Non-rigid structure from motion using non-parametric tracking and non-linear optimization

Alessio Del Bue Fabrizio Smeraldi Lourdes Agapito

Department of Computer Science
Queen Mary, University of London
London, E1 4NS, U.K.
{alessio,fabri,lourdes}@dcs.qmul.ac.uk

Abstract

In this paper we address the problem of estimating the 3D structure and motion of a deformable non-rigid object from a sequence of uncalibrated images. It has been recently shown that if the deformation is modelled as a linear combination of basis shapes both the motion and the 3D structure of the object may be recovered using an extension of Tomasi and Kanade's factorization algorithm for affine cameras. The main drawback of the existing methods is that the non-rigid factorization algorithm does not provide a correct estimate of the motion: the motion matrix has a repetitive structure which is not respected by the factorization algorithm. This also affects the estimation of the 3D shape. In this paper we present a non-linear optimization method which minimizes image reprojection error and imposes the correct structure onto the motion matrix by choosing an appropriate parameterization. In addition, we propose a novel non-rigid tracking algorithm based on the use of ranklets, a multiscale family of rank features. Finally, we show that improved motion and shape estimates are obtained on a real image sequence of a person's face which is moving and changing expression.

1. Introduction

Recent work in non-rigid factorization [3, 2, 11] has proved that under weak perspective viewing conditions it is possible to infer the principal modes of deformation of an object alongside its 3D shape, within a structure from motion estimation framework. These non-rigid factorization methods stem from Tomasi and Kanade's factorization algorithm for rigid structure [10] developed in the early 90's. The key idea is the use of rank-constraints to express the geometric invariants present in the data. This allows the factorization of the matrix containing the image feature tracks into its shape and motion components. Crucially, these new factorization methods work purely from video in an unconstrained case: a single uncalibrated camera viewing an arbitrary 3D surface which is moving and articulating.

The main drawback of existing non-rigid factorization methods is that they minimize an algebraic cost function and they do not provide correct motion and structure estimates. The motion matrix recovered via non-rigid factorization should have a replicated block structure which is not preserved by these algorithms resulting in an ambiguity between the motion and structure parameters.

In this paper, we propose to overcome these problems using a bundle adjustment step to refine an initial solution by minimizing image reprojection error, which, contrary to other approaches, is a geometrically meaningful error function. Aanæs and Kahl first proposed the use of bundle-adjustment in the non-rigid case [1], however our approach differs in the choice of initialization and more notably in the parameterization of the problem and the quality of the results.

We also introduce a novel tracking algorithm based on ranklets, a recently developed family of multiscale rank features that present an orientation selectivity pattern similar to Haar wavelets [9]. Ranklets are used to encode an appearance based description of the neighbourhood for each of a sparse set of tracked points. The use of filter adaptation and dynamic model updating makes this algorithm particularly suitable for tracking deformable structures. In our experiments, we employ the coordinates of the tracked points as inputs for the factorization algorithm we introduce here.

This is, to the best of our knowledge, the first work to show that 3D shape and motion estimates can be succesfully disambiguated using bundle adjustment and to present comparative results with previous non-rigid factorization methods using real image sequences with automatically tracked points.

The paper is organized as follows. In section 2 we describe the use of rank constraints to compute motion and 3D shape within the factorization framework. We briefly outline the factorization algorithm and then describe the existing non-rigid factorization methods. In section 3 we present the non-linear optimization scheme based on the bundle adjustment framework while section 4 describes our non-parametric tracking algorithm based on ranklets. Finally in section 5 we present two set of experimental results comparing our approach with former methods and then showing the reconstruction quality of our unsupervised system for non-rigid structure from motion.

2 Non-rigid factorization

Tomasi and Kanade's factorization algorithm for rigid structure [10] has recently been extended to the case of non-rigid deformable 3D structure [3, 2, 11]. Here, a model is needed to express the deformations of the 3D shape in a compact way. A simple linear model is chosen where the 3D shape of any specific configuration of a non-rigid object is approximated by a linear combination of a set of K basis-shapes which represent the K principal modes of deformation of the object for P points. A perfectly rigid object would correspond to the situation where K=1. Each basis-shape $(\mathtt{S}_1, \mathtt{S}_2, ..., \mathtt{S}_K)$ is a 3 × P matrix which contains the 3D locations of P object points for that particular mode of deformation. The 3D shape of any configuration can be expressed in terms of the basis-shapes $\mathtt{S}_i$ and the deformation weights l_i in the following way:

$$\mathtt{S} = \sum_{i=1}^{K} l_i \mathtt{S}_i \qquad \mathtt{S}, \mathtt{S}_i \in \Re^{3\times P} \quad l_i \in \Re$$

