Surface Reconstruction of Rotating Objects from Monocular Video

Conference Paper · August 2011

DOI: 10.1007/978-3-642-23687-7_63 · Source: DBLP

2 authors, including:

Abhir Bhalerao
The University of Warwick
103 PUBLICATIONS **1,500** CITATIONS

SEE PROFILE

Surface Reconstruction of Rotating Objects from Monocular Video

Charlotte Boden and Abhir Bhalerao

Department of Computer Science,
University of Warwick, Coventry, CV4 7AL, UK

Abstract. The ability to model 3D objects from monocular video allows for a number of very useful applications, for instance: 3D face recognition, fast prototyping and entertainment. At present there are a number of methods available for 3D modelling from this and similar data. However many of them are either not robust when presented with real world data, or tend to bias their results to a prior model. Here we use energy minimisation of a restricted circular motion model to recover the 3D shape of an object from video of it rotating. The robustness of the algorithm to noise in the data and deviations from the assumed motion is tested and a 3D model of a real polystyrene head is created.

Keywords: Structure from Motion, 3D Modelling, Face Modelling, Turntable Sequence Reconstruction

1 Introduction

The ability to model the shape of a 3D object from monocular video would allow for many useful applications to be developed, for example: 3D face recognition, fast prototyping for industry and entertainment, and transmission of exact 3D information over the internet for video conferencing or reproduction.

As much work has been completed on this and related areas such as multi-view stereo, with great success, one might suppose that this is a solved problem. However in practical situations a robust unbiased solution is often difficult to obtain.

The method proposed here is practically applicable and does not overbias the reconstructed shape towards a strong prior. Points are tracked from frame to frame and their 3D positions found by minimising a cost function based on the assumption that the object being modelled is rotating about an axis perpendicular to the optical axis. The internal camera parameters are known a priori. The shape of the object is not constrained.

The many modelling techniques which exist already will be discussed in section 2. In section 3 the method proposed here involving minimising a cost function obtained from a circular motion model will be described. This method will be used to create models from synthetic and real data of rotating heads and the reconstructions obtained will be compared with ground truth in section 4.

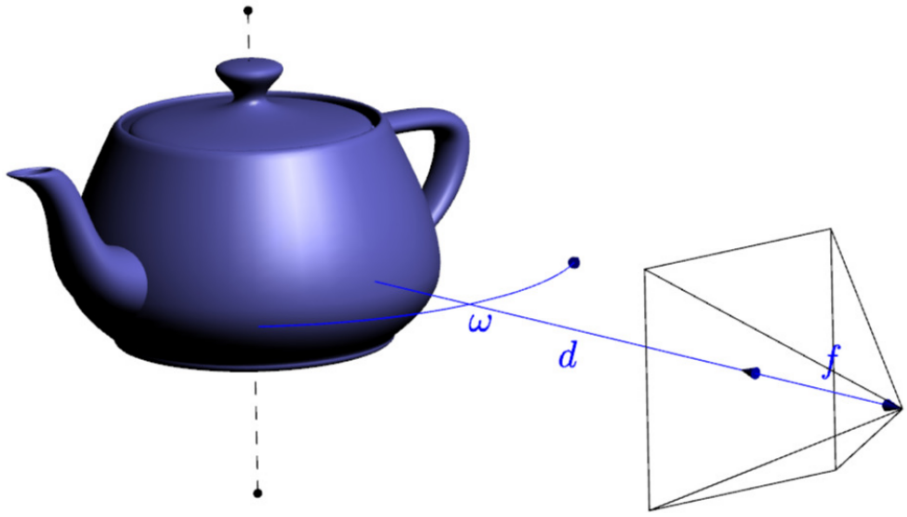

Fig. 1. A camera, with focal length f, is placed a distance d away from the axis of rotation of an object that is rotating at angular velocity ω.

2 Previous Work

There are many ways of reconstructing a surface from images. Feature points or textured patches can be tracked from frame to frame and their motions used to find the 3D structure of the object using such methods as [8], [6] or [10]. These methods use autocalibration techniques and so very little has to be known a priori. However such techniques can produce unpredictable results due to 'critical motions', which are very common in practice, for which there are multiple possible solutions. It is possible to avoid this issue by using prior information either about the structure of the object, as in the case of 3D Morphable Models [1] and appearance models [3], about the motion that is being undertaken [4, 5], about the cameras internal parameters, or about both the motion and internal parameters of the camera, which is similar to multi-view stereo [9]. For point based methods, once a good initial estimate has been found the solution can be refined by minimising the reprojection error, i.e. bundle adjustment (see [11]), by such means as sparse Levenberg-Marquardt minimisation.

