

Modeling 3D Objects from Range Maps and Color Images using a Warping-based Approach

Faysal Boughorbel, Andreas Koschan, and Mongi Abidi

Imaging, Robotics and Intelligent Systems Lab, Department of Electrical and Computer Engineering, 409 Ferris Hall, The University of Tennessee, Knoxville, TN, 37996, United States, Phone: 1-865-974-9213, Fax: 1-865-974-5459, e-mail: fboughor@utk.edu

Keywords: 3D modeling, warping, shape from motion, range, video.

1. Introduction

Two important means of acquiring 3D models of real world scenes and objects are range scanners and stereovision systems. The first approach is usually implemented through laser-range mapping techniques, which result in dense depth maps, with relatively high accuracy. On the other hand, stereo methods reconstruct the scene geometry by finding corresponding points in two or more images, then triangulating these matches. Stereovision is one of the earliest methods for the recovery of scene structure from imagery. Unfortunately, building accurate dense depth maps using the latter approach is made difficult by the correspondence problem. Nevertheless, stereovision is still a popular method for many applications. One of its advantages is the possibility of real time implementation in video-based modeling pipelines [1]. In fact, despite the increase in the acquisition speed of high-resolution laser scanners, real time systems are still rare and expensive.

In this paper we describe an approach that integrates both range scanning and stereo-motion techniques. Our goal is to reconstruct the 3D geometry of objects from video streams, without using dense depth matching. To achieve dense reconstruction, we start with a generic model of the object of interest, acquired by a high-resolution laser range scanner, then, a small number of high-confidence feature points are tracked throughout an image sequence of the object of interest and matched with the range data. The 3D world points corresponding to the image feature-points are reconstructed using a structure from motion (SFM) technique. This, in effect, leads to a set of 3D-3D correspondences between the range and stereo data. Using these matches, the model of the object is obtained by 3D warping of the generic model. We applied this approach to the reconstruction of faces from video streams. In section 2 we will describe the method adopted for the recovery of 3D feature points from image sequences. Section 3 will present the thin-plate splines interpolation technique that we used for 3D warping. Section 4 will describe the face modeling results, and finally section 5 will summarize our conclusions.

2. Structure from image sequences

The reconstruction of the structure of a scene from a sequence of images, acquired by a moving camera, is a classic problem in computer vision, known as Shape from Motion (SFM) [2]. Most

algorithms are based on point-feature correspondences along the image sequence [3]. In the following, we will consider calibrated cameras. The problem of modeling scenes using uncalibrated cameras received large attention during the last years [4], but is not our focus in this research.

Some methods, such as the factorization algorithm [5], attempt the recovery of motion and structure in a single step by searching for a closed form solution. These approaches, however, require assumptions about the camera models that are not always realized. More common are techniques that first recover the camera's motion from the correspondences and then reconstruct the 3D points using triangulation. The basis of these algorithms is the coplanarity constraint, relating a 3D world point M and its projection in two images. If the second camera's reference frame is related to the first one by the rigid transformations (\mathbf{R}, \mathbf{t}) , the 3D rotation and translation, then the coplanarity constraint imposes that the rays \mathbf{m}_1 and \mathbf{m}_2 , pointing from the cameras optical centers to the world point, be located in the same plane. This can be expressed as follows:

$$\mathbf{m}_1 \cdot (\mathbf{t} \times \mathbf{R}\mathbf{m}_2) = 0 \quad (1)$$

In our system we first used a linear method for the recovery of (\mathbf{R}, \mathbf{t}) followed by a non-linear refinement step [4][6]. In the linear method (1) is rewritten as:

$$\mathbf{m}_1^T \mathbf{E} \mathbf{m}_2 = 0 \quad (2)$$

\mathbf{E} is a 3x3 matrix, known as the essential matrix, that embeds the motion parameters: $\mathbf{E} = [\mathbf{t}]_x \mathbf{R}$, with $[\mathbf{t}]_x$ being the anti-symmetric matrix representing the cross product. A system of homogeneous equations in the entries of \mathbf{E} is then set from the point matches. The usually over-constrained system (for more than 8 matches) is solved using singular value decomposition (SVD), and the motion parameters are easily computed from the essential matrix. The scale of the translation, and hence of the reconstruction is inherently ambiguous in the SFM problem and is set arbitrarily.

A non-linear refinement step is normally required. Several formulations were proposed, where the rotation matrix was parametrized as orthonormal matrices or unit quaternions. In this work we used a classic method based on the least squares minimization [6]:

$$\min_{\mathbf{R}, \mathbf{t}} \sum_i \left\| \mathbf{m}_1^i \cdot (\mathbf{t} \times \mathbf{R}\mathbf{m}_2^i) \right\|^2 \quad (3)$$

After the recovery of the motion parameters, 3D points are reconstructed through robust triangulation from matches in two or more images.