If we assume a scaled orthographic projection model for the camera, the coordinates of the 2D image points observed at each frame f are related to the coordinates of the 3D points according to the following equation:

$$\mathtt{W}_f = \begin{bmatrix} u_{f,1} & \dots & u_{f,P} \\ v_{f,1} & \dots & v_{f,P} \end{bmatrix} = \mathtt{R}_f \left(\sum_{i=1}^{K} l_{f,i} \mathtt{S}_i \right) + \mathbf{T}_f \quad (1)$$

where $\mathtt{R}_f$ is a 2 × 3 orthonormal matrix which contains the first and second rows of the camera rotation matrix and $\mathbf{T}_f$ contains the first two components of the camera translation vector. Weak perspective is a good approximation when the depth variation within the object is small compared to the distance to the camera. The weak perspective scaling (f/Z_{avg}) is implicitly encoded in the $l_{f,i}$ deformation coefficients. We may eliminate the translation vector $\mathbf{T}_f$ by registering image points to the centroid in each frame. In this way, the 3D coordinate system will be centered at the centroid of the shape $\mathtt{S}$. If all P points can be tracked throughout an image sequence we may stack all the points tracks from frame 1 to F into a s $2F \times P$ measurement matrix W and we may write:

$$\mathtt{W} = \begin{bmatrix} u_{1,1} & \dots & u_{1,P} \\ v_{1,1} & \dots & v_{1,P} \\ \vdots & & \vdots \\ u_{F,1} & \dots & u_{F,P} \\ v_{F,1} & \dots & v_{F,P} \end{bmatrix} = \begin{bmatrix} l_{11}\mathtt{R}_1 & \dots & l_{1K}\mathtt{R}_1 \\ \vdots & & \vdots \\ l_{F1}\mathtt{R}_F & \dots & l_{FK}\mathtt{R}_F \end{bmatrix} \begin{bmatrix} \mathtt{S}_1 \\ \vdots \\ \mathtt{S}_K \end{bmatrix} = \mathtt{MS} \quad (2)$$

Since $\mathtt{M}$ is a $2F \times 3K$ matrix and $\mathtt{S}$ is a $3K \times P$ matrix, the rank of $\mathtt{W}$ when no noise is present must be $r \leq 3K$.

2.1 Previous work on non-rigid factorization

The rank constraint on the measurement matrix $\mathtt{W}$ can be easily imposed by truncating the SVD of $\mathtt{W}$ to rank 3K. This will factor $\mathtt{W}$ into a motion matrix $\tilde{\mathtt{M}}$ and a shape matrix $\tilde{\mathtt{S}}$. Note that in the non-rigid case the matrix $\tilde{\mathtt{M}}$ needs to be further decomposed into the 3D pose matrices $\mathtt{R}_f$ and the deformation weights l_{fk} since their values are mixed inside the motion matrix $\tilde{\mathtt{M}}$.

A further issue is that the result of the factorization of $\mathtt{W}$ into $\tilde{\mathtt{M}}$ and $\tilde{\mathtt{S}}$ is not unique since any invertible $3K \times 3K$ matrix $\mathtt{Q}$ can be inserted in the decomposition leading to the alternative factorization: $\mathtt{W} = (\tilde{\mathtt{M}}\mathtt{Q})(\mathtt{Q}^{-1}\tilde{\mathtt{S}})$. The problem is to find a transformation matrix $\mathtt{Q}$ that renders the appropriate replicated block structure of the motion matrix $\tilde{\mathtt{M}}$ shown in (2) and that removes the affine ambiguity upgrading the reconstruction to a metric one. Whereas in the rigid case the problem of computing the transformation matrix $\mathtt{Q}$ to upgrade the reconstruction to a metric one can be solved linearly [10], in the non-rigid case imposing the appropriate repetitive structure to the motion matrix $\tilde{\mathtt{M}}$ results in a non-linear problem. Various methods to recover the transformation matrix $\mathtt{Q}$ have been proposed so far in the literature [2, 3, 11] but they fail to provide a completely satisfactory solution.

