Non-rigid 3D Factorization for Projective Reconstruction

X. Lladó, A. Del Bue and L. Agapito
Vision Group, Department of Computer Science
Queen Mary, University of London. London, E1 4NS, U.K.
{llado,alessio,lourdes}@dcs.qmul.ac.uk

Abstract

In this paper we address the problem of projective reconstruction for deformable objects. Recent work in non-rigid factorization has proved that it is possible to model deformations as a linear combination of basis shapes, allowing the recovery of camera motion and 3D shape under weak perspective viewing conditions. However, the performance of these methods degrades when the object of interest is close to the camera and strong perspective distortion is present in the data.

The main contribution of this work is the proposal of a practical method for the recovery of projective depths, camera motion and non-rigid 3D shape from a sequence of images under strong perspective conditions. Our approach is based on minimizing 2D reprojection errors, solving the minimization as four *weighted least squares* problems. Results using synthetic and real data are given to illustrate the performance of our method.

1 Introduction

The simultaneous recovery of camera motion and shapes from multiple images has been one of the fundamental problems in computer vision in recent years. Numerous techniques have been proposed to solve the *structure from motion* problem and one of the most successful approaches has been Tomasi and Kanade's factorization algorithm [11] developed in the early 90's. The key idea of their method 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 (*measurement matrix*) into its shape and motion components. Tomasi and Kanade's algorithm works for rigid scenes viewed under weak perspective conditions but the algorithm was later extended to work with more general camera models [8, 13].

It was only recently that the factorization framework was extended to deal with non-rigid objects. Most biological objects and natural scenes vary their shape, for instance, a tree, a moving animal or a face which is undergoing different facial expressions. The main challenge in non-rigid structure from motion is to disambiguate the contributions to the image motion caused by the shape deformations and the rigid motion.

Bregler et al. [2] were the first to extend the factorization framework to the non-rigid case. They introduced a representation for non-rigid 3D shape where any configuration

BMVC 2005 doi:10.5244/C.19.32

can be expressed as a linear combination of basis shapes that define the principal modes of deformation of the object. They proposed a factorization method that exploits the rank constraint on the measurement matrix and enforces orthonormality constraints on camera rotations to recover the motion and the non-rigid 3D shape. Torresani et al. [12] extended the method of Bregler et al. to a trilinear optimization problem by minimizing 2D image reprojection error using Alternating Least Squares. Brand [1] proposed an alternative optimization method and added an extra constraint on the basis shapes: the deformations should be as small as possible relative to the rigid shape. Xiao et al. [16] later proved that the orthogonality constraints were insufficient to disambiguate rigid motion and deformations. They identified a new set of constraints on the shape bases which, when used in addition to the rotation constraints, provide a closed form solution to the problem of non-rigid structure from motion. However, their solution requires that there be K frames (where K is the number of basis shapes) in which the shapes are independent. Recently, Del Bue et al. [3] have proposed a further non-linear optimization step that minimizes image reprojection error.

Note that all these 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 but problematic when the images are taken at closer distances and perspective distortions appear, such as when using webcams.

The main objective of this paper is to extend non-rigid factorization to the full perspective camera model. Many works have addressed projective factorization [13, 8, 15, 4, 10] but they assume that the objects in the scene are rigid. The rest of this work is organized as follows: The notation we use to formulate the problem is described in Section 2. In Section 3 we present the factorization framework used to recover projective depths, camera motion and 3D non-rigid structure. Some results using synthetic and real data are described in Section 4 to illustrate the performance of our method. Finally we give concluding remarks and the future direction of our work.

