Recovering Euclidean Deformable Models from Stereo-motion

Xavier Lladó
IIiA, University of Girona
Girona, Spain
llado@eia.udg.edu

Alessio Del Bue
Instituto Superior Técnico
Lisboa, Portugal
adb@isr.ist.utl.pt

Lourdes Agapito
Queen Mary, University of London
London, U.K
lourdes@dcs.qmul.ac.uk

Abstract

In this paper we present a novel Structure from Motion (SfM) approach able to infer 3D deformable models from uncalibrated stereo images. Using a stereo setup dramatically improves the 3D model estimation when the observed 3D shape is mostly deforming without undergoing strong rigid motion. Our approach first calibrates the stereo system automatically and then computes a single metric rigid structure for each frame. Afterwards, these 3D shapes are aligned to a reference view using a RANSAC method in order to compute the mean shape of the object and to select the subset of points on the object which have remained rigid throughout the sequence without deforming. The selected rigid points are then used to compute frame-wise shape registration and to extract the motion parameters robustly from frame to frame. Finally, all this information is used in a global optimization stage with bundle adjustment which allows to refine the frame-wise initial solution and also to recover the non-rigid 3D model. We show results on synthetic and real data that prove the performance of the proposed method even when there is no rigid motion in the original sequence.

1. Introduction

The recovery of deformable 3D structure from uncalibrated image sequences is still a complex problem. Recently, several SfM factorization approaches have been able to estimate non-rigid 3D models in the case of a deforming shape viewed by affine cameras [1, 6, 9] and full perspective cameras [3, 7, 10]. The main constraint of these monocular SfM approaches is that a reliable model can only be extracted if the motion performed by the observed 3D shape has a strong rigid component. In the deformable case, this constraint is even more critical since deformations must be properly disambiguated from the motion component given by the imaging device (i.e. perspective distortion and camera motion). Using a stereo rig is a straightforward solution which may overcome this limitation and improve the 3D estimation when the shape exhibits weak rigid motion. The problem of recovering 3D structure using a stereo-rig moving in time or a stereo rig looking at a moving object has been defined for the rigid case as the stereo-motion problem [8]. Ho and Chung [5] were the first to formulate this problem within the factorization scenario. Recently, a stereo-motion approach with deformable shapes was successfully used for the affine camera case [2]. However, a method which deals with the full perspective case has not yet been proposed.

In this paper we present a novel approach for the 3D Euclidean reconstruction of deformable objects observed by an uncalibrated stereo rig. In a first step, the stereo system is automatically calibrated and used to compute the metric rigid shape from each pair of stereo views. Adopting the assumption that some of the object points remain rigid over the sequence [3, 7], we register all the 3D shapes to a reference view using a RANSAC algorithm in order to compute the mean shape of the object and also to select the set of rigid points. These selected rigid points are then used to compute frame-wise registration and to extract the motion parameters robustly. All this information – stereo camera parameters, mean shape, and motion between frames – is then used to initialise a non-linear optimization stage. This bundle adjustment (BA) step allows to refine the initial solution and also to recover the non-rigid 3D model of the deformable object. We present different synthetic experiments in order to evaluate the performance of our approach when using different ratios of rigid/non-rigid points in the object, different degrees of deformation, and different rigid motion in the sequence. Experimental results when using real data from a human face performing different facial expressions are also presented.

2. Non-rigid factorization: single camera

Assuming a perspective projection camera model a 3D point $\mathbf{X}_j$ is projected onto an image frame i according to $\mathbf{x}_{ij} = \frac{1}{\lambda_{ij}} \mathtt{P}_i \mathbf{X}_j$ where $\mathbf{x}_{ij}$ and $\mathbf{X}_{ij}$ are both expressed in homogeneous coordinates, $\mathtt{P}_i$ is the projection camera matrix and λ_{ij} is the projective depth for that point. The projection matrix may be parameterized as $\mathtt{P}_i = \mathtt{K}_i[\mathtt{R}_i|\mathbf{T}_i]$ where $\mathtt{K}_i$ is the calibration matrix, $\mathtt{R}_i$ the rotation matrix and $\mathbf{T}_i$ the translation vector. When an object is deforming, the non-rigid 3D structure can be approximated by a linear combination of a set of D basis shapes $\mathtt{B}_d$ which represent