In some cases, for instance when the object to be modelled is untextured, it is difficult to track points. In such cases an alternative approach may be to find the silhouette of the object in the various images and deduce what 3D shape could have produced such an outline [12] [5]. One such method [5], which requires the object to be rotating and all the internal parameters of the camera apart from the focal length to be known, can produce excellent results. However if only a limited number of views are available, or if it is difficult to obtain a silhouette, then the results will be more limited. This may arise if the object being modelled is a head turning, in which case there will be a limited number of views, or if

the object is a component part of another object, in which case it may be hard to obtain a silhouette.

Another related approach is space-carving. This method fills space and then carves away voxels until it is possible for the remaining shape to produce images which are consistent with those input (by reprojection). In this way a 'photo-hull' is produced. This technique can be very successful when the motion of the object is known.

When it is easily possible to control the lighting environment, shape from shading can also be used to find the shape of the object. This can work well but is not always applicable if the lighting is not easy to control, for instance for outdoor cloudy scenes, or where there are multiple light sources and reflections.

This work uses a circular motion model such as that used by [12] for silhouettes and by [4] for point tracking. The method is based on point tracking and bears some similarity to the works of Fitzgibbon et al. [4]. However ours uses internally precalibrated cameras and uses all frames and all points rather than merging triplets of frames. We assume that the object is rotating about an axis which is perpendicular to the optical axis (see fig. 1). This obviously is not useful for reconstructing objects from archive footage or from surveillance videos, but is easily achievable for cooperative subjects. It is also assumed that a video camera is used so that there is necessarily continuity of motion between frames. As we are primarily concerned with the surface reconstruction problem rather than the camera calibration problem, these assumptions are not unreasonable and allow for a more accurate solution to be found without over biasing the result.

3 Method

The camera is placed on the z axis at a distance d from the origin, at which it points. The intrinsic parameters are known, the focal length being f and aspect ratio a (see fig. 1). The object to be reconstructed is placed at the origin and allowed to rotate about the Y axis. A point on the object, $\mathbf{X_i} = (X_i, Y_i, Z_i, 1)^T$, will therefore be projected onto a point in the image, $\mathbf{x_i} = (x_i, y_i, 1)^T$, as follows:

$$\mathbf{x_i} = \begin{pmatrix} f & 0 & 0 \\ 0 & af & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & -d \end{pmatrix} \begin{pmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{pmatrix}, \tag{1}$$

in other words:

$$x_i = \frac{fX_i}{d - Z_i} \tag{2}$$

$$y_i = \frac{afY_i}{d - Z_i}, \tag{3}$$

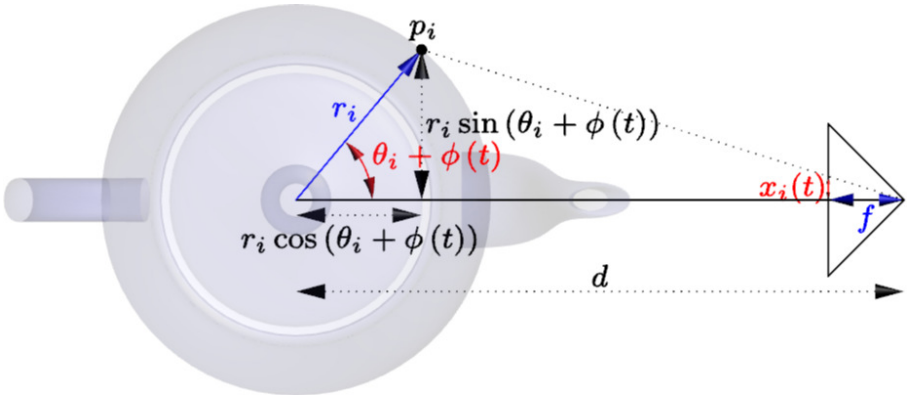

Fig. 2. At time t the point p_i lies on a rotating rigid object at a distance r_i from the centre of rotation at angle $\theta_i + \phi(t)$. The point is imaged by a camera a distance d away with focal length f.