2 Background

2.1 Rigid Projective Factorization

The perspective projection equation models the projection of a 3D point $\mathbf{X}_j = [x_j, y_j, z_j, 1]^T$ on an image as

$$\lambda_{ij}\mathbf{x}_{ij} = \mathtt{P}_i\mathbf{X}_j \tag{1}$$

where $\mathbf{x}_{ij} = [u_{ij}, v_{ij}, 1]^T$ are the image coordinates of point j in the i^{th} view, λ_{ij} is the *projective depth* of the point and $\mathtt{P}_i$ is a 3×4 projection matrix. If we consider the projection of n scene points on m views we can construct a $(3m \times n)$ scaled measurement matrix

$$\mathtt{W} = \begin{bmatrix} \lambda_{11}\mathbf{x}_{11} & \dots & \lambda_{1n}\mathbf{x}_{1n} \\ \vdots & & \vdots \\ \lambda_{m1}\mathbf{x}_{m1} & \dots & \lambda_{mn}\mathbf{x}_{mn} \end{bmatrix} \tag{2}$$

which contains the image coordinates of all the points in all the views scaled by their projective depths. The matrix $\mathtt{W}$ can also be defined as the product $\mathtt{W} = \mathtt{PX}$ where $\mathtt{P} = [\mathtt{P}_1^T \mathtt{P}_2^T \dots \mathtt{P}_m^T]^T$ is the $(3m \times 4)$ matrix which gathers the projection matrices $\mathtt{P}_i$ in all m

frames and $\mathtt{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n]$ is the $(4 \times n)$ shape matrix that contains the projective coordinates of all n scene points.

Since $\mathtt{P}$ and $\mathtt{X}$ are at most rank 4, the rank of the scaled measurement matrix $\mathtt{W}$ is constrained to be $r \leq 4$. This constraint can be easily imposed by taking the Singular Value Decomposition of the measurement matrix and truncating it to rank 4. Therefore, if the projective depths $\{\lambda_{ij}\}$ were known it would be possible to factorize the measurement matrix into two rank-4 matrices, $\hat{\mathtt{P}}$ and $\hat{\mathtt{X}}$. However, the result of the factorization would not be unique since any invertible (4×4) matrix $\mathtt{Q}$ and its inverse can be inserted in the decomposition leading to the alternative camera and shape matrices $\{\hat{\mathtt{P}}\mathtt{Q}\}$ and $\{\mathtt{Q}^{-1}\hat{\mathtt{S}}\}$. Therefore, without assuming any additional constraints on the cameras or on the scene the reconstruction will be up to an overall projective transformation [6]. In general, the true projective depths λ_{ij} are unknown so the essence of projective factorization methods is to deal with the estimation of projective depths $\tilde{\lambda}_{ij}$ in order to obtain a measurement matrix which could be decomposed into camera motion and shape in 3D projective space using the rank constraint described above.

Various projective factorization methods have been proposed so far for the case of scenes with rigid objects. The first work to extend Tomasi and Kanade's algorithm to the perspective camera case was by Sturm and Triggs [9] who proposed a non-iterative factorization method for uncalibrated cameras. The method solves for the projective depths by calculating the fundamental matrices and epipoles between pairs of views. The overall accuracy of the algorithm depends greatly on the estimation of the epipolar geometry as large errors in the fundamental matrix would affect the measurement matrix and result in errors in the shape and motion. On the other hand, Han and Kanade [4] perform a projective reconstruction using a bilinear factorization algorithm without calculating the fundamental matrices. Heyden's method uses a different approach. It relies on using subspace constraints to perform projective structure from motion [7]. Ueshiba and Tomita [15] presented a method by which the projective depths are iteratively estimated so that the measurement matrix is made close to rank 4. The authors also derived metric constraints for a perspective camera model in the case where the intrinsic camera parameters are available. Recently, Tang and Hung [10] proposed an iterative algorithm for projective reconstruction based on minimizing the 2D reprojection errors. They show that 2D reprojection errors can be approximated by weighting each term of SVD reprojection errors by an appropriate weighting factor.

2.2 Non-rigid Factorization

Tomasi and Kanade's factorization has recently been extended to the case of non-rigid 3D structure, assuming affine viewing conditions [2, 1, 12, 3]. The model used to express the deformations is point-wise and the 3D shape of any specific configuration $\mathtt{S}$ is approximated by a linear combination of a set of d basis shapes $\mathtt{S}_k$ which represent the principal modes of deformation of the object:

$$\mathtt{S} = \sum_{k=1}^{d} l_k \mathtt{S}_k \qquad \mathtt{S}, \mathtt{S}_k \in \Re^{3\times n} \quad l_k \in \Re \tag{3}$$