3. 3D data Warping

After correspondences are established between the points reconstructed by SFM and the generic range data, we obtain a set of 3D-3D matches. Our approach to dense shape recovery is to find a warping transformation that maps the generic model into the interest object, visible in the image sequences. To account for local shape variations, we used thin-plate splines interpolation [7], a

method popular in many non-rigid registration applications, particularly in medical imaging [8]. The method provides an interpolation function f , which maps one set of points, the source set, to another corresponding set, the target set. Let $\{P_i(x_i, y_i, z_i), i = 1, \dots, n\}$ be the set of source points, and $\{Q_i(x_i', y_i', z_i'), i = 1, \dots, n\}$ be the set of target points. If $r_{ij} = |P_i - P_j|$ is the distance between two source points, then the function f is a function that maps the corresponding points exactly, while minimizing the quadratic semi-norm $|f|^2$, interpreted physically as the bending energy of a thin plate of infinite extent.

The resulting function is the sum of two terms: an affine term representing its behavior at infinity, and a second term, which is asymptotically flat:

$$f(x, y, z) = a_1 + a_x x + a_y y + a_z z + \sum_{j=1}^n \mathbf{w}_j U(|P_j - (x, y, z)|) \quad (4)$$

The basis function U is the fundamental solution to the biharmonic equation $\Delta^2 f = 0$. In 3D the function is $U(r) = |r|$, whereas $U(r) = r^2 \ln r$ in 2D, and $U(r) = |r|^3$ in 1D. For each coordinate, three functions: f_x , f_y , and f_z expressed as in (4) are to be recovered.

If $a = (a_1, a_x, a_y, a_z)$ and $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_n)$, the interpolating function f is obtained by solving the linear system:

$$\begin{cases} K\mathbf{w} + Pa = v \\ P^T \mathbf{w} = 0 \end{cases} \quad (5)$$

Where $K = (U(r_{ij}))$, $P = \begin{bmatrix} 1 & x_1 & y_1 & z_1 \\ \dots & \dots & \dots & \dots \\ 1 & x_n & y_n & z_n \end{bmatrix}$, and v is the vector of one coordinate of the

target set (e.g. $v = (x_1', \dots, x_n')$ in the case of the x coordinate).

4. Results

We applied the reconstruction technique, described above, to the problem of face reconstruction from video streams. This task has several applications in the fields of computer vision and graphics, such as 3D face recognition and facial animation [9]. A generic face model, displayed in Fig. 1(c), was reconstructed from laser range scans of the mannequin's head shown in Fig. 1(a). Strategically located landmark points are chosen in the range image by the user (Fig. 1(b)). An image sequence of the moving face of interest was captured, and the feature-points were tracked throughout the sequence (Fig. 1(d)). To perform the tracking, pre-defined facial features templates were used and matched with the images. After reconstructing the tracked points and warping the range data as described in section 3, we obtained a 3D model of the face, rendered in Fig. 2(a). Texture can be mapped into this model using the color images as shown in Fig. 2(b). It

can be seen from these results that the method was able to capture a significant level of detail from the modeled object. In particular, the use of thin-plate splines interpolation allowed for a good approximation of both global and local shape variations.

5. Conclusion

A technique for modeling objects from both range maps and color image sequences was described. The method is particularly useful in modeling objects for which the generic shape is known, such as human faces. It was shown experimentally, that a combination of a Shape from Motion algorithm and 3D thin-plate splines warping allows for the recovery of accurate face geometry. The proposed method combined the high-density of laser range scanning with the speed of stereo-motion algorithms.

Faysal Boughorbel received his B.S. and M.S. degrees from The National School of Engineers of Tunis, Tunisia, in 1997 and 1999, respectively, both in Electrical Engineering. He has also worked as a visiting research scholar with the University of Tennessee, Knoxville. He is currently a Ph.D. candidate in Electrical and Computer Engineering at the University of Tennessee, Knoxville, and serves as a research assistant in the Imaging, Robotics, and Intelligent Systems Laboratory. His research interests include computer vision, image processing and pattern recognition.

References

- [1] ZISSERMAN, A., FITZGIBBON, A. W., and CROSS, G.: 'VHS to VRML: 3D graphical models from video sequences'. Proc. IEEE Int. Conf. on Multimedia and Systems, Florence, June, 1999, pp. 51-57.
- [2] JEBARA, T., AZARBAYEJANI, A., PENTLAND, A.: '3D structure from 2D motion'. Signal Processing Magazine, 1999, pp. 66-84.
- [3] SCHMID, C., MOHR, R., BAUCKHAGE, C.: 'Evaluation of interest point detectors'. International Journal of Computer Vision, 37(2), 2000, pp. 151-172.
- [4] HARTLEY, R. and ZISSERMAN, A.: 'Multiple view geometry in computer vision' (Cambridge University Press, June, 2000).
- [5] TOMASI, C., and KANADE, T.: 'Shape and motion from image streams under orthography---a factorization method', International Journal on Computer Vision, 9(2), Nov. 1992, pp. 137-154.
- [6] FAUGERAS, O.: 'Three-Dimensional computer vision' (The MIT Press, Cambridge Massachusetts, 1993).