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

the principal modes of deformation of the object [1]. The non-rigid 3D points at each frame i are expressed in homogeneous coordinates as:

$$\mathtt{X}_i = \begin{bmatrix} \sum_{d=1}^{D} l_{id}\mathtt{B}_d \\ \mathbf{1} \end{bmatrix} \qquad \mathtt{X}_i \in \Re^{4\times N} \quad \mathtt{B}_d \in \Re^{3\times N} \quad (1)$$

where $\mathtt{B}_d$ are the $3 \times N$ basis shapes (N is the number of points), l_{id} are the linear deformation coefficients and $\mathbf{1}$ is a N-vector of ones. The projection of a 3D point j at any frame i onto the image plane is then governed by:

$$\mathbf{x}_{ij} = \mathtt{P}_i \mathtt{X}_{ij} = \mathtt{P}_i \begin{bmatrix} \sum_{d=1}^{D} l_{id}\mathbf{B}_{dj} \\ \mathbf{1} \end{bmatrix} \qquad (2)$$

where each 3-vector $\mathbf{B}_{dj}$ is given such that $\mathtt{B}_d = [\mathbf{B}_{1j} \ldots \mathbf{B}_{Dj}]$.

3. Our non-rigid stereo factorization approach

As mentioned in the introduction, the main problem of SfM methods is the requirement of a sufficient overall rigid motion in order to correctly estimate the reconstruction parameters. Note that in real situations this may not be possible. For instance in a human face performing different facial expressions, the undergoing rigid motion — mainly rotation — is usually very small. Moreover, in the full perspective camera case, the perspective distortion may be wrongly considered as deformations (and viceversa). Aiming to solve this problem, we propose a novel approach for recovering non-rigid models from a stereo rig, where the two cameras remain fixed relative to each other throughout the sequence. This stereo case requires not only the temporal tracks of points in the left and right image sequences but also the stereo correspondences between left and right image pairs. In this paper the correspondence issue is not tackled, assuming that the complete stereo measurements are correctly matched and available.

In the first step, our stereo system is automatically calibrated, computing the fundamental matrices from each pair of views and using the Kruppa equations to recover the intrinsic camera parameters $\mathtt{K}_i$ (focal lengths) [4]. Since the relative orientation and position between the left and right cameras is fixed, we have expressed the rotation and translation of the right camera in terms of the relative rotation $\mathtt{R}_{rel}$ and translation $\mathbf{T}_{rel}$. Exploiting the relationship between the fundamental matrix and the essential matrix, both $\mathtt{R}_{rel}$ and $\mathbf{T}_{rel}$ are recovered [4]. Once the calibration is obtained, we then compute the metric rigid shape for each frame by applying triangulation. It is important to remark that one could not apply epipolar geometry at each single camera to recover the frame-wise motion (i.e. rotation and translation) since the points on the structure are varying with time and therefore violating the epipolar constraints.

3.1. Frame-wise motion estimation

In order to solve for the motion between frame to frame we adopt the reasonable assumption that some of the object points remain rigid over the sequence. Our idea behind this assumption is twofold. Firstly, to use a RANSAC algorithm which considers non-rigid points as outliers in order to register all the shapes to a reference view. This way we are able to compute the mean shape over the sequence which will be then used as initialization of the first basis shape $\mathtt{B}_1$ of our non-rigid model. Secondly, to select a set of rigid points from the 3D shapes which will do the frame-wise motion estimation more robust.

The procedure to select a set of rigid points from all the shapes works as follows. Once the shapes are aligned to a reference frame, we perform a segmentation between rigid and non-rigid points analyzing the 3D registration errors obtained per point. Since the structure of deforming parts varies from frame to frame, the mean registration error of these deforming points will be much larger than the one of the rigid points. Thus a set of rigid points can be easily distinguished from the obtained registration errors. A similar strategy to perform a point deformation detection from 3D views has been recently proposed by Wang et al. [7]. Notice that in this step we are not looking for a perfect segmentation among all rigid and non-rigid points. Our goal is only to select a good set of rigid points for helping the frame-wise motion estimation. Once the rigid points have been selected, we used them to compute the frame-wise registration and to robustly extract the motion parameters.