where each basis shape $\mathtt{S}_k$ is a $3 \times n$ matrix which contains the 3D locations of n object points for that particular mode of deformation. The shape is then projected onto an image

frame i giving n image points:

$$\left[\begin{array}{cccc} \mathbf{x}_{i1} & \ldots & \mathbf{x}_{in} \end{array} \right] = \mathtt{R}_i \left(\sum_{k=1}^{d} l_{ik} \mathtt{S}_k \right) \tag{4}$$

where 2D and 3D points are expressed in non-homogeneous coordinates referred to the centroid of the object and $\mathtt{R}_i$ is the 2×3 orthographic camera matrix for a specific frame i. If all n points are tracked in m image frames we may construct the $2m \times n$ measurement matrix $\mathtt{W}$ whose rank is constrained to be at most $3d$, where d is the number of deformations. This rank constraint can be exploited to factorize the measurement matrix to obtain the 3D pose, configuration coefficients and a pre-specified number of 3D basis shapes. A summary of existing methods was given in section 1.

3 Our Perspective Factorization Framework

If we now assume a perspective projection model for the camera, the 3D shape will be projected onto image frame i according to the following equation:

$$\left[\begin{array}{cccc} \lambda_{i1}\mathbf{x}_{i1} & \ldots & \lambda_{in}\mathbf{x}_{in} \end{array} \right] = \mathtt{P}_i \left(\sum_{k=1}^{d} l_{ik} \mathtt{S}_k \right) \qquad \mathtt{S}_k \in \Re^{4\times n} \quad l_{ik} \in \Re. \tag{5}$$

Here, λ_{ij} are the projective depths, $\mathbf{x}_{ij}$ are the image coordinates of the n 3D points expressed in homogeneous coordinates, $\mathtt{P}_i$ is the 3×4 perspective projection matrix corresponding to frame i and each $\mathtt{S}_k = [\mathtt{S}_{k1} \ \ldots \ \mathtt{S}_{kn}]$ is now a $4 \times n$ matrix which contains the 3D locations in homogeneous coordinates of n object points for the k^{th} mode of deformation. We can now write the equation for the perspective camera case as:

$$\mathtt{W} = \left[\begin{array}{ccc} \lambda_{11}\mathbf{x}_{11} & \ldots & \lambda_{1n}\mathbf{x}_{1n} \\ \vdots & & \vdots \\ \lambda_{m1}\mathbf{x}_{m1} & \ldots & \lambda_{mn}\mathbf{x}_{mn} \end{array} \right] = \left[\begin{array}{ccc} l_{11}\mathtt{P}_1 & \ldots & l_{1d}\mathtt{P}_1 \\ \vdots & & \vdots \\ l_{m1}\mathtt{P}_m & \ldots & l_{md}\mathtt{P}_m \end{array} \right] \left[\begin{array}{c} \mathbf{S}_1 \\ \vdots \\ \mathbf{S}_d \end{array} \right] \tag{6}$$

Clearly, the rank of the measurement matrix is at most $4d$ for the projective case. If the projective depths λ_{ij} were known the measurement matrix could be decomposed into the motion and shape matrices using SVD. However, our strategy is to compute the projective depths and the 3D shape and motion simultaneously.

3.1 Estimating Projective Depths and Non-rigid 3D Structure

In order to estimate projective depths λ_{ij} we propose to minimize the following cost function:

$$\min_{\hat{\mathtt{P}}\hat{\mathbf{S}}\hat{\mathtt{L}}\tilde{\lambda}} \sum_{i,j=1}^{m,n} \gamma_{ij} \left\| \tilde{\lambda}_{ij}\mathbf{x}_{ij} - \hat{\mathtt{P}}_i \sum_{k=1}^{d} \hat{l}_{ik}\hat{\mathbf{S}}_{kj} \right\|^2 \tag{7}$$

which expresses the SVD reprojection error for each point in each view weighted by a factor $\gamma_{ij} = 1/(\tilde{\lambda}_{ij})^2$. The purpose of this factor is simply to approximate the residual for each point to the 2D image reprojection error $\| \mathbf{x}_{ij} - \hat{\mathbf{x}}_{ij} \|^2$. (See [14, 10] for details).