or in cylindrical polar coordinates:

$$x_i(t) = \frac{f r_i \sin(\theta_i(t))}{d - r_i \cos(\theta_i(t))} \tag{4}$$

$$y_i(t) = \frac{a f h_i}{d - r_i \cos(\theta_i(t))}. \tag{5}$$

The object is constrained to rotate about the Y axis. Therefore r is constant and θ is a function of time. As the object is rotating rigidly the change in θ from frame to frame will be the same for all points. So $\theta_i(t)$ can be expressed as $\theta_i + \phi(t)$ (see fig. 2). Therefore:

$$x_i(t) = \frac{f r_i \sin(\theta_i + \phi(t))}{d - r_i \cos(\theta_i + \phi(t))} \tag{6}$$

which gives

$$\frac{dx_i}{dt} = \frac{d\phi}{dt}\left(\frac{x_i(t)}{\tan(\theta_i + \phi(t))} - \frac{x_i(t)^2}{f}\right). \tag{7}$$

This expression has a singularity at $\theta_i = t = 0$, however it can be rearranged to give:

$$x_i(t) = \left(\frac{dx_i}{dt}\right)\left(\frac{d\phi}{dt}\right)^{-1}\left(\frac{f \tan(\theta_i + \phi(t))}{f - x_i \tan(\theta_i + \phi(t))}\right) \tag{8}$$

The following energy function can then be formed:

$$E = \sum_i \sum_j \left(x_i(j) - \left(\frac{dx_i}{dt}\right)_{t=j}\left(\frac{d\phi}{dt}\right)^{-1}_{t=j}\left(\frac{f \tan(\theta_i + \phi(j))}{f - x_i(j)\tan(\theta_i + \phi(j))}\right)\right)^2, \tag{9}$$

which can then be minimised with respect to θ_i and ϕ to give the most likely solution.

In practice points are seeded at random positions in the foreground of an initial frame. Points are tracked backwards and forwards using pyramid based normalised cross correlation. The points are normalised such that the principal point is at zero. Each θ_i is then initialised at $\theta_i = \frac{\pi}{8}$ if x_i is greater than 0, or at $\theta_i = \frac{-\pi}{8}$ otherwise. The θ_i and $\frac{d\phi}{dt}$ values are then adjusted using the MATLAB implementation of the interior-point algorithm [2] until the cost function is minimised. Point positions, the focal length and the distance to the camera must be input to the minimisation procedure. Once θ_i has been found an estimate of $r_i(j)$ and $h_i(j)$ can then be calculated at each frame as follows:

$$r_i(j) = \frac{x_i(j)\, d}{x_i(j)\cos(\theta_i + \phi(j)) + f\sin(\theta_i + \phi(j))} \tag{10}$$

$$h_i(j) = \frac{y_i(j)\, d - y_i(j)\, r_i \cos(\theta_i + \phi(j))}{af} \tag{11}$$

The final estimate of r_i and h_i is the median over all frames of these estimates. These can then be converted into cartesian coordinates if desired.

4 Experimental Results

In order to judge the effectiveness of this algorithm, ground truth data were generated from the vertices of a laser-scanned head obtained from the Basel face database [7]. The positions of the projections of these vertices onto a virtual camera were then used as input to the algorithm. To test the robustness of the algorithm to noise in the data further tests were completed where Gaussian noise was added to the projected positions. The error was then computed as the mean Euclidean distance between the reconstructed points and their corresponding ground truth positions. A graph showing these errors is shown in figure 4.

Another likely problem when applying this technique to real data is that the axis is not perpendicular to the optical axis. To test the robustness of the algorithm to this scenario, sequences where the Basel data were tilted towards the camera were generated and the error computed. The results for these tests are shown in figure 3.

The algorithm was used to reconstruct a model from real data of a polystyrene head ('Poly') on a turntable, see figure 6 for sequence and 7 for the reconstruction. The video was shot with a standard consumer video camera, a Sony DCR-SR90, with a resolution of 640×480 pixels. Five hundred points were tracked. The location of these points in five frames in the middle of the sequence were then used as the input to the reconstruction algorithm (so the profile and close to profile views shown in fig. 7 were not used to make the reconstruction).