3.2. Estimating the non-rigid model

In order to estimate the complete 3D non-rigid shape model we minimize the geometric distance between the measured image points and the estimated reprojected points $\sum_{i,j} \parallel \mathbf{x}_{ij} - \hat{\mathbf{x}}_{ij} \parallel^2$. Therefore, our cost function being minimized is:

$$\min_{\mathtt{K}_i\mathtt{R}_i\mathbf{T}_i\mathtt{R}_{rel}\mathbf{T}_{rel}\mathbf{B}_d l_{id}} \sum_{i,j} \parallel \mathbf{x}_{ij}^{L} - \mathtt{K}_i[\mathtt{R}_i|\mathbf{T}_i]\begin{bmatrix} \sum_{d=1}^{D} l_{id}\mathbf{B}_{dj} \\ 1 \end{bmatrix} \parallel^2$$

$$+ \parallel \mathbf{x}_{ij}^{R} - \mathtt{K}_i[\mathtt{R}_{rel}\mathtt{R}_i|\mathbf{T}_i + \mathbf{T}_{rel}] \begin{bmatrix} \sum_{d=1}^{D} l_{id}\mathbf{B}_{dj} \\ 1 \end{bmatrix} \parallel^2 \quad (3)$$

The goal of this minimization is to refine and correctly estimate the left and right camera matrices, the intrinsic camera parameters $\mathtt{K}_i$, the configuration weights l_{id} and the basis-shapes $\mathbf{B}_{dj}$ such that the distance between the measured image points $\mathbf{x}_{ij}^{L}$ and $\mathbf{x}_{ij}^{R}$ and the estimated image points $\hat{\mathbf{x}}_{ij}^{L}$ and $\hat{\mathbf{x}}_{ij}^{R}$ is minimized. This minimization is accomplished with a bundle adjustment step that uses as initialization the estimated parameters of the geometry of the stereo rig $\mathtt{K}_i$, $\mathtt{R}_{rel}$ and $\mathbf{T}_{rel}$, the estimated frame-wise motion $\mathtt{R}_i$ and $\mathbf{T}_i$, and the obtained mean shape $\mathtt{B}_1$. The remaining

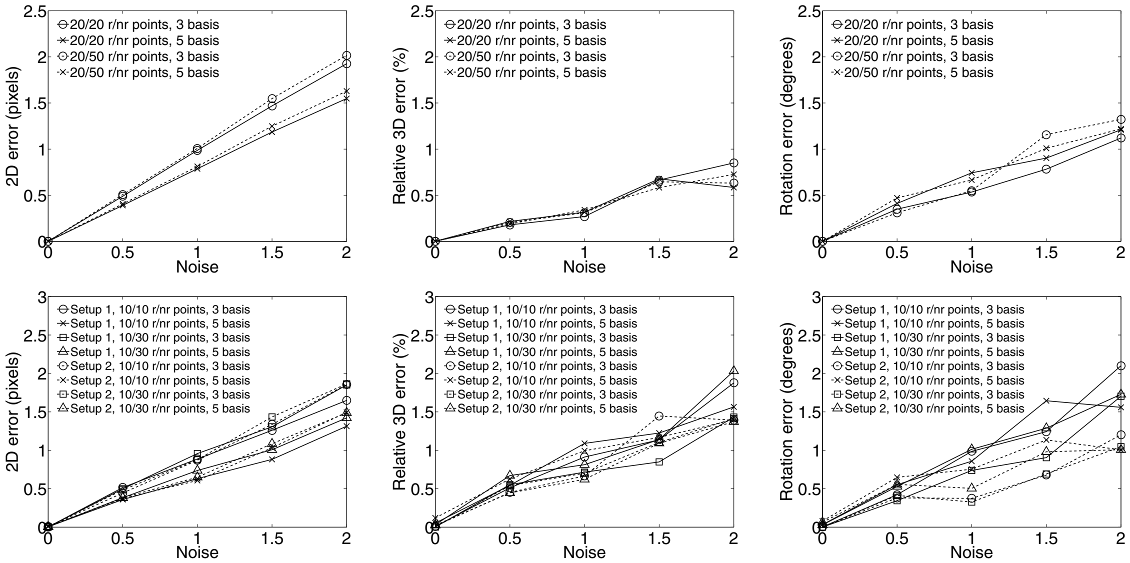

Figure 1. 2D, 3D and rotation error curves. First row: results when not rigid motion was applied. Second row: results when the object was deforming while doing a rigid motion transformation.