Most of them do not respect the replicated block structure of the motion matrix $\mathtt{M}$ expressed in (2). It is important to notice that the replicated structure does not affect the estimation of the motion of image points, which makes these factorization algorithms very well suited to non-rigid tracking [11, 2]. However, if the main goal is to recover the correct camera matrices and the 3D non-rigid structure, preserving the replicated block structure of the motion matrix $\mathtt{M}$ after factorization becomes crucial. If this is not achieved, there follows an ambiguity between the motion parameters and the estimated 3D structure. Our solution to this issue consists in a non-linear optimization step obtained by minimizing a meaningful geometric cost function.

3 Bundle adjustment

Our approach is to obtain an initial solution for the non-rigid shape and 3D pose and then to perform a non-linear optimization step by minimizing image reprojection error.

The goal is to estimate the camera matrices $\mathtt{R}_i$ and the 3D structure parameters $l_{ik}, \mathbf{S}_k$ such that the distance between the measured image points $\mathbf{x}_{ij}$ and the estimated image points $\hat{\mathbf{x}}_{ij}$ is minimized:

$$\min_{\mathtt{R}_i \mathbf{S}_k l_{i,k}} \sum_{i,j} \| \mathbf{x}_{ij} - \hat{\mathbf{x}}_{ij} \|^2 = \min_{\mathtt{R}_i \mathbf{S}_k l_{i,k}} \sum_{i,j} \| \mathbf{x}_{ij} - (\mathtt{R}_i \sum_k l_{i,k} \mathbf{S}_k) \|^2 \tag{3}$$

This method is generically termed bundle-adjustment in the computer vision and photogrammetry communities and it provides a Maximum Likelihood estimate provided that the noise can be modelled with a Gaussian distribution. 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 [12].

The work presented here is most closely related to the work by Aanæs and Kahl with bundle adjustment [1]. However, their approach differs in some fundamental aspects. Firstly, the initial estimate for the non-rigid shape was obtained by estimating the mean and variance of 3D calibrated data obtained directly from image measurements. The authors state that their approach would work without this constraint, but they do not give experimental evidence. In contrast, we consider a scenario based on pure uncalibrated data from a generic video sequence. The second main difference is in the parameterization of the problem. In [1] the camera rotations are parameterized by the elements of the rotation matrix. Instead we have used quaternions which has proved to lead to better behaved results for the motion estimates as will be shown in the experimental section.

In terms of their experimental evaluation, Aanæs and Kahl do not provide an analysis of the recovered parameters, only some qualitative results of the 3D reconstruction. In contrast, our quantitative experimental analysis shows that it is possible to decouple motion parameters from deformation parameters and to obtain a sensible improvement in the quality of the reconstruction (see Section 5 for a detailed description).

3.1 Our approach

We have chosen to parameterize the camera matrices $\mathtt{R}_f$ using unit quaternions [5] giving a total of $4 \times F$ rotation parameters, where F is the total number of frames. Quaternions ensure that there are no singularities and that the orthonormality of the rotation matrices is preserved by merely forcing the normality of the 4-vector. This would not be the case with the Euler angle or the rotation matrix parameterization where orthonormality of the rotations is more complex to preserve. In our initial implementation, we parameterized the 3D pose using the 6 entries of the rotation matrices $\mathtt{R}_f$, however the use of quaternions led to improved convergence and to much better results for the rotation parameters and the 3D pose. The structure was parameterized with the $3 \times K \times P$ coordinates of the $\mathbf{S}_k$ basis shapes and the $K \times F$ deformation weights l_{ik}.

A non-rigid factorization method essentially similar to Brand's [2] (see Section 5 for details) was used for initialization purposes. However, the camera matrices were initialized with the motion corresponding to the rigid component, since it encodes the most significant part of the motion. This assumption works well in the scenario of human facial motion analysis, but would not be valid for highly deformable objects such as a hand or the human body.

The basis shapes were initialized with the values obtained using the non-rigid factorization method essentially similar to Brand's as were the weights associated with the rigid component. However, the weights associated with the basis shapes that account for the non-rigid motion were initialized to a very small value. The reason for this choice is that it was observed that this initial estimate, which effectively uses the rigid component of the shape and motion, led to a smaller value of the initial error function and to better convergence. Note that a significant component of rigid motion is required to estimate the 3D structure. We suggest for a scenario with a nearly static subject a stereo factorization approach [4] followed by an analogous non-linear refinement of the motion and shape components.