Fig. 3 shows that the reconstruction error increases increasingly for increasing tilt angle. The error at 10° being 3.8 times the error at 0°. This shows that the reconstruction does degrade with increasing tilt angle as expected, however for

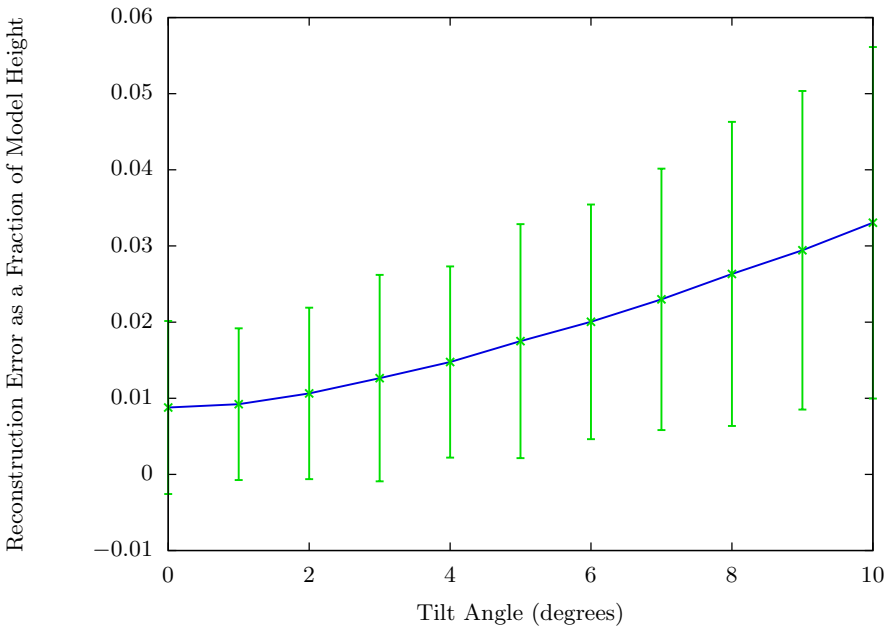

Fig. 3. Vertices of a 3D face model were projected into a virtual camera and their projections used to reconstruct the 3D positions of the vertices. Various angular deviations of the rotation axis from the plane perpendicular to the optical axis were used. Here the error in the reconstruction is plotted against these tilt angles.

small tilt angles the error is not great and so the reconstruction may still be useable, depending on the desired application.

Fig. 4 shows that the error generally increases as the noise level increases, the error at a noise level of 2 pixels being ≈ 7.3 times the error for zero noise. For a model with height 100 pixels the reconstruction error at a plausible track error of 0.5 pixels standard deviations is approximately 2 pixels. Here Gaussian noise was added to every point independently at each frame. However in practice errors are usually non-random and so these values are perhaps not an ideal indication of the visual result and do not show how the effect of noise may vary depending on track length, speed of motion, or other characteristics.

In fig. 5 the reconstruction of ground truth tracks generated from the model face from the 'Basel' database is shown. The spheres represent the reconstructed points and the mesh is the input mesh. This shows that for good data the reconstruction algorithm is very accurate.

The reconstruction of 'Poly' shown in fig. 7 shows that this technique performs promisingly on real data. A limited number of close to frontal frames were used as input and the reconstruction has the correct shape and looks convincing from the profile view. However the reconstruction is quite noisy and so from