Since this is a non-linear minimization on $\hat{\mathtt{P}}$, $\hat{\mathtt{S}}$, $\hat{\mathtt{L}}$ and $\tilde{\lambda}$ it can be expressed as four different WLS problems where $\hat{\mathtt{P}}$, $\hat{\mathtt{S}}$, $\hat{\mathtt{L}}$ and $\tilde{\lambda}$ are evaluated one by one iteratively while keeping the others unchanged. Appendix A details how we rearrange the general expression defined in equation (6) to solve these minimizations in a least-squares sense. Note that our work can be seen either as an extension of Tang and Hung's projective factorization [10] to the non-rigid case or as an extension of Torresani et. al's non-rigid factorization [12] to the projective case.

In order to ensure good numerical conditioning we work with normalized image coordinates as described in [5]. In terms of the initialization, if initial guesses for $\hat{\mathtt{P}}$ and $\hat{\mathtt{S}}$ are not provided, they are initialized by the rank 4 approximation of $\{\tilde{\lambda}_{ij}\mathbf{x}_{ij}\}$. The depths $\tilde{\lambda}_{ij}$ are initialized to 1 and the configuration weights can be initialized to small values, otherwise the weight associated to the rigid component can be estimated by an initial rigid factorization.

4 Experimental Results

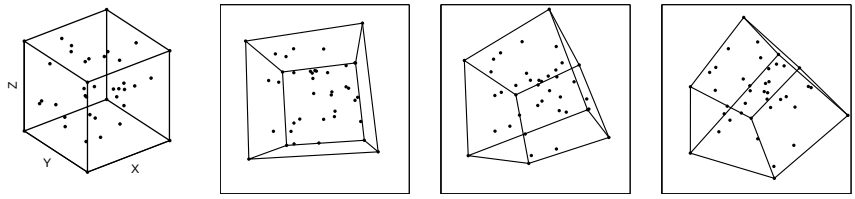
4.1 Synthetic data

The synthetic 3D data consisted of a set of 40 random points sampled inside a cube of size $50 \times 50 \times 50$ units. We used two configurations of the 3D data points: one in which 30 points remained rigid throughout the sequence (including the 8 vertices of the cube) and 10 were deforming and the second one in which only the 8 vertices of the cube were rigid while the remaining 32 points deformed. Our aim is to show the performance of our approach under different degrees of non-rigidity. The deformations for the non-rigid points were generated using random basis shapes as well as random deformation weights. The first basis shape had the largest weight equal to 1. Different number of basis shapes ($d = 2$, 3, 4 and 5) were used to show the performance of the algorithm with respect to the number of basis shapes. The data was then rotated and translated over 20 frames. The overall maximum rotation about any axis was 90 degrees.

The 3D data was then projected onto the images using 4 different camera setups varying the distance of the object to the camera and the focal length to achieve increasing levels of perspective distortion (Setup 1: z=250, f=1000; Setup 2: z=200, f=1000; Setup 3: z=150, f=800; Setup 4: z=100, f=500). Figure 1 shows an example of a 3D shape and 3 different perspective views of the synthetic scene using Setup 3. We show results for increasing levels of Gaussian noise where σ varied from 0 to 2 pixels with 0.5 pixels increments.

Figure 2 shows the r.m.s. 2D reprojection errors expressed in pixels for both configurations of 3D points and for the 4 camera setups with different number of basis shapes and varying levels of noise. Note that the plots show the mean values corresponding to 5 trials for each level of noise. Figure 2 also illustrates the 3D reconstruction errors expressed in percentage relative to the scene size which we defined as the maximum of the x, y and z coordinates. The 3D error was computed after aligning the projective reconstruction with the Euclidean 3D model.

The algorithm appears to perform well in the presence of image noise. Note that the 3D reconstruction error is well below 4% even for large perspective distortions (Setups 3 and 4) and for large levels of image noise. The 2D error is also small and it appears to be of the same order as the image noise. Note also that when the number of basis



(a) (b) (c) (d)

Figure 1: Synthetic sequence. (a) Example of ground truth of the 3D shape. (b)-(d) Different views of the sequence for z=150 and f=800. These images show the strong perspective effects present in the 2D data.

shapes increases the 2D error decreases but the 3D error increases. The algorithm usually converges within around one hundred iterations.