Occasionally the non-linear optimization leads to the solution corresponding to a local minimum. In particular, we have found that occasionally the 3D points tend to lie on a plane. To overcome this situation, a prior on the 3D shape has been added to the cost function. Our prior states that the depth of the points on the object surface will not change significantly from one frame to the next since the images are closely spaced in time adding the term $\sum_{i=2,j=1}^{i=F,j=P} \| S_z^{i-1,j} - S_z^{i,j} \|^2$ to the cost function. In this way we can preserve the relief present in the 3D data. Similar regularization terms have also been reported in [11, 2].

4 Non-rigid tracking using ranklets

In our experiments we employ a novel tracking algorithm for non-rigid objects based on the use of ranklets, a recently developed family of multiscale rank features with an orientation selectivity pattern similar to Haar wavelets (see [9] for a detailed description). The fiducial points used for 3D reconstruction are automatically selected based on a saliency criterion (specified below), and subsequently tracked throughout the image sequence.

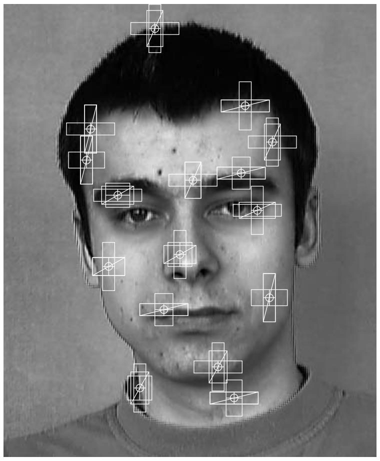


Figure 1: Support of the adapted filters for a few tracked points (lowest frequency channel). The orientation selectivity of each filter is indicated by a horizontal, vertical or diagonal line drawn across the corresponding box.

4.1 Adaptive appearance-based modelling

In analogy with classic approaches involving multiscale features [6, 8], we choose to encode the local image neighborhood of each point by means of a vector of ranklets consisting of a total of 9 filters arranged in 3 frequency channels and 3 orientation channels (corresponding, as in the case of Haar wavelets, to horizontal, vertical and diagonal edges). Saliency is proportional, for each point, to the norm of the corresponding ranklet vector; points are selected for tracking in decreasing saliency order (for the sequence in Figure 6, we decided to track 110 points).

For each tracked point, an optimization step is performed to adapt the shape of the filters to the appearance of the specific image neighborhood. This is done by independently varying the aspect ratio of the support of each ranklet in order to obtain the largest possible response (the area of the support is kept constant in the process). The support of a few adapted filters is shown in Figure 1.

Tracking is performed by using, for each tracked point, the adapted ranklet vector as a model. In each subsequent frame, a gradient descent algorithm is employed in a neighborhood of the previous position of the point in order to find the location that gives the best match.

4.2 Updating the models

Due to the deformations and changes in perspective of the tracked object, the quality of the match between the features extracted at the location of each point and the corresponding model generally deteriorates with time. This eventually results in failure to track the points.

To alleviate this problem, a new model is acquired from the current best estimate of the position of a point whenever the distance from the original model exceeds a threshold. The filter adaptation step is repeated and the new model is kept alongside the previous one in a “model stack” for the point. This procedure can be repeated up to a given maximum number of models. For each point, tracking is performed at first using the most recently acquired model, which is the last on the stack. A further gradient descent is then initiated in an attempt to match the previous model on the stack. If this succeeds, the last model is discarded, and the point is assumed to have recovered an appearance more similar its original one.

The model stack provides a mechanism for tracking a point across a range of deformations and perspective variations, during which the point may occasionally revert to the original appearance (for a related approach see [7]). Upon creation of a new model, an added check is performed to allow “grafting” it next to the most similar model already present in the stack; thus a point is not required to return to its original appearance by going through the same series of deformations in reverse order (although this will often be the case, for instance, for the points of a face that is rotating left to right and then right to left). Points are discarded when thresholds for the maximum number of models or the maximum frame-to-frame drift are exceeded.

5 Experimental results

Our experimental analysis has been devised with two main objectives in mind. Firstly we aim to compare the performance of our non-linear optimization approach – which minimizes image reprojection error – with previous non-rigid factorization methods, which minimize an algebraic error and do not impose the correct structure on the motion matrix. Our aim is to prove that the ambiguity between the rotation and the 3D structure parameters can be solved using our proposed bundle adjustment refinement step, which results in improved estimates. Since the focus of this test is a comparison with former methods, the features on the subject’s face were marked manually throughout the sequence.

Our second objective is to jointly test the novel non-rigid tracker described above and the non-linear optimization method in a realistic scenario, where the feature points are automatically obtained by the tracker and fed through the optimization procedure, in a fully unsupervised system. Thus the quality of the reconstruction confirms, on the one hand, the reliability of the tracked points, while on the other hand it shows the ability of the factorization algorithm to cope with the levels of noise inherent in automatically generated point traces.

