Pattern Recognition Letters 32 (2011) 1020–1028

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Reconstruction of non-rigid 3D shapes from stereo-motion

Xavier Lladó [a,*], Alessio Del Bue [b], Arnau Oliver [a], Joaquim Salvi [a], Lourdes Agapito [c]

[a] Institute of Informatics and Applications, University of Girona, Campus de Montilivi, 17071 Girona, Spain
[b] Istituto italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy
[c] School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK

ARTICLE INFO

Article history:
Received 11 May 2010
Available online 16 February 2011
Communicated by Y. Liu

Keywords:
Structure from motion
Stereo-motion
Non-rigid structure

ABSTRACT

Several non-rigid structure from motion methods have been proposed so far in order to recover both the motion and the non-rigid structure of an object. However, these monocular algorithms fail to give reliable 3D shape estimates when the overall rigid motion of the sequence is small. Aiming to overcome this limitation, in this paper we propose a novel approach for the 3D Euclidean reconstruction of deformable objects observed by an uncalibrated stereo rig. Using a stereo setup drastically improves the 3D model estimation when the observed 3D shape is mostly deforming without undergoing strong rigid motion. Our approach is based on the following steps. Firstly, the stereo system is automatically calibrated and used to compute metric rigid structures from pairs of views. 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 which have remained rigid throughout the sequence. The selected rigid points are then used to compute frame-wise shape registration and to robustly extract the motion parameters from frame to frame. Finally, all this information is used as initial estimates of a non-linear optimization which allows us to refine the initial solution and also to recover the non-rigid 3D model. Exhaustive results on synthetic and real data prove the performance of our proposal estimating motion, non-rigid models and stereo camera parameters even when there is no rigid motion in the original sequence.

1. Introduction

Most biological objects and natural scenes vary their shape under the effect of external and internal forces, i.e. human bodies may articulate and deform, sheets of paper and clothes may bend under pressure and strains, living cells may change abruptly their topology during biological interactions. During the last years several works have presented extensions of the Tomasi and Kanade's structure from motion (SfM) algorithm to deal with the reconstruction of non-rigid objects (Bregler et al., 2000; Brand, 2001; Xiao et al., 2004; Torresani et al., 2008; Paladini et al., 2009). These methods are based on the fact that any configuration of the shape can be explained as a linear combination of basis shapes that define the principal modes of deformation of the object. Moreover, almost all these non-rigid methods assume the case of images acquired under weak perspective viewing conditions, useful when the relief of the object is small compared to the distance to the object. In this work we are interested in the case when the images are acquired at closer distances, with a wide field of view or when the scene is large in space and perspective distortions appear. In this situation, the assumption of weak perspective projection no longer holds and these methods provide distorted reconstructions. Different monocular approaches have already been proposed to deal with the reconstruction of 3D deformable shapes under the full perspective camera case (Xiao and Kanade, 2005; Vidal and Abretske, 2006; Hartley and Vidal, 2008). However, the main constraint of all these SfM approaches is that a reliable model can only be extracted if the image sequence includes a large rotation component. In the deformable case, this constraint is even more critical since deformations have to 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. Nowadays, the two-view stereo cameras start to be available in the market and the development of image analysis and 3D reconstruction methods for these systems is growing fast with applications such as image-based modeling, human–computer interaction and vision-based control. Not to mention the recent revamp of stereo professional cameras in the film industry which are already affecting a novel set of consumer products (Ronfard and Taubin, 2010). The problem of recovering 3D

* Corresponding author. Tel.: +34 972 418878; fax: +34 972 418259.
E-mail addresses: llado@eia.udg.edu (X. Lladó), alessio.delbue@iit.it (A. Del Bue), aoliver@eia.udg.edu (A. Oliver), qsalvi@eia.udg.edu (J. Salvi), lourdes@dcs.qmul.ac.uk (L. Agapito).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.02.010

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 (Waxman and Duncan, 1986; Stein and Shashua, 1998; Dornaika and Sappa, 2009). As stated by Ho and Chung (1997) there are two visual cues that have to be taken into account when formulating the stere-motion reconstruction problem: the motion cue, in which 3D structure is recovered from the relative motion between the scene and the camera, and the stereo vision cue, in which 3D structure is recovered from the stereo pair images of the same scene. Ho and Chung (1997) were the first to formulate this stere-motion problem within the rigid SfM scenario, presenting a framework that combined the advantages of both cues to provide more accurate reconstructions. Recently, a stereo approach with deformable shapes was successfully used for the affine camera case (Del Bue et al., 2006). This SfM method imposes the extra constraints that arise from the fact that both stereo cameras are viewing the same non-rigid 3D structure. However, a method which deals with the full perspective camera case has not yet been proposed.

With the aim to overcome the well-known intrinsic limitation of the monocular SfM methods of having enough overall rigid motion in the sequence, which can occur frequently (e.g. modeling the deformations of a human face), we present in this paper a projective approach for the 3D Euclidean reconstruction of deformable objects observed by a stereo rig. We show that the use of two cameras allows us to recover 3D non-rigid models even when the overall rigid motion of the sequence is small.

The main idea of our proposal is the following. Our approach first calibrates automatically the stereo system and it computes the metric 3D rigid structure from every pair of views (i.e. using two images). Afterwards, the algorithm computes a rigid registration of all the 3D shapes to a reference view using a RANSAC algorithm (Fischler and Bolles, 1987). This step requires the assumption that some of the object points remain rigid over the sequence (Del Bue et al., 2006; Lladó et al., 2010; Wang and Wu, 2008). Given this registration, we can compute the mean shape of the object and also to select a set of rigid points which will be employed to perform a frame-wise registration and to robustly extract the motion parameters from frame to frame. Note that epipolar geometry can not be directly applied from frame to frame due to the fact that the object is deforming at each frame. Instead of computing the deformable metric model from the independent rigid shapes obtained for each stereo pair, we propose in this work a non-linear optimization process which allows us to integrate both motion and shape constraints from the spatial and temporal acquisitions (i.e. structure and deformation coefficients are shared by left and right cameras). All the extracted information – stereo camera parameters, mean shape, and motion between frames – is used as initial estimates of the non-linear optimization where the objective function to be minimized is the image reprojection error. This bundle adjustment (BA) step allows us to refine the initial solution and also to recover the non-rigid 3D model of the deformable object which benefits from the integration of spatial and temporal stereo acquisitions. We present different synthetic experiments in order to evaluate the behavior 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 using real data are also presented. The rest of the paper is organized as follows. In Section 2 we introduce the background on non-rigid perspective SfM methods, presenting the non-rigid model obtained by these methods. Section 3 describes our proposal (introduced in Lladó et al., 2008) for the non-rigid metric shape and motion recovery from stereo sequences. In Section 4 we show experiments on different synthetic and real data sets which validate our approach. Finally, conclusions are given in Section 5.