basis shapes $\mathtt{B}_d$ which encode the $(D-1)$ non-rigid components are initialized to small random values. Finally, the deformation weights l_{i1} associated with the mean shape are initialized to 1 while the rest are initialized to small values. A similar initialization has previously been used in [3, 6].

4. Experimental results

4.1. Synthetic data

The synthetic 3D data consisted of a set of random points sampled inside a cube of size $50 \times 50 \times 50$ units. In order to evaluate our method we used two different types of image sequences: the first one in which the object was not performing any rigid motion, and the second one where the object was deforming and rigidly moving at the same time. For both situations, several sequences were generated using different ratios of rigid points (which included the vertices of the cube) and non-rigid points. Different deformations for the non-rigid points were generated using random basis shapes and random deformation weights. We also created different sequences varying the number of basis shapes ($D = 3$ and $D = 5$) for the different ratios of rigid/non-rigid points. We consider cameras with zero skew, unit aspect ratio, and known principal points. Finally, gaussian noise of increasing levels of variance was added to the image coordinates.

4.1.1 Deforming object without rigid motion

For this particular experiment we used a fixed set of 20 rigid points while using 20 and 50 non-rigid points generated using 3 and 5 different basis shapes. The 3D data was then projected onto 20 pairs of views using a perspective camera model and without applying any rotations and translations to the object. The distance of the object to the cameras was z=100 and the focal length was fixed to be f=500. We then applied our 3D reconstruction algorithm to all the experimental configurations described before. The results are summarized on the first row of Figure 1 where we show the r.m.s. 2D image reprojection error (pixels), 3D metric reconstruction error (percentage relative to the scene size) and the absolute rotation error (degrees). The plots show the mean values of 5 different random trials per level of noise. Our approach appears to perform well in the presence of noise. The 3D reconstruction error is low even for a large proportion of non-rigid versus rigid points. The sequences had also large perspective distortions due to the chosen camera setup. Figure 1 also illustrates that the rotations are correctly estimated. Reliable estimates for the internal camera parameters (focal length, relative camera rotation and translation) were also obtained even in the presence of noise.

4.1.2 Deforming object undergoing rigid motion

For this experiment, the 3D data was also projected onto 20 pairs of views using a perspective camera model but now applying random rotations and translations over all the axes. We used here a set of 10 rigid points while using 10 and 30 non-rigid points. In order to evaluate different levels of perspective distortion, we used 2 different camera setups in which we varied the distance of the object to the cameras and the focal length (Setup1: z=80,f=400; Setup2: z=100,f=500). The obtained results are summarized on the second row of Figure 1. Observe, that our proposed algorithm performed well even when using a minimal set of rigid points. Regarding the algorithm convergence, the nonlinear optimization step for all these experiments usually converged within around 30 iterations. Note also that the algorithm always converges in the absence of noise.

4.2. Experiments with real data

In this experiment we use real 3D data of a human face performing different facial expressions. The 3D data was captured using a VICON motion capture system by tracking a subject wearing 37 markers on the face. First row of Figure 2 shows three key-frames showing the positions of the markers and the range of deformations of some expressions in the tested sequence. The 3D points were then projected synthetically onto a stereo image sequence 22 frames long using a perspective camera model and fixing the relative rotation and translation of the stereo pair. The size of the face model was $169 \times 193 \times 102$ units and the stereo camera setup was such that the subject was at a distance of 150 units from the cameras and the focal length was 300 pixels so the perspective effects were significant. As in the synthetic experiments we applied our method when the object was not performing any rigid motion, and when the object was rotating and translating during the sequence. For both cases – and without introducing noise – our algorithm converged to small errors. When introducing Gaussian noise of 2 pixels and for the case in which the face was also rotating and translating, the obtained 2D reprojection error was 1.44 pixels, the absolute 3D error was 2.51 units, the absolute rotation error was 1.17 degrees, while the estimated focal length was 310.54. The number of basis shapes was fixed to $D = 5$. Figure 2 shows the ground truth (squares) and reconstructed shapes (crosses) from front and side views of frames 1, 14 and 22. The selected set of rigid points obtained using the RANSAC algorithm is highlighted in the frontal view of the first frame. Notice that these rigid points are situated mainly on the nose and on the temples of the face. Interestingly, the deformations are very well captured by the model even for the frames in which the facial expressions are more exaggerated.