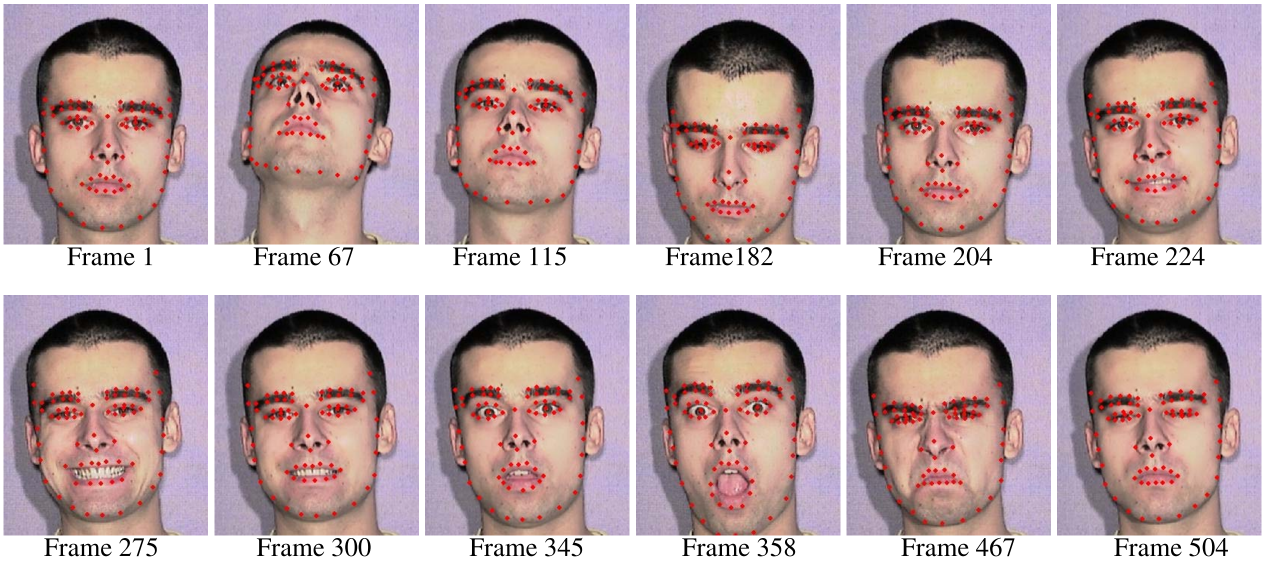
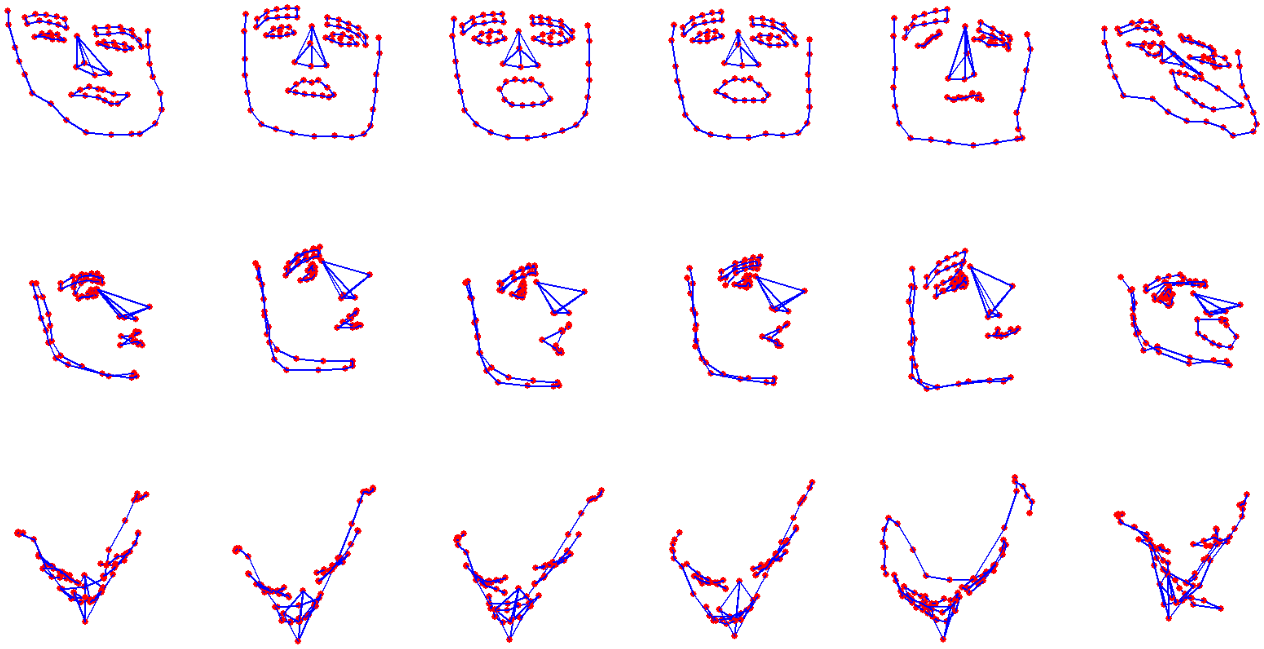