To illustrate the performance of our method, in Figure 3 we compare the recovered reconstruction with one obtained using an orthographic non-rigid factorization method which uses bundle-adjustment to recover 3D non-rigid shape and motion. As expected, the results obtained with the affine factorization are unsatisfactory as it fails to model the strong perspective distortions. Note that the projective shape was aligned with the Euclidean 3D model while the affine reconstruction was upgraded to euclidean using orthonormality constraints on the camera matrices.

4.2 Real data

In this experiment we use real 3D data of a human face undergoing rigid motion while performing different facial expressions. The 3D data was captured using a VICON motion capture system by tracking the subject wearing 37 markers on the face. The 3D points were then projected synthetically onto an image sequence 931 frames long. The size of the face model was $169 \times 193 \times 102$ units, and the camera setup was such that the subject was at a distance of 300 units from the camera and the focal length was 600 units so the perspective effects were significant.

Figure 4 shows front, side and top views of the ground truth and 3D reconstructions obtained using our projective method for frames 1, 501, 821 and 930. To estimate the quality of the projective reconstruction, we aligned it with the Euclidean model of the scene. The number of basis shapes was set to $d = 6$. The average 2D reprojection error was 0.49 pixels while the absolute 3D error was 2.71 units. Note that the reconstructed 3D model has a natural looking shape and the deformations model faithfully the different expressions. However, some extreme deformations are not well recovered. See for example the reconstruction of the open mouth in frame 930. In that case the shape of the mouth is close to the ground truth in the frontal view, but the depth is not recovered accurately. We repeated the experiment using the same configuration but introducing an error of 0.5 pixels in the original sequence. In that case the 2D reprojection error was 0.79 pixels while the 3D reconstruction error was 4.6 units.

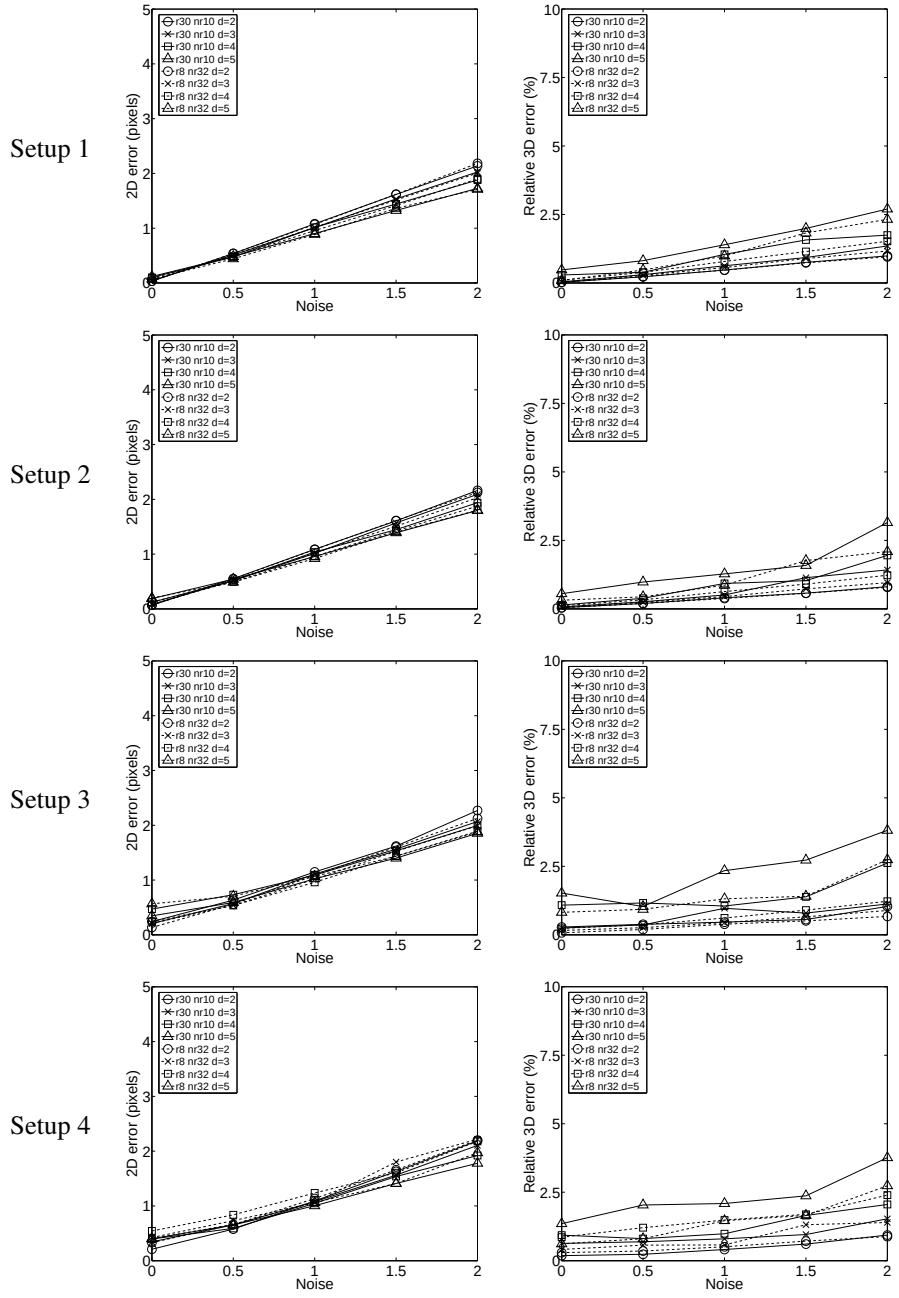
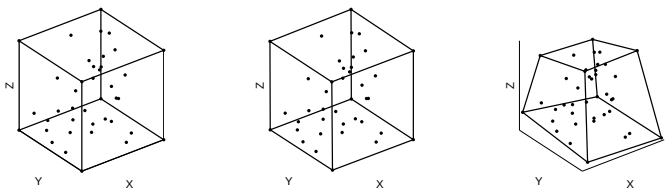


