smoothing splines produced efficient estimates of the location of the rest of the nodes and the reconstructed 3D model is very close to the original. As expected, the mean error for all the nodes of the model is considerably small (0.06). The experimental procedure revealed that only 5 out of the 116 nodes tend to produce errors larger than 0.01. These 5 nodes correspond to nodes in the forehead area where the mean error is larger than the rest of the face due to the lack of 3D reconstructed points in the area.

An alternative way to evaluate the effectiveness of our method is the re-projection error for all the images in an image set. Experiments affirmed that the re-projection error is significant small since it does not exceed 4 pixels for images of dimensions 2048×1536.

Through experimentation we have reached the conclusion that the quality of the results is mainly affected by the selection of the first two views. These views should correspond to views taken from angles that are as far apart as possible to capture adequately the depth of the face. Furthermore, the accuracy of the feature selection procedure largely affects the accuracy of the reconstruction, since the reconstruction algorithm is based on the determination of the epipoles and hence is very sensitive to measurement noise.

5 CONCLUSIONS

In this paper, we have presented a two-step technique to deal with the challenging task of face reconstruction in three dimensions from a set of uncalibrated images. In the first step of the proposed approach, 15 salient features are manually identified in all the images of the set and their 3D coordinates are retrieved using an uncalibrated 3D reconstruction algorithm. In the second step, a generic face model is deformed according to the 3D points produced in the first step. Since the 3D coordinates of only 15 points are known, the 3D coordinates for the rest nodes of the generic model are retrieved through interpolation and extrapolation, with extra provisions to ensure a proper face reconstruction. The experimental results prove that the combination of a robust SfM algorithm with a generic mesh model deformed using thin plate smoothing splines can yield very satisfactory 3D reconstructions.

ACKNOWLEDGEMENTS

This work was supported by the ”SIMILAR” European Network of Excelence on Multimodal Interfaces of the IST Programme of the European Union (www.similar.cc).

REFERENCES

P. R. S. Mendonca and R. Cipolla. A simple technique for self-calibration. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 500–505, 1999.

R. I. Hartley. Euclidean reconstruction from uncalibrated views. In *Proc. of the 2nd Joint European - US Workshop on Applications of Invariance in Computer Vision*, pages 237–256. Springer-Verlag, 1994.

M. Pollefeys and L. Van Gool. A stratified approach to metric self-calibration. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recongition (CVPR)*, pages 407–418, 1997.

S. Basu A. Pentland J. Strom, T. Jebara. Real time tracking and modeling of faces: An ekf-based analysis by synthesis approach. Technical Report 506, MIT Media Laboratory, 1999.

Z. Zhang Y. Shan, Z. Liu. Model-based bundle adjustment with application to face modeling. In *Proc. of IEEE Internation Conference on Computer Vision (ICCV)*, volume 2, pages 644–651, Vancouver, BC, Canada, 2001.

P. Fua. Regularized bundle-adjustment to model heads from image sequences without calibration data. *International Journal of Computer Vision*, 38(2):153–171, 2000.

S. Krishnamurthy T. Vo A.R. Chowdhury, R. Chellappa. 3d face reconstruction from video using a generic model. In *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, volume 1, pages 449– 452, 2002.

V. Atalay R. Hassanpour. Delaunay triangulation based 3d human face modeling from uncalibrated images. In *Proc. of Computer Vision and Pattern Recognition Workshop (CVPR)*, pages 75– 75, June 2004.

V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proc. of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 187–194, New York, NY, USA, 1999.

P. Fua M. Dimitrijevic, S. Ilic. Accurate face models from uncalibrated and ill-lit video sequences. In *Proc of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1034–1041, 2004.

T. Sim M. Zhao, T. Seng Chua. Morphable face reconstruction with multiple images. In *Proc. of the International Conference on Automatic Face and Gesture Recognition (AFGR)*, pages 597–602. IEEE Computer Society, 2006.

R. Koch M. Pollefeys and L. Van Gool. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *Proc. of the International Conference on Computer Vision (ICCV)*, pages 90–95, 1998.

J. Ahlberg. Candide 3 - an updated parameterized face. Technical Report LiTH-ISY-R-2326, Dept. of Elegxtrial Engineering, Linkoping University, Sweden, 2001.

C. Boyd J.L. Mitchell A. Vlachos, J. Peters. Curved pn triangles. In *Proc. of the 2001 Symposium on Interactive 3D graphics*, pages 159–166. ACM Press, 2001.

F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE trans. Pattern Analysis and Machine Intelligence (PAMI)*, 11 (6):567–585, 1989.