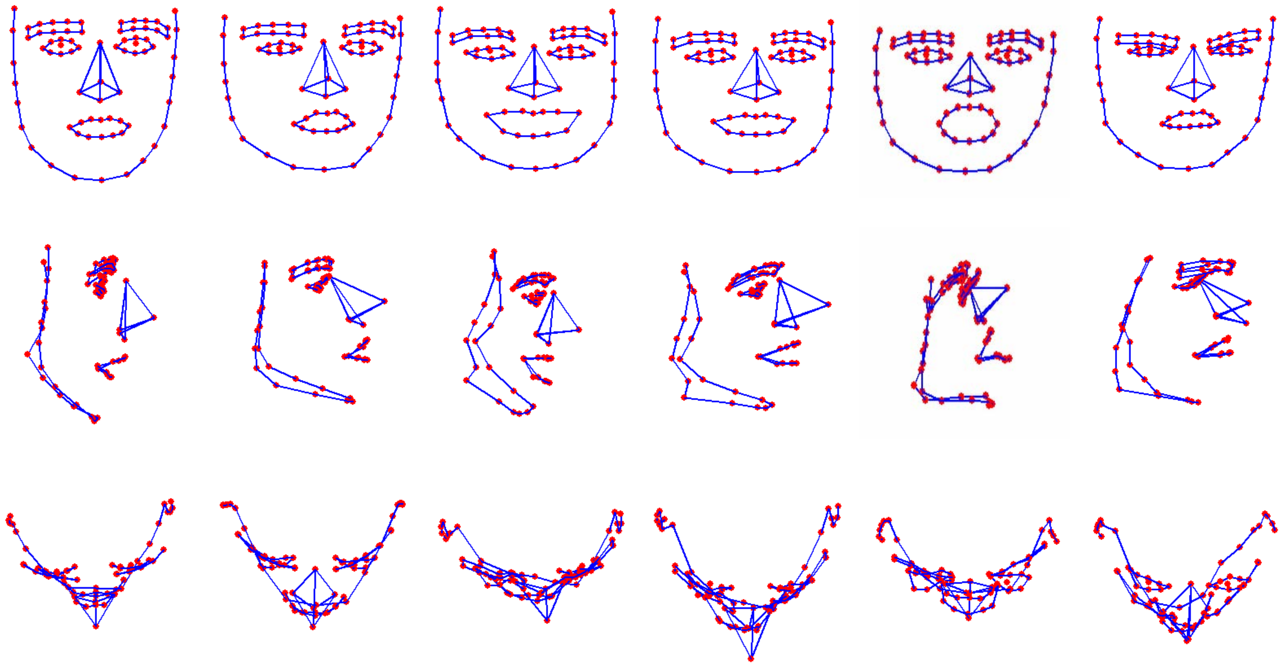