Figure 2: 2D error and 3D error curves for each camera setup. Each plot shows the results obtained using our projective WLS over the sequences with $(30/10)$ and $(8/32)$ rigid/non-rigid points respectively. Each plot also show the results of the experiments using different number of basis shapes (2, 3, 4 and 5).



(a) (b) (c)

Figure 3: Example of shape reconstruction. (a) Ground truth. (b) Reconstruction using the projective method. (c) Reconstruction using the orthographic method.

5 Conclusions

In this paper we have proposed a non-rigid factorization method able to recover camera motion and non-rigid structure from a sequence of images under strong perspective distortions. The recovery of projective depths, camera motion and non-rigid 3D shape has been achieved minimizing the 2D reprojection error using an alternate least squares scheme for the minimization. Our results on synthetic and real data have proved the performance of our method even for cases with significant deformation and strong perspective effects. A further nonlinear optimization step by using bundle adjustment could be used in order to refine the final motion and structure estimates. Furthermore, handling missing data could be treated simply by deleting the equations referring to the missing points.

Acknowledgments

This work has been supported by EPSRC grant GR/S61539/01. ADB holds a Queen Mary Studentship Award.

References

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

[2] 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.

[3] A. Del Bue, F. Smeraldi and L. Agapito. Non-rigid structure from motion using non-parametric tracking and non-linear optimization. In *IEEE Workshop on Computer Vision and Pattern Recognition Workshop, CVPRW'04, Washington DC, USA*, 2004.

[4] M. Han and T. Kanade. Scene reconstruction from multiple uncalibrated views. Technical Report CMU-RI-TR-00-09, Carnegie Mellon University, Pittsburgh, PA, January 2000.

[5] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[6] R.I. Hartley, R. Gupta, and T. Chang. Stereo from uncalibrated cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–764, 1992.

[7] A. Heyden. Projective structure and motion from image sequences using subspace methods. In *Proc. 10th Scandinavian Conference on Image Analysis, Lappeenranta, Finland*, pages 963–968, June 1997.

[8] C. J. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. In *Proc. 3rd European Conference on Computer