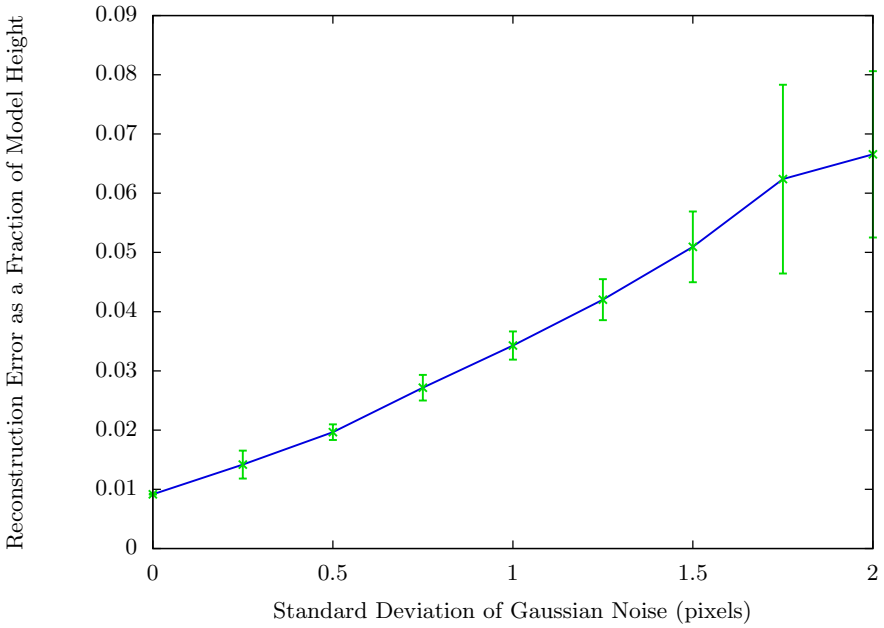

Fig. 4. Perfect data were generated by projecting vertices of a face model onto a virtual camera. Gaussian noise of various standard deviations was then added to these projections. Here the error in the reconstructed locations of the vertices (which is the mean Euclidean distance between reconstructed and ground truth positions) is plotted against the standard deviation of this noise.

certain angles looks quite poor. In further work we will attempt to remedy this by introducting a smoothing term to the cost function.

5 Conclusions

Here a method was devised for generating 3D models of rotating objects from video sequences. It was assumed that either the cameras were internally calibrated or that the height and width of the object being modelled was known. The motion of the object was restricted to be a rotation about an axis perpendicular to the optical axis. The sensitivity of the algorithm to noise in the data and tilts of the axis towards the camera was tested. Reconstructions were made of synthetic ground truth tracks from a model head and of real video data of a polystyrene head rotating (the 'Poly' sequence).

It has been found that the performance of the algorithm degrades reasonably slowly when the rotation axis tilts away from its assumed position. Therefore for situations where the motion can be controlled the method will produce reasonable results even if it is difficult to align the axis exactly. However it will not be

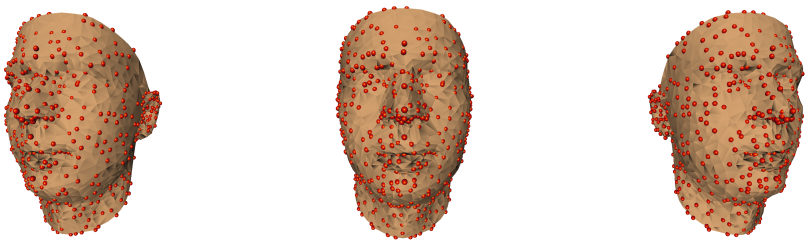

Fig. 5. The surface shown is the ground truth face from the 'Basel' database [7]. The points are the vertices of the reconstructed model. See http://www.dcs.warwick.ac.uk/∼cboden/reconstruction.html for a video of the entire sequence.

Fig. 6. 'Poly' sequence: Video sequence of a model head, 'Poly', rotating on a turntable. 200 frames were taken with a Sony DCR-SR90 camcorder at a resolution of 640×480 pixels. Feature points on these images were tracked and used to reconstruct the surface (see fig. 7).

of use for reconstructing points which rotate about an arbitrary axis relative to the camera.

The method performs well at modelling rotating objects when there is little or no noise in the tracks. When there is a greater level of noise the reconstruction is adversely affected. This could be remedied by including a smoothing term in the energy function, for instance a shape prior or surface constraints. Further work will involve attempting to make such improvements in a manner that does not introduce too much bias to the result.

The reconstruction of 'Poly' highlights the above. A good reconstruction was obtained from five frames. However the obtained model was slightly noisy and so improvements could be made by using a smoothing term. Reconstructions of real heads will be created in future work.