2. Background on monocular non-rigid SfM

The key idea of the monocular structure from motion methods is to gather the 2D image coordinates of a set of P points tracked throughout F frames into a measurement matrix $\mathtt{W}_{2F\times P}$. Assuming affine viewing conditions, the measurement matrix can be expressed analytically as a bilinear product of two matrices: $\mathtt{W} = \mathtt{M}\,\mathtt{S}$ where $\mathtt{M}$ is a $2F \times 3$ motion matrix which expresses the pose of the camera and $\mathtt{S}$ is the $3 \times P$ shape matrix which contains 3D locations of the reconstructed scene points. Therefore, the rank of the centered measurement matrix – where the translation is removed – is constrained to be $r \leqslant 3$. Exploiting this rank constraint and enforcing metric constraints on the rotation matrices one can recover the motion and the 3D shape (Tomasi and Kanade, 1992).

Using a perspective camera model, a 3D point $\overline{\boldsymbol{X}}_j$[1] is projected onto image frame i according to the equation $\bar{\boldsymbol{w}}_{ij} = \frac{1}{\lambda_{ij}}\mathtt{P}_i\overline{\boldsymbol{X}}_{ij}$, where $\bar{\boldsymbol{w}}_{ij}$ and $\overline{\boldsymbol{X}}_{ij}$ are both expressed in homogeneous coordinates (i.e. $\bar{\boldsymbol{w}}_{ij} = [u_{ij}\,v_{ij}1]^T = \left[\boldsymbol{w}_{ij}^T 1\right]^T$ and $\bar{\boldsymbol{X}}_j = [X_j Y_j Z_j 1]^T$), $\mathtt{P}_i$ is the 3×4 projection matrix and λ_{ij} is the projective depth for that point. The projection camera is defined mathematically as $\mathtt{P}_i = \mathtt{K}_i[\mathtt{R}_i|\boldsymbol{T}_i]$, where the 3×3 rotation matrix $\mathtt{R}_i$ and the translation vector $\boldsymbol{T}_i$ represent the Euclidean transformation between the camera and the world coordinate system respectively and $\mathtt{K}_i$ is a 3×3 upper triangular matrix which contains the intrinsic camera parameters. Scaling the image coordinates of all the points in all the views by their corresponding projective depth gives a $3F \times P$ rescaled measurement matrix $\bar{\mathtt{W}} = \mathtt{MS}$, where $\mathtt{S} = [\bar{\boldsymbol{X}}_1, \ldots, \bar{\boldsymbol{X}}_P]$ is a $4 \times P$ shape matrix which contains the homogeneous coordinates of the P 3D rigid points and $\mathtt{M}$ is a $3F \times 4$ matrix which contains the perspective cameras for each frame (Sturm and Triggs, 1996). In the case of rigid structure and assuming the projective depths λ_{ij} are known $\mathtt{M}$ and $\mathtt{S}$ are at most rank *4*. Therefore, the rank of the scaled measurement matrix $\bar{\mathtt{W}}$ is constrained to be $r \leqslant 4$.

When an object is deforming, its 3D structure changes from frame to frame where $\bar{\mathtt{X}}_i = [\bar{\boldsymbol{X}}_{i1}, \ldots, \bar{\boldsymbol{X}}_{iP}]$ is a $(4 \times P)$ matrix representing the shape at frame i in homogeneous coordinates. In order to express the deformations of the 3D shape in a compact way, Bregler et al. (2000) introduced a simple linear model where the 3D shape of any specific configuration is approximated by a linear combination of a set of D basis shapes $\mathtt{B}_d$ with $d = 1, \ldots, D$, which represent the principal modes of deformation of the object. In the projective case the 3D vectors are expressed in homogeneous coordinates and so the shape may be written (Xiao and Kanade, 2005) as:

$$\bar{\mathtt{X}}_i = \begin{bmatrix} \sum_{d=1}^{D} l_{id}\mathtt{B}_d \\ \mathbf{1}^T \end{bmatrix} \quad \bar{\mathtt{X}}_i \in \Re^{4\times P} \quad \mathtt{B}_d \in \Re^{3\times P}, \tag{1}$$

where $\mathtt{B}_d$ are the $3 \times P$ basis shapes, l_{id} are the corresponding deformation coefficients and $\mathbf{1}$ is a P-vector of ones. Note that the first basis shape corresponds to the mean shape of the 3D model.

Xiao and Kanade (2005) were the first proposing a two step SfM algorithm for reconstruction of 3-D deformable shapes under the full perspective camera model. From this initial work several approaches such as the works of Vidal and Abretske (2006), Del Bue et al. (2006), Wang et al. (2007), Wang and Wu (2008), Bartoli et al. (2008) and Hartley and Vidal (2008) have also been proposed. All these projective SfM methods are able to recover non-rigid shape models from a single video camera. However, all of them share the well-known SfM constraint of having enough rigid motion during the sequence in order to provide reliable 3D deformable models.

[1] Vectors are represented in bold font.

3. Our non-rigid stereo approach

In real situations the monocular SfM requirement of having a sufficient overall rigid motion 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. In this stereo case all the measurement 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 (see Fig. 1). In this paper the correspondence issue is not tackled, assuming that the complete stereo measurements are correctly matched and available. Hence, we adopt the assumption that all scene elements are visible from both cameras and we do not deal with the problems associated with occlusions and missing data. However, we will show in the experimental results a real sequence where the measurement matrix is automatically obtained by using a stereo tracker algorithm (Ferrer and Garcia, 2008). It is important to mention that a stereo approach requires both cameras to be synchronized. However, if this were not the case, it could be elegantly solved using the solution proposed by Tresadern and Reid (2009) for the synchronization of stereo video sequences in an uncalibrated scenario.

The use of a calibrated stereo pair is a common and practical solution to obtain reliable 3D reconstructions. In its simpler formulation, once the stereo rig is calibrated, the depth of points in the image is estimated by applying triangulation (Hartley and Zisserman, 2000). In order to obtain accurate depth estimates, the cameras are usually separated from each other by a significant baseline thus creating widely spaced observations of the same object. The disadvantage of this configuration though, is that having a wide baseline makes the matching of features between pairs of views a more challenging problem (Delponte et al., 2006). Note that in this situation features often have a very different appearance between views or are not even visible in both cameras, thus making spatial correspondence difficult.

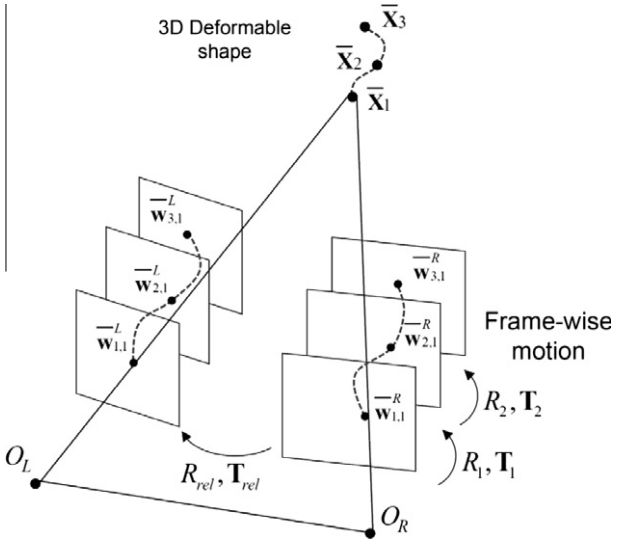

Fig. 1. Stereo-motion setup. A point is moving in space and its position in 3D is shown for each time instance as $\overline{\mathbf{X}}_1$, $\overline{\mathbf{X}}_2$, and $\overline{\mathbf{X}}_3$. The three points are then projected onto the respective image frames obtaining the image coordinates $\bar{\mathbf{w}}^L_{1,1}$, $\bar{\mathbf{w}}^L_{2,1}$ and $\bar{\mathbf{w}}^L_{3,1}$ for the left camera and $\bar{\mathbf{w}}^R_{1,1}$, $\bar{\mathbf{w}}^R_{2,1}$ and $\bar{\mathbf{w}}^R_{3,1}$ for the right one. The dotted lines connecting the points represent the 2D trajectory in time of the point in the left and right images. Since the position of the cameras is fixed, the relative orientation $\mathbf{R}_{rel}$ and camera displacement $\boldsymbol{T}_{rel}$ between the camera centers O_L and O_R are considered constants in time.

On the other hand, the task of computing temporal tracks from the single camera sequences is relatively easier since the images are closely spaced in time. As a drawback, disparities between consecutive frames may be insufficient to obtain a reliable depth estimation and, as a result, longer sequences are needed to infer the 3D structure. Particularly, in the case of non-rigid structure, a sufficient overall rigid motion is necessary to allow the monocular algorithms to correctly estimate the reconstruction parameters. Hence, a question of relevant interest is the feasibility of an approach that efficiently fuses the positive aspects of both methods. In this sense, Del Bue et al. (2006) have recently proposed a stereo-motion approach able to recover deformable shapes for the affine camera case.

Affine cameras are only an approximation of the real viewing conditions affecting the projection of a body onto the image plane. Therefore, these models are generally effective when the relief of the object is small compared to the distance from the camera center. When these assumptions weaken, for instance in the case when the images are acquired at closer distances, the use of a perspective camera model is necessary to obtain a correct 3D reconstruction of the object. Our stereo approach presented in the following sections deals with the recovery of deformable shapes for the projective camera case, aiming also to overcome the limitation of having a sufficiently large overall rigid motion for the monocular SfM methods.

3.1. Obtaining estimates from the stereo rig

In the first step, our stereo system is automatically calibrated using the captured data. This is done computing the fundamental matrices from each pair of views and using the Kruppa equations to recover the intrinsic camera parameters $\mathbb{K}^L$ and $\mathbb{K}^R$ (focal lengths for both left and right cameras) (Faugeras and Luong, 2001). 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 $\mathbb{R}_{rel}$ and translation $\boldsymbol{T}_{rel}$ (see Fig. 1). Exploiting the relationship between the fundamental matrix and the essential matrix, both $\mathbb{R}_{rel}$ and $\boldsymbol{T}_{rel}$ are automatically recovered (Faugeras and Luong, 2001). Once the calibration of the stereo system is obtained, we then compute the metric rigid shape for each pair of views by applying triangulation (Hartley and Zisserman, 2000). Note that one fundamental matrix yields two Kruppa equations; therefore we are able to recover two intrinsic parameters assuming that the rest are known. In this work, we consider cameras which have zero skew, unit aspect ratio, and known principal points for doing the automatic calibration. This automatic calibration step could be avoided if the stereo calibration is already known by using a separate calibration procedure based on standard techniques such as the ones described in (Armangué et al., 2002). This will be shown in the experimental section with a real stereo sequence where the full stereo system calibration was available beforehand. The key issue that is important to remark is 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 non-rigidly with time and therefore violating the epipolar constraints.

3.2. 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. This assumption was already introduced in (Del Bue et al., 2006; Wang et al., 2007). The idea

behind this assumption is twofold. Firstly, to use a RANSAC algorithm (Fischler and Bolles, 1987) 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 mean shape (first basis shape $\mathtt{B}_1$) of the non-rigid model. Secondly, to select a set of rigid points from the 3D shapes in order to compute the frame-wise motion estimation more robustly.

The procedure to align the 3D shapes to a reference frame is based on a RANSAC process over Horns' absolute orientation algorithm (Horn, 1987). During each RANSAC iteration, we randomly draw pairs of corresponding 3D points and use these sets to hypothesize the transformation matrix. The estimation of the transformation is achieved up to a certain input probability (in our implementation fixed to 0.99) of having an outlier-free set containing only rigid points. In our RANSAC implementation the minimum theoretical number of points was 3, although, in all our experimental tests bigger sets were selected. Once all the shapes are aligned, we perform the segmentation between rigid and non-rigid points analyzing the 3D registration errors obtained for all the points in each frame. The registration error is defined here as the Euclidean distance between each pair of point correspondences. Since the structure of deforming parts varies from frame to frame, the accumulated registration error of these deforming points across the sequence 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 by using a threshold. This can be seen in Fig. 2 where the accumulated 3D registration error obtained for a synthetic sequence that contains 30 rigid and 50 non-rigid points respectively is shown. Note that for different levels of image noise (e.g. noise = 0.5 and 1) one can clearly separate the set of rigid (first 30 points) from the non-rigid ones (the remaining 50 points). We automatically determine the segmentation threshold using the well-known Otsu's method (Otsu, 1979) which is able to select the best threshold value minimizing the intra-class variance of the registration error distribution. This thresholding technique has provided good performances as we will see in Section 4. One could also fix this threshold experimentally being even more strict on the decision and therefore avoiding possible misclassifications of deformable points into rigid ones. One should notice also that the ability to separate rigid and non-rigid points may be affected by the ratio of rigid and non-rigid points, noise level and degree of non-rigidity of the object. These issues will be analyzed in the experimental section.

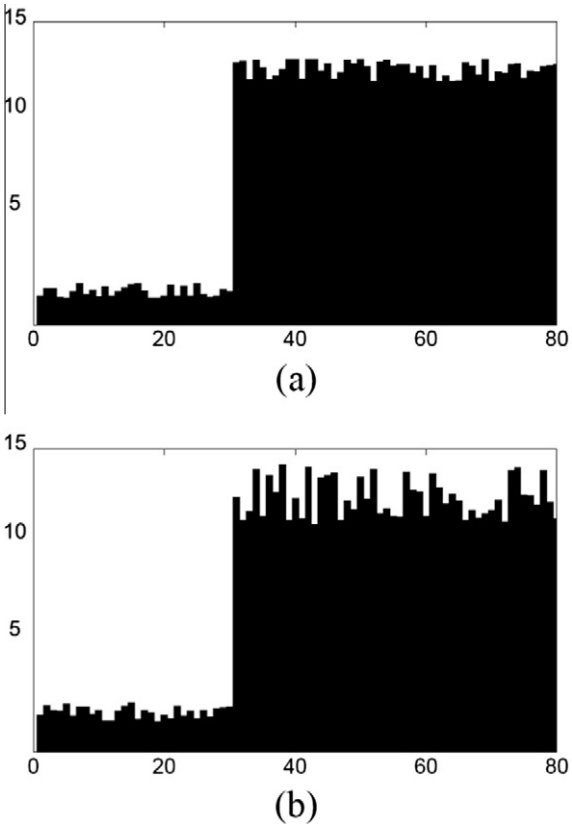

Fig. 2. Two examples of the accumulated 3D registration error obtained for every point across a synthetic stereo sequence. The sequence contains 30 and 50 rigid and non-rigid points respectively. (a) Image noise level = 0.5 and (b) image noise level = 1. For simplicity in the visualization we ordered the points so that the rigid are the first ones, showing that we are able to classify these points.

A similar strategy to perform a point deformation detection from 3D views has been recently proposed by Wang and Wu (2008). Moreover, Del Bue et al. (2007a) also proposed a method to detect and segment the rigid points from the non-rigid ones from the 2D measurements. Instead of using the 3D shapes, their approach is based on the fact that rigid points will satisfy the epipolar geometry while the non-rigid points will give a high residual in the estimation of the fundamental matrix between pairs of views. They use a RANSAC algorithm to estimate the fundamental matrices from pairwise frames in the sequence and to segment the scene into rigid and non-rigid points (Del Bue et al., 2007a). However, as argued by Wang and Wu (2008), addressing the problem from the 3D geometrical information and accumulating the errors of the deformations parts frame by frame may provide more accurate results. Nevertheless, it is important to remark that in this step we are looking for a good set of rigid points which will help the frame-wise motion estimation, even though the complete segmentation among all rigid and non-rigid points may not be achieved. Once the final set of rigid points had been selected, we used them to compute the frame-wise registration via a RANSAC procedure and to robustly extract the frame-wise motion parameters ($\mathtt{R}_i$ and $\boldsymbol{T}_i$).

All this information: mean shape, stereo camera parameters (relative rotation and translation) and the motion between frame to frame, is then used as initial estimates of a non-linear optimization where the objective function to be minimized is the image reprojection error: a geometrically meaningful error function. This step takes the parameters of the geometry of the stereo rig, rigid structure and estimated frame-wise motion into account in order to refine the initial solution and also to recover the non-rigid 3D model of the deformable, integrating both motion and shape constraints from the spatial and temporal 2D image acquisitions.

3.3. Non-rigid 3D model estimation: bundle adjustment

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 $\sum_{ij}\|\bar{\boldsymbol{w}}_{ij}-\hat{\boldsymbol{w}}_{ij}\|^2$. Therefore, our cost function being minimized is

$$\min_{\mathtt{K}^L\mathtt{K}^R\mathtt{R}_i\boldsymbol{T}_i\mathtt{R}_{rel}\boldsymbol{T}_{rel}\boldsymbol{B}_d l_{id}} \sum_{ij} \left\| \bar{\boldsymbol{w}}^L_{ij} - \mathtt{K}^L[\mathtt{R}_i \mid \boldsymbol{T}_i] \begin{bmatrix} \sum_{d=1}^{D} l_{id}\boldsymbol{B}_{dj} \\ 1 \end{bmatrix} \right\|^2 + \left\| \bar{\boldsymbol{w}}^R_{ij} - \mathtt{K}^R[\mathtt{R}_{rel}\mathtt{R}_i \mid \mathtt{R}_{rel}\boldsymbol{T}_i + \boldsymbol{T}_{rel}] \begin{bmatrix} \sum_{d=1}^{D} l_{id}\boldsymbol{B}_{dj} \\ 1 \end{bmatrix} \right\|^2, \quad (2)$$

where, the goal of this minimization is to refine and correctly estimate the left and right camera matrices, the intrinsic camera parameters $\mathtt{K}^L$ and $\mathtt{K}^R$, the configuration weights l_{id} and the basis shapes $\boldsymbol{B}_{dj}$ such that the distance between the measured image

points $\bar{\boldsymbol{w}}_{ij}^{L}$ and $\bar{\boldsymbol{w}}_{ij}^{R}$ and the estimated image points $\hat{\boldsymbol{w}}_{ij}^{L}$ and $\hat{\boldsymbol{w}}_{ij}^{R}$ is minimized.

Following the rigidity priors introduced by Del Bue et al. (2006), we also impose priors on our set of rigid points obtained from the segmentation. If the motion of a point j is completely rigid for the entire sequence, the structure referring to that point is expressed entirely by the mean shape (first basis $d = 1$, and zero value for the non-rigid component). From this it follows that for a rigid point $\boldsymbol{B}_{dj} = \mathbf{0} \quad \forall d > 1$ where $\boldsymbol{B}_j = \left[\boldsymbol{B}_{1j}^T, \ldots, \boldsymbol{B}_{Dj}^T 1\right]^T$. Note that $\boldsymbol{B}_j$ is a $3D + 1$ vector which encodes the D basis shapes for point j and $\boldsymbol{B}_{dj}$ is the 3-vector which contains 3D coordinates of basis shape d for point j. Notice that this rigidity prior forces $3(D-1)$ zeros in each column of the shape matrix corresponding to a rigid point. We write these rigidity constraints as priors on the coordinates of the basis vectors $\boldsymbol{B}_{dj}$. The benefit of including this rigidity prior was previously demonstrated in (Lladó et al., 2010), showing that priors helped to avoid local minima and improved the reconstruction results.

The minimization of Eq. (2) is accomplished with a bundle adjustment step. Bundle adjustment algorithms require careful initialization since they may fail to converge to the global minimum unless they are initialized close enough to it. The non-linear optimization of the cost function was achieved using a Levenberg–Marquadt minimization scheme modified to take advantage of the sparse block structure of the matrices involved in the process (Triggs et al., 2000). We have chosen to parameterize the camera matrices using quaternions. Quaternions ensure that there are no singularities and that the orthonormality of the rotation vectors is preserved. Given the large number of parameters involved in the non-linear minimization, the objective function is highly non-linear and so it is crucial to provide an initial estimate that is sufficiently close to the global minimum. In this sense, our initialization comes from the estimated parameters of the geometry of the stereo rig $\mathbb{K}^L$ and $\mathbb{K}^R$, $\mathbb{R}_{rel}$ and $\boldsymbol{T}_{rel}$, the estimated frame-wise motion $\mathbb{R}_i$ and $\boldsymbol{T}_i$, and the obtained mean shape $\mathbb{B}_1$. The remaining basis shapes $\mathbb{B}_d$ 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. Note that this shape initialization is equivalent to initializing the shape with a strong mean component and small deviations from it to explain the deformations, which is a fair assumption in most deformable objects. This assumption and also this initialization procedure has been seen to produce satisfactory results in the case of affine and perspective cameras as shown in (Torresani et al., 2001; Del Bue et al., 2007b; Lladó et al., 2010). However, we want to emphasize that algorithms based on BA do not guarantee the convergence to the global minima so a good initialization continues to be crucial.

4. Results

This experimental section validates our proposal for 3D non-rigid metric reconstruction with synthetic and real experiments. The synthetic tests are designed in order to verify the performance of the method when using different ratios of rigid/non-rigid points and when using different camera setups. After the synthetic tests, this section presents the performance of our approach in the case of a real deforming object. In particular, we use a real face undergoing different facial expressions. The image measurements for the stereo set are generated from the data acquired by using a VICON motion capture system which also provides the ground truth for comparing the 3D reconstructions. Finally, we present another experiment using the image acquisitions from a real calibrated stereo rig. This experiment allows also to illustrate the improvements of our perspective approach compared to the affine stereo case presented by Del Bue et al. (2006).

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 proposal we used two different setups of image sequences: the first one in which the object was not performing any rigid motion, and the second one in which the object as well as deforming was doing a certain rigid motion transformation. 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 deformation degrees for the non-rigid points were generated using random basis shapes and random deformation weights. The first basis shape had the largest weight equal to 1. We also created different sequences varying the number of basis shapes ($D = 3$ and $D = 5$) for different ratios of rigid/non-rigid points. The 3D data was then projected onto 20 pairs of views using a perspective camera model. Different stereo camera configurations were also used, varying the intrinsic parameters, the relative rotation (between 10° and 30°) and the baseline of the stereo pair. The sequences had also large perspective distortions due to the chosen camera setup. For all these experiments we assumed that the focal lengths of the cameras were unknown but constant, the aspect ratios and the principal points were known and constant, while the skew was set to be 0. 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 assumed that all the camera parameters, including the relative camera orientation and the baseline of the stereo pair remained constant over the sequences.

We then applied our 3D reconstruction algorithm to all the experimental setups described before. The results are summarized on the first row of Fig. 3 where we show the root mean square (RMS) 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 performs well in the presence of noise. The 3D reconstruction error is low even for a large proportion of non-rigid versus rigid points. The 2D error is also small and it appears to be of the same order as the image noise. Fig. 3 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. We want to emphasize that in these experiments the Euclidean non-rigid 3D model was obtained when neither the camera and the object was doing any rigid motion, a situation in which the non-rigid SfM methods presented in Section 2 will fail.

4.1.2. Deforming object with 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 between 10° and 50° and translations over all the axes. We used here 2 sets of 10 and 30 fixed 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 and third rows of Fig. 3 where we show again the RMS 2D image reprojection error,

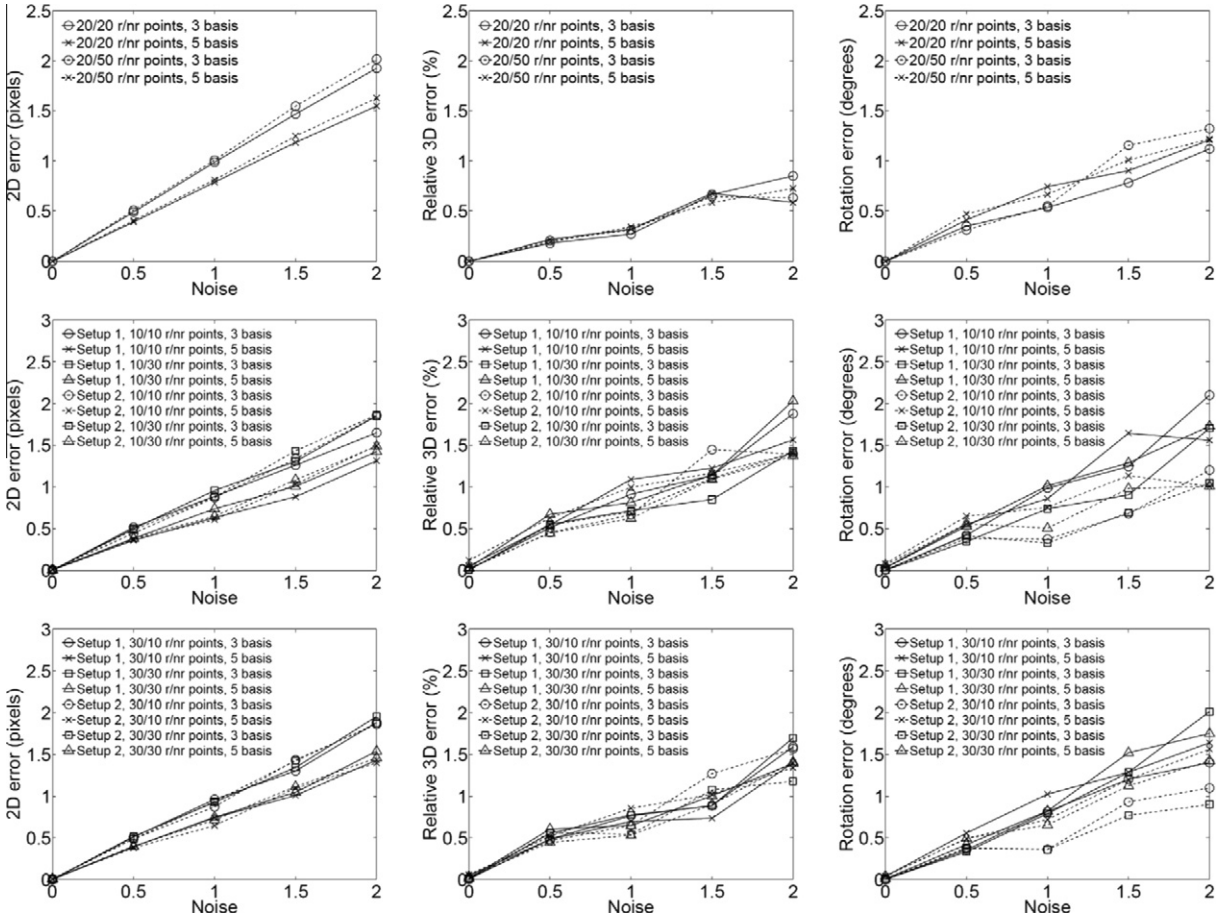

Fig. 3. 2D error, 3D error and rotation error curves. First row: results obtained when not rigid motion was applied to the object. Second and third rows: results obtained when the object was deforming while doing a rigid motion transformation.

3D metric reconstruction error and the absolute rotation error. The plots show the mean values of 5 different random trials per level of noise. Observe, that our proposed algorithm perform well even when using a reduced set of rigid points. The 3D reconstruction error and the rotations are correctly estimated for different proportions of rigid versus non-rigid points. Note also that in general we obtained better estimates (i.e. smaller 3D and rotation errors) in the experiments without any rigid motion (Section 4.1.1). We have also noticed that widely separated stereo views allow a more reliable estimation of motion and deformation parameters.

Regarding to the algorithm convergence, the non-linear optimization step for all these experiments usually converged within around 30 iterations. Moreover, results in Fig. 3 show the algorithm converges in the absence of noise. In this sense, the good initial estimates for the stereo camera setup, the motion and the 3D structure are fundamental to avoid local minima.

It is important to mention that the result of the rigid/non-rigid point segmentation can provide some misclassified points. These outliers may introduce error in the initial rotation and translation estimates. However, as we show in these experiments, after applying bundle adjustment the results are satisfactory, providing acceptable motion and structure estimates. In the following section we analyze and discuss some results about the rigid and non-rigid point segmentation.

4.1.3. Evaluating the segmentation of rigid and non-rigid points

In order to evaluate the 3D rigid and non-rigid point segmentation, we used the same synthetic experimental setup described above. Therefore, we analyzed sequences with 10 and 30 rigid points while using 10, 30 and 50 non-rigid points generated using 3 and 5 different basis shapes. 30 different trials in which the object was undergoing a rigid motion were done per each configuration. Table 1 shows the degree of misclassification (measured as number of misclassified points) for varying ratios of rigid/non-rigid points and for increasing levels of noise. Note that in general a good behaviour is achieved for all the setups, although, the misclassification error was higher when having a large proportion of non-rigid points (i.e. using 10/50 rigid/non-rigid points). This was

Table 1
Mean misclassification error for different levels of noise with variance $\sigma = 0.5, 1, 1.5, 2$ pixels. The experimental setups use different number of bases ($D = 3, 5$) and ratios of rigid (10/30) versus non-rigid points (10/30/50). The mean error is computed over 30 tests for each setup and level of noise.

Experiments	Noise				
	0	0.5	1	1.5	2
$D = 3$, 10/10	0	0	0.1	0.1	0.3
$D = 3$, 10/30	0.6	0.8	0.9	1.1	1.6
$D = 3$, 10/50	1.1	1.4	2.1	2.4	2.6
$D = 3$, 30/10	0	0	0	0	0
$D = 3$, 30/30	0.2	0.4	0.4	0.3	0.6
$D = 3$, 30/50	1.2	1.2	1.8	1.8	3.2
$D = 5$, 10/10	0	0	0	0	0
$D = 5$, 10/30	0	0	0	0	0
$D = 5$, 10/50	0.3	0.4	0.4	0.8	0.8
$D = 5$, 30/10	0	0	0	0	0
$D = 5$, 30/30	0	0	0	0	0.2
$D = 5$, 30/50	0.4	0.4	0.5	0.6	0.5

due to the fact that we had more probability to synthetically generate non-rigid points with very small deformations and therefore more misclassified points. Another interesting point is that we have noticed a better performance in the case of stronger deformations compared to weaker ones due to the fact that the segmentation is less ambiguous (see the examples when using 5 basis shapes). Obviously, a misclassification error could affect the RANSAC estimation of the rotations and translations, and therefore the final recovery of the 3D structure and camera parameters. However, we want to remark that the effect of misclassification in this stage is not significant on the final motion estimates since only a subset of the rigid points is needed in the RANSAC algorithm to compute the frame-wise motion estimates. The effect of misclassified points in the optimization process and in the inclusion of priors was previously studied in Del Bue et al. (2006) showing that a small set of outliers (i.e. 2 misclassified points) was not significant in the final estimates of BA.

4.2. Experiments with reprojected data

In this experiment we use 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 Fig. 4 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 an stereo image sequence 22 frames long using a perspective camera model and fixing the relative rotation and translation of the stereo set. Gaussian noise of 2 pixels was also added to the image coordinates. 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 camera and the focal length was 300 pixels so the perspective effects were considerable. We assumed that the focal lengths of the cameras were unknown but constant, while the aspect ratios and the principal points were known and constant.

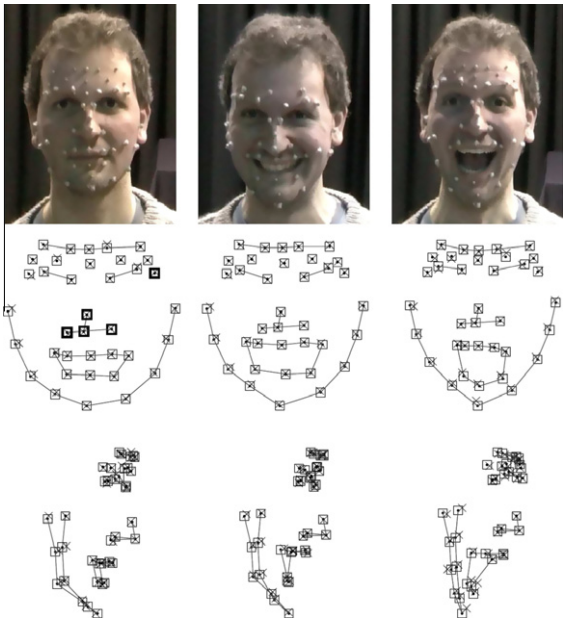

Fig. 4. Front and side views of the reconstructed face for pixel noise = 2. Reconstructions are shown for frames 1, 14 and 22. Crosses are used to indicate our estimated reconstructed points while squares refer to the ground truth. Highlighted marks on the frontal view of frame 1 indicate the selected set of rigid points.

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 in the image coordinates – our algorithm converged to the optimum. 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 after 5 tests with different random noise was 1.54 ± 0.01 pixels, the absolute 3D error was 2.26 ± 0.32 units, the absolute rotation error was $1.12 \pm 0.20^\circ$, while the estimated focal length was 301.55 ± 20.43. The number of basis shapes was fixed to $D = 5$. Fig. 4 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. Note that these rigid points are situated mainly on the nose and the side of the face. Observe that the obtained results are satisfactory since the deformations are very well captured by the model even for the frames in which the facial expressions are more exaggerated.

4.2.1. Comparison with monocular approaches

Using this data set we also performed a comparison with the perspective monocular SfM approaches of Del Bue et al. (2006) and Xiao and Kanade (2005). As expected, for the first experiment in which the face was not performing any rigid motion, it was not possible to obtain any reliable reconstruction with these SfM approaches using only information from one camera. However, when the object was doing a sufficient rigid motion during the sequence, these monocular methods provided satisfactory 3-D models. In particular, for a sequence with 2 pixels noise, the obtained 3D errors were 2.73 and 3.89 units, while the absolute rotation errors were 1.69 and 2.71°, respectively. The former results were obtained with the Del Bue et al. (2006) approach fixing $D = 5$, while the latter results were obtained with the Xiao and Kanade (2005) approach, where the number of independent basis shapes was automatically selected by the method to be $D = 3$.

4.3. Experiments with a real stereo rig

In this experiment we show qualitative results with measurements obtained from a real static stereo system composed by two digital cameras (Canon EOS 50D) properly synchronized and calibrated (Bouguet, 2009). In order to facilitate the image feature tracking and matching, the stereo setup was such that the cameras were looking in the same direction separated only by a baseline of 15 cm. The captured sequence was composed of 10 stereo frames with a resolution of 2352×1568 pixels. The first and the last stereo frames are shown in the first and third row of Fig. 5. Notice that the background of the sequence is composed by three static books while there are two moving objects that move linearly in 2 different directions. These 2 moving objects are treated here as the non-rigid part of the scene. As shown in Fig. 5, 475 tracked features are automatically established across the stereo sequence, where 389 features belong to the static background and 17 and 69 features belong to the two moving objects, respectively. In order to deal with the feature detection, tracking, matching, and triangulation we used the approach of Ferrer and Garcia (2008).

After applying our non-rigid stereo approach into the obtained image measurements, all the rigid points were automatically segmented from the non-rigid ones using our segmentation scheme. Furthermore, the recovered 3D non-rigid metric structure of the scenario was satisfactory as can be seen in Fig. 5(a), (b), (d) and (e) where frontal and top views of the reconstructed 3D scene for the first and last frames are shown. Observing the top views of the reconstructions, one can clearly appreciate the reconstructed planes belonging to the books of the background and the preserved

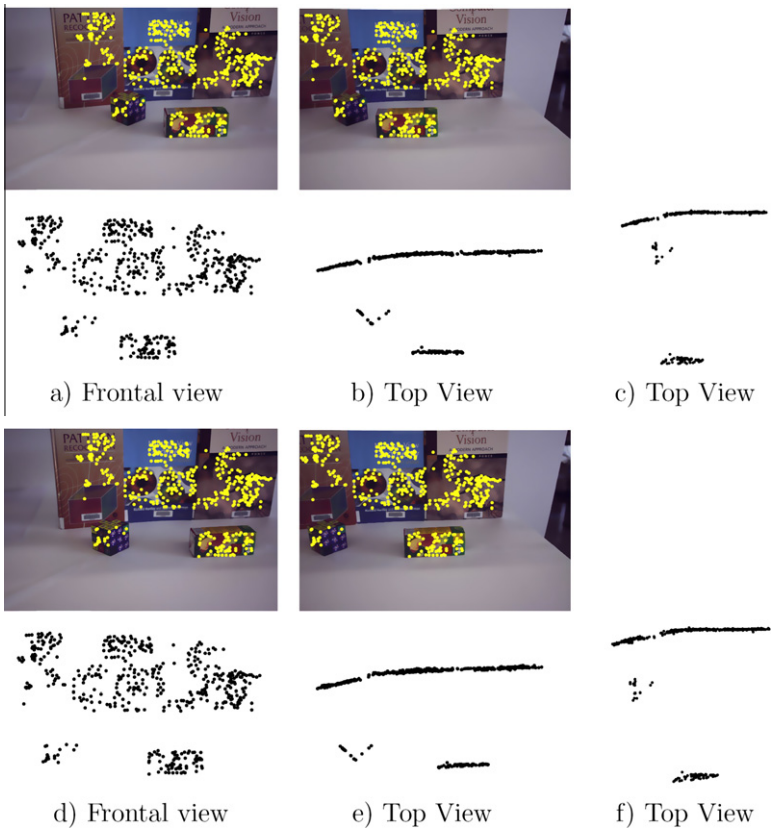

Fig. 5. Experimental results using a real stereo system. First and third row shows the pair of stereo images (left and right respectively) for frames 1 and 10. 2D image measurements are shown in yellow dots. Second and fourth row – images (a), (b), (d) and (e), respectively – shows the frontal and top views of the reconstructed 3D shapes using our projective stereo approach, (c) and (f) show the top views of the obtained 3D shapes when using the affine stereo approach of Del Bue et al. (2006). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

orthogonality of the planes belonging to the box object (left moving object). Moreover, the deformable model of the scene – including the non-rigid structure – is correctly recovered even though the stereo sequence was acquired without frame-wise motion. Finally, we also applied the affine stereo approach proposed by Del Bue et al. (2006) to this testing data where the images were acquired at closer distance. Top views of the reconstructed scene for the first and last frames are shown in Fig. 5(c) and (f). Observe that even though the main 3D structure of the scene is recovered, it contains distortions. Note that the orthogonality of the planes belonging to the box and also the distance between objects is better reconstructed using our perspective approach.

5. Conclusion

A novel approach for the estimation of 3D Euclidean non-rigid models observed by an uncalibrated stereo rig has been proposed. Our approach computes first metric rigid structures from pairs of views by using the stereo system. The obtained 3D shapes are then used to compute the mean shape of the object and to select a subset of rigid points which are used to compute frame-wise shape registration and to extract the motion parameters robustly from frame to frame. Finally, given all the initial estimates, the problem of recovering the non-rigid 3D shape is formalized as a non-linear optimization which benefits from the integration of spatial and temporal stereo acquisitions.

The experimental results on synthetic and real data have proven the performance of our proposal estimating motion, non-rigid 3D models and stereo camera parameters even when there is no rigid motion in the original sequence and with a small set of rigid points. Using a stereo setup drastically improves the 3D model estimation when the observed 3D shape is mostly deforming without undergoing strong rigid motion.

The main assumptions of our method are that the deformable object should contain a small set of points remaining rigid over the sequence (as in Del Bue et al., 2006 and Wang and Wu, 2008), and that cameras must be synchronized and stereo matches be available. In this sense, nowadays it is common to obtain synchronized video from stereo cameras while stereo matching with deformable objects may be better tackled by extending current techniques to deal with the non-rigid case.

Acknowledgements

This work has been supported by Spanish MEC project DPI2007-66796-C03-02, EPSRC, Grant No. GR/S61539/01. The authors thank J. Ferrer and Dr. R. Garcia for sharing the code of their stereo tracker system, and Dr. J. Xiao and Prof. T. Kanade for sharing their non-rigid SfM code.

References

Armangué, X., Salvi, J., Batlle, J., 2002. A comparative review of camera calibrating methods with accuracy evaluation. Pattern Recognition 35, 1617–1635.

Bartoli, A., Gay-Bellile, V., Castellani, U., Peyras, J., Olsen, S., Sayd, P., 2008. Coarse-to-fine low-rank structure-from-motion. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 1–8..

Bouguet, J., 2009. Camera calibration toolbox for matlab. <http://www.vision.caltech.edu/bouguetj/calib_doc/index.html>.

Brand, M., 2001. Morphable models from video. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Kauai, Hawaii, pp. 456–463.

Bregler, C., Hertzmann, A., Biermann, H., 2000. Recovering non-rigid 3d shape from image streams. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Hilton Head, South Carolina, pp. 690–696.

Del Bue, A., Agapito, L., 2006. Stereo non-rigid factorization. Internat. J. Comput. Vision 66 (2), 193–207.
Del Bue, A., Lladó, X., Agapito, L., 2006. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, New York, pp. 1191–1198.
Del Bue, A., Lladó, X., Agapito, L., 2007a. Segmentation of rigid motion from non-rigid 2d trajectories. In: Iberian Conf. on Pattern Recognition and Image Analysis, LNCS, 4477, 491–498.
Del Bue, A., Smeraldi, F., Agapito, L., 2007b. Non-rigid structure from motion using ranklet–based tracking and non-linear optimization. Image Vision Comput. 25 (3), 297–310.
Delponte, E., Isgrò, F., Odone, F., Verri, A., 2006. Svd-matching using sift features. Graph. Models 68 (5), 415–431.
Dornaika, F., Sappa, A., 2009. A featureless and stochastic approach to on-board stereo vision system pose. Image Vision Comput. 27 (9), 1382–1393.
Faugeras, O., Luong, Q., 2001. The Geometry of Multiple Images. The MIT Press, Cambridge, Massachusetts.
Ferrer, J., Garcia, R., 2008. Optical seafloor mapping, Research report, M.Sc. Thesis, University of Girona.
Fischler, M.A., Bolles, R.C., 1987. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In: Fischler, M.A., Firschein, O., (Eds.), Readings in Computer Vision: Issues, Problems, Principles, and Paradigms, Los Altos, CA, pp. 726–740.
Hartley, R., Vidal, R., 2008. Perspective nonrigid shape and motion recovery. In: Proc. 8th European Conf. on Computer Vision, Prague, Czech Republic.
Hartley, R.I., Zisserman, A., 2000. Multiple View Geometry in Computer Vision. Cambridge University Press.
Ho, P.K., Chung, R., 1997. Stereo-motion that complements stereo and motion analysis. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Puerto Rico, pp. 213–218.
Horn, B.K.P., 1987. Closed-form solution of absolute orientation using unit quaternions. J. Opt. Soc. Am. A 4 (4), 629–642.
Lladó, X., Del Bue, A., Agapito, L., 2008. Recovering euclidean deformable models from stereo-motion. In: Internat. Conf. on Pattern Recognition, Tampa.
Lladó, X., Del Bue, A., Agapito, L., 2010. Non-rigid metric reconstruction from perspective cameras. Image Vision Comput. 28 (9), 1339–1353.
Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Systems Man Cybernet. 2 (1), 62–66.
Paladini, M., Bue, A.D., Stosic, M., Dodig, M., Xavier, J., Agapito, L., 2009. Factorization for non-rigid and articulated structure using metric projections. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2898–2905.
Ronfard, R., Taubin, G., 2010. Image and geometry processing for 3-d cinematography: An introduction. In: Ronfard, R., Taubin, G. (Eds.), Image and Geometry Processing for 3-D Cinematography, Geometry and Computing, vol. 5. Springer, Berlin, Heidelberg, pp. 1–8.
Stein, G., Shashua, A., 1998. Direct estimation of motion and extended scene structure from a moving stereo rig. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Santa Barbara, pp. 211–218.
Sturm, P., Triggs, B., 1996. A factorization based algorithm for multi-image projective structure and motion. In: Proc. 4th European Conf. on Computer Vision, Cambridge, pp. 709–720.
Tomasi, C., Kanade, T., 1992. Shape and motion from image streams under orthography: A factorization approach. Internat. J. Comput. Vision 9 (2), 137–154.
Torresani, L., Yang, D., Alexander, E., Bregler, C., 2001. Tracking and modeling non-rigid objects with rank constraints. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Kauai, Hawaii, pp. 493–500.
Torresani, L., Hertzmann, A., Bregler, C., 2008. Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. IEEE Trans. Pattern Anal. Machine Intell. 30 (5), 878–892.
Tresadern, P.A., Reid, I.D., 2009. Video synchronization from human motion using rank constraints. Comput. Vis. Image Und. 113 (8), 891–906.
Triggs, B., McLauchlan, P., Hartley, R.I., Fitzgibbon, A., 2000. Bundle adjustment – A modern synthesis. In: Triggs, W., Zisserman, A., Szeliski, R. (Eds.), Vision Algorithms: Theory and Practice, LNCS. Springer-Verlag, pp. 298–375.
Vidal, R., Abretske, D., 2006. Nonrigid shape and motion from multiple perspective views. In: Proc. European Conf. on Computer Vision, LNCS 3952, 205–218.
Wang, G., Wu, Q.M.J., 2008. Stratification approach for 3-d euclidean reconstruction of nonrigid objects from uncalibrated image sequences. IEEE Trans. Systems Man Cybernet. 38 (1), 90–101.
Wang, G., Tsui, H., Hu, Z., 2007. Structure and motion of nonrigid object under perspective projection. Pattern Recognition Lett. 28 (4), 507–515.
Waxman, A., Duncan, J., 1986. Binocular image flows: Steps toward stereo-motion fusion. IEEE Trans. Pattern Anal. Machine Intell. 8 (6), 715–729.
Xiao, J., Kanade, T., 2005. Uncalibrated perspective reconstruction of deformable structures. In: Proc. 10th Internat. Conf. on Computer Vision, Beijing, China, pp. 1075–1082.
Xiao, J., Chai, J., Kanade, T., 2004. A closed-form solution to non-rigid shape and motion recovery. In: Proc. 4th European Conf. on Computer Vision, Cambridge, pp. 573–587.