Figure 2: *Key frames in the sequence used in the experiments. The subject performed an almost rigid motion for the first 200 frames moving the head sideways and then then changed facial expression for the next 400 frames.*



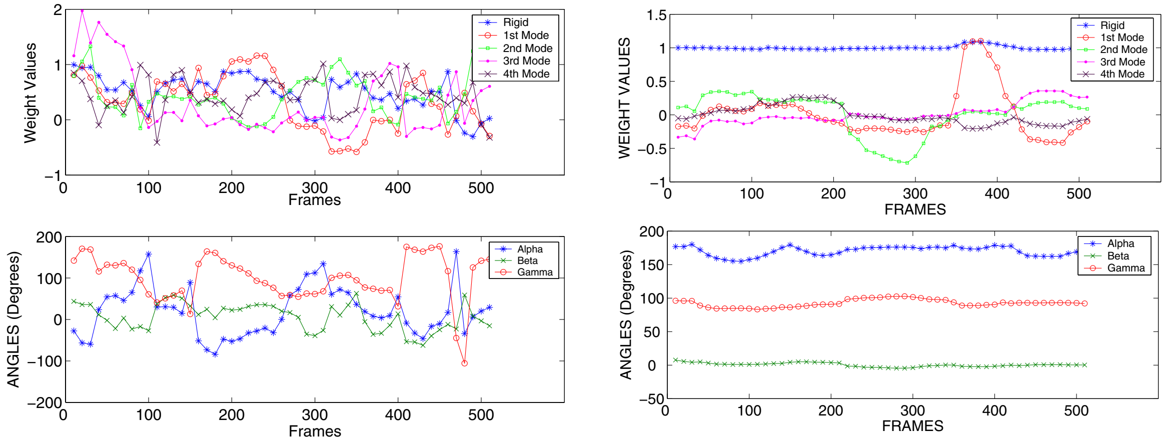
a) Frame 1 b) Frame 67 c) Frame 275 d) Frame 300 e) Frame 358 f) Frame 467

Figure 3: Front, side and top views of the 3D reconstructions obtained from the non-rigid factorization algorithm without bundle adjustment for some of the key frames in the sequence



a) Frame 1 b) Frame 67 c) Frame 275 d) Frame 300 e) Frame 358 f) Frame 467

Figure 4: Front, side and top views of the 3D reconstructions after bundle adjustment for some of the key frames in the sequence



(A) Results from non-rigid factorization (B) Results after bundle adjustment

Figure 5: This figure shows the values obtained for the configuration weights (top) and for the rotation angles (bottom) with the initial factorization algorithm (A) and after bundle adjustment (B). Bundle adjustment provides smoother and better behaved solutions for all the parameters.

We have used real video sequences of a person's face moving and changing his facial expression in both sets of experiments.

5.1 Analysis of bundle adjustment results

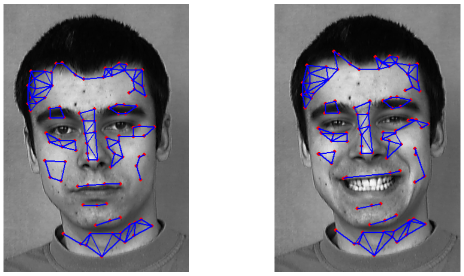
Here we compare the results obtained using one of the previous non-rigid factorization methods with the results obtained adding our non-linear optimization step. In this experiment the subject performed an almost rigid motion for the first 200 frames, moving his head sideways. The subject then changed facial expression with his head facing front for the next 400 frames (see Figure 2).

We first present the results using a non-rigid factorization scheme which in essence is very similar to Brand's method. The results of the 3D reconstructions obtained in this case for the some of the key frames in the sequence are shown in Figure 3. The recovered 3D shape does not reproduce the facial expressions very accurately. This can be seen by inspecting the front views of the 3D plots. The depth is not recovered accurately either and this is evident by looking at the top views of the reconstruction. Notice the asymmetry of the left and right side of the face.

In Figure 4 we show the reconstructed 3D shape recovered for the same sequence after applying the bundle adjustment refinement step. The facial expressions in the 3D plots reproduce the original ones reliably: notice for example the motion of the eyebrows in the frowning expression or the opening of the mouth for the surprise expression. Finally, the top views show that the overall relief appears to be well preserved as is the symmetry of the face.

Figure 5 shows the results obtained for the motion parameters. The graphs show the rotation angles about the X, Y and Z axes (up to an overall rotation) recovered for each frame of the 600 frame sequence. We show results obtained using the initial factorization method and the improved results after bundle adjustment. It can be observed how the results before bundle adjustment are very noisy and incorrect. However, after bundle adjustment the rotation angles are smooth and they represent the real motion accurately. In particular note how the "alpha" parameter, which corresponds to a tilt, varies smoothly throughout the first 200 frames capturing the up and down tilt of the head of about 50 degrees in total. The rotation angles about the other 2 axes do not vary significantly throughout the sequence which is what was expected.

The evolution throughout the sequence of the values of the deformation weights associated with the 5 modes of deformation is also shown to be smoother after bundle adjustment in Figure 5.