5. Conclusions

We have proposed a new approach for the 3D Euclidean reconstruction of deformable objects observed by an uncalibrated stereo rig. The experimental results on synthetic and real data have proven the performance of our proposal even when there is no rigid motion in the original sequence and with a minimal set of rigid points.

Acknowledgments

This work has been supported by MEC grant DPI2007-66796-C03-02, EPSRC grant GR/S61539/01, and by FCT through the POS_Conhecimento Program and grant PTDC/EEA-ACR/72201/2006.

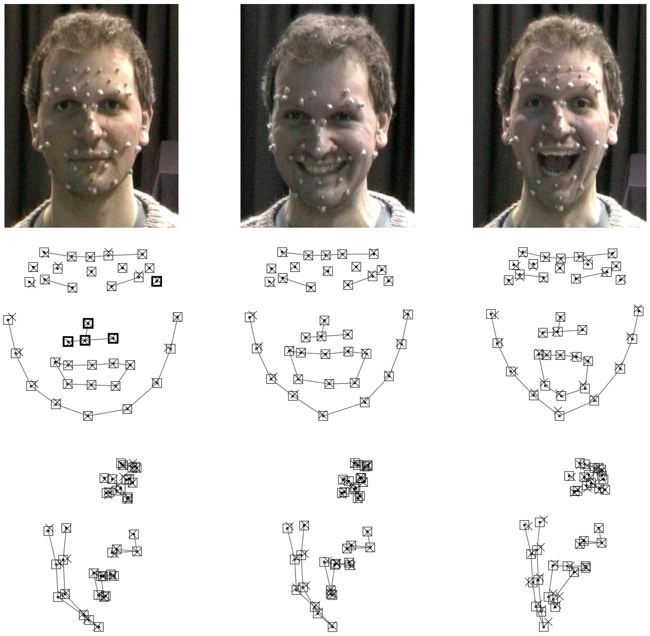

Figure 2. Front and side views for noise = 2. Reconstructions for frames 1, 14 and 22.

References

[1] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *Proc. IEEE CVPR*, pages 690–696, June 2000.

[2] A. Del Bue and L. Agapito. Stereo non-rigid factorization. *IJCV*, 66(2):193–207, February 2006.

[3] A. Del Bue, X. Lladó, and L. Agapito. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In *Proc. IEEE CVPR*, pages 1191–1198, June 2006.

[4] O. Faugeras and Q. Luong. *The Geometry of Multiple Images*. The MIT Press, Cambridge, Massachusetts, 2001.

[5] P. K. Ho and R. Chung. Stereo-motion that complements stereo and motion analysis. In *Proc. IEEE CVPR*, pages 213–218, 1997.

[6] L. Torresani, D. Yang, E. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *Proc. IEEE CVPR*, pages 493–500, December 2001.

[7] G. Wang and Q. M. J. Wu. Stratification approach for 3-d euclidean reconstruction of nonrigid objects from uncalibrated image sequences. *IEEE Trans. Syst., Man, Cybern.*, 38(1):90–101, February 2008.

[8] A. Waxman and J. Duncan. Binocular image flows: steps toward stereo-motion fusion. *PAMI*, 8(6):715–729, 1986.

[9] J. Xiao, J. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. In *Proc. ECCV*, pages 573–587, May 2004.

[10] J. Xiao and T. Kanade. Uncalibrated perspective reconstruction of deformable structures. In *Proc. IEEE ICCV*, pages 1075–1082, October 2005.