[7] BOOKSTEIN, F. L.: 'Principal warps: Thin-plate splines and the decomposition of deformations'. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(6), 1989. pp. 567-585.

[8] TOGA A. W.: 'Brain Warping' (Academic Press, San Diego, 1999).

[9] LIU, Z., ZHANG, Z., JACOBS, C., and COHEN, M.: 'Rapid modeling of animated faces from video'. Technical Report, Microsoft Research 99-21, April 1999.

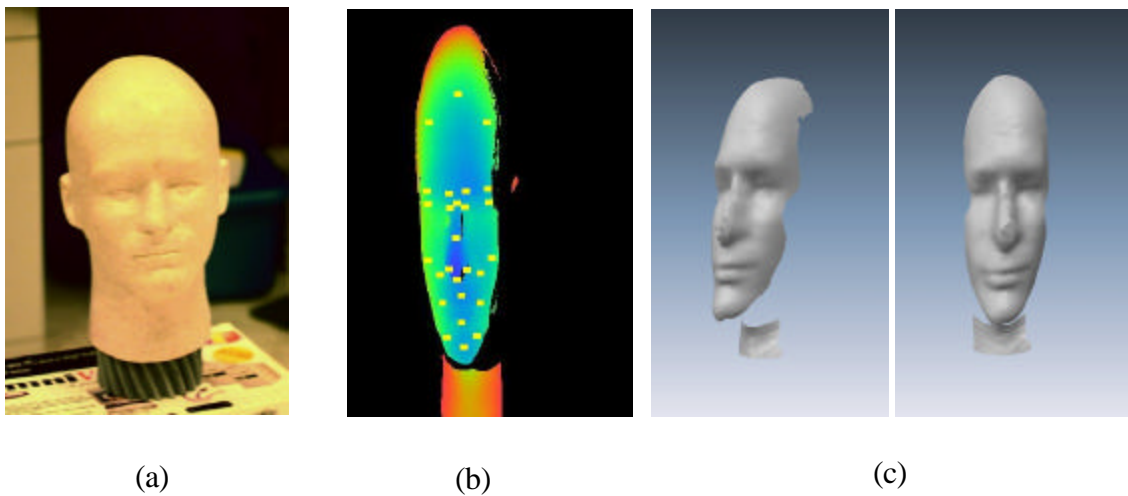
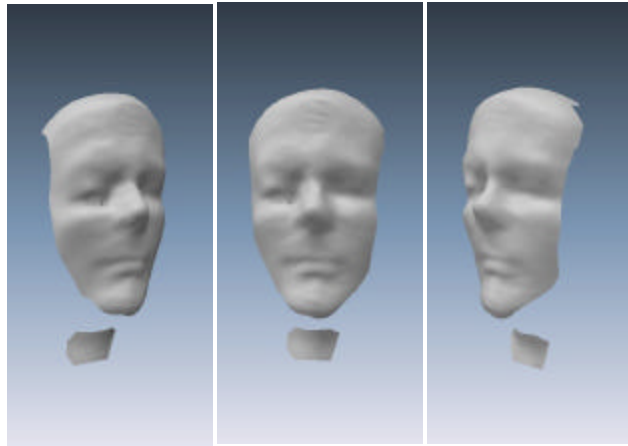
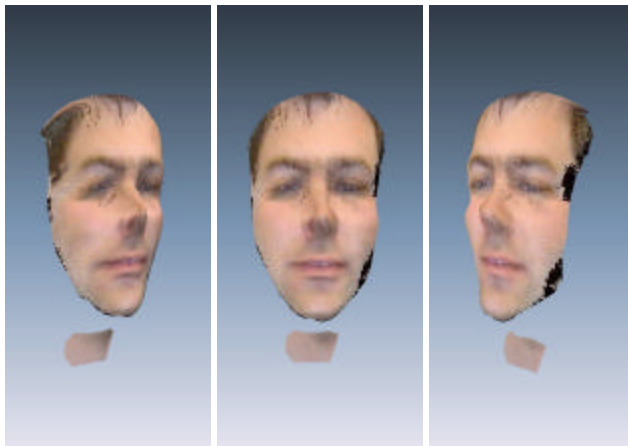


Fig. 1 A range scan of the mannequin's head (a) is shown with the selected landmark points (b), the generic model is reconstructed from this range map (c). Corresponding image points are tracked along the image sequence (d).



(a)



(b)

Fig. 2 Rendering of the reconstructed face, after the warping. Shaded rendering in (a), and with texture map in (b).