Frame 1 Frame 250

Figure 6: Key frames in the sequence used for unsupervised 3D reconstruction

5.2 Results with automatic tracking

We have also tested our non-linear optimization algorithm on a real scenario with automatic tracking of feature points as described in Section 4. The tracker has to cope with a complex long sequence where the subject is performing at the same time rigid motion and different facial expressions. For the 990 frame sequence only a total of 10 points out of 110 were lost showing that the tracker can adapt to the face deformations and to the perspective change due the rigid motion. A certain amount of points initialized on homogenous texture are to be considered outliers and affect evidently the 3D shape estimation.

We present in Figure 7 the 3D reconstruction of two frames from the video sequence. The overall depth is generally correct: notice the point belonging to the neck in comparison to the face position and the nose pointing out from the structure of the face. Generally the face symmetry is well preserved, as it is possible to notice from the top views of the reconstruction. Outliers are evident mostly on the cheeks and neck area where the tracker performs poorly, such feature points are wrongly reconstructed by our non-rigid model.

6. Summary and Conclusions

In this work we have introduced a novel non-linear optimization method for non-rigid structure from motion analysis, as well as an adaptive point tracking algorithm based on ranklets.

We have demonstrated that the quality of motion and structure recovery in non-rigid factorization is significantly improved with the addition of a bundle adjustment step. Moreover, the proposed solution is able to successfully disambiguate the motion and deformation components as shown in our experimental results.

The tracking algorithm we have introduced appears to handle effectively the deformations and changes in perspective of a non-rigid object. By integrating it with our factor-

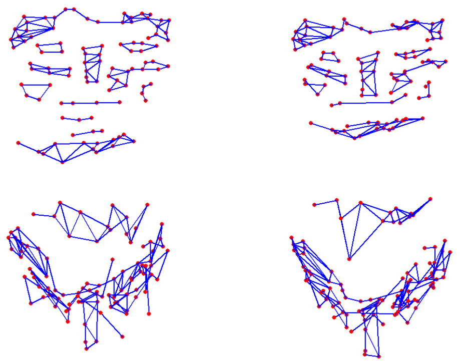


Figure 7: Front and top views of the 3D unsupervised reconstructions after bundle adjustment for two key frames in the sequence

ization algorithm, we have obtained a fully unsupervised system that can generally estimate a correct shape depth, although the occasional unreliable point traces result in a somewhat coarse approximation. We are currently working on ways to improve the robustness of tracking and factorization separately, as well as to harness the information extracted by the structure from motion algorithm itself in order to deal with the uncertainty in the tracked feature points.

Acknowledgements

The authors would like to thank Andrew Fitzgibbon for the matlab functions for bundle adjustment. Lukas Zalewski provided the sequences and the manual tracks used for the reconstruction.

References

[1] H. Aanæs and F. Kahl. Estimation of deformable structure and motion. In *Workshop on Vision and Modelling of Dynamic Scenes, ECCV'02, Copenhagen, Denmark*, 2002.

[2] M. Brand. Morphable models from video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, December 2001.

[3] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, pages 690–696, June 2000.

[4] A. Del Bue and L. Agapito. Non-rigid 3d shape recovery using stereo factorization. *Asian Conference of Computer Vision*, 1:25–30, January 2004.

[5] B.K.P. Horn. Closed form solutions of absolute orientation using unit quaternions. *J. Optical Soc. of America A.*, 4(4):629–642, 1987.

[6] B. S. Manjunath, C. Shekhar, and R. Chellappa. A new approach to image feature detection with applications. *Pattern Recognition*, 31:627–640, 1996.

[7] I. Matthews, T. Ishikawa, and S. Baker. The template update problem. In R. Harvey and J. A. Bangham, editors, *Proceedings of the British Machine Vision Conference, Norwich, UK*, volume II, pages 649–658, 2003.

[8] R. P. N. Rao and D. H. Ballard. An active vision archíteture based on iconic representations. *Artificial Intelligence Journal*, 78:461–505, 1995.

[9] F. Smeraldi. Ranklets: orientation selective nonparametric features applied to face detection. In *Proc. of the 16th ICPR*, volume 3, pages 379–382, Quebec QC, August 2002.

[10] C. Tomasi and T. Kanade. Shape and motion from image streams: a factorization method. *International Journal in Computer Vision*, 9(2):137–154, 1991.

[11] L. Torresani, D. Yang, E. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, 2001.

[12] Bill Triggs, Philip McLauchlan, Richard Hartley, and Andrew Fitzgibbon. Bundle adjustment – A modern synthesis. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, LNCS, pages 298–375. Springer Verlag, 2000.