References

1. Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive*

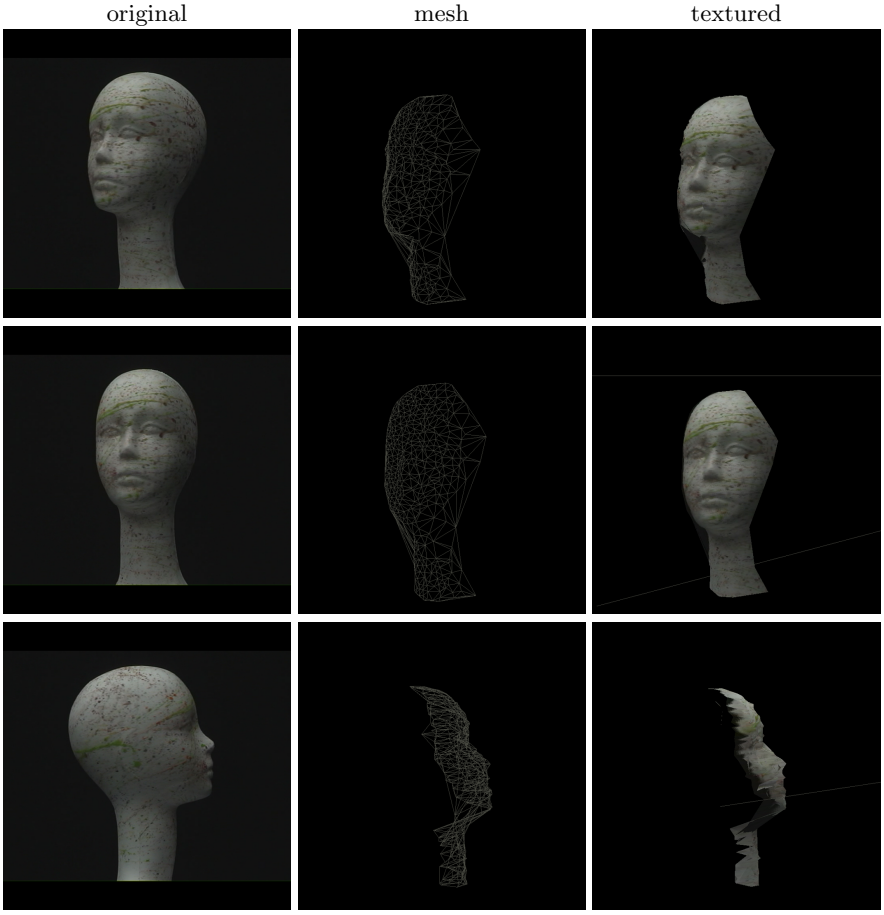

Fig. 7. Reconstruction of Poly: The first column shows the original video sequence; the second shows the reconstructed mesh (created by Delaunay Triangulation from the middle view); and the third a textured version of the reconstructed mesh. The middle row is a frame which was used in the reconstruction, the other two rows show frames which are far from those used in the reconstruction. See http://www.dcs.warwick.ac.uk/∼cboden/reconstruction.html for video of entire sequence.

techniques, SIGGRAPH '99, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

2. Richard H. Byrd, Jean Charles Gilbert, and Jorge Nocedal. A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming*, 89:149–185, 2000. 10.1007/PL00011391.
3. Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23:681–685, June 2001.
4. Andrew W. Fitzgibbon, Geoff Cross, and Andrew Zisserman. Automatic 3d model construction for turn-table sequences. In *Proceedings of the European Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, SMILE'98, pages 155–170, London, UK, 1998. Springer-Verlag.
5. C. Hernandez, F. Schmitt, and R. Cipolla. Silhouette coherence for camera calibration under circular motion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(2):343 –349, 2007.
6. David Nister. Untwisting a projective reconstruction. *International Journal of Computer Vision*, 60:165–183, 2004. 10.1023/B:VISI.0000029667.76852.a1.
7. P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments*, Genova, Italy, 2009. IEEE.
8. Marc Pollefeys, Reinhard Koch, and Luc Van Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. *International Journal of Computer Vision*, 32:7–25, 1999. 10.1023/A:1008109111715.
9. Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:519–528, 2006.
10. B. Triggs. Autocalibration and the absolute quadric. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:609, 1997.
11. Bill Triggs, Philip Mclauchlan, Richard Hartley, and Andrew Fitzgibbon. Bundle adjustment a modern synthesis. In *Vision Algorithms: Theory and Practice, LNCS*, pages 298–375. Springer Verlag, 2000.
12. K-Y. K. Wong and R. Cipolla. Reconstruction of sculpture from its profiles with unknown camera positions. *IEEE Transactions On Image Processing*, 13:381–389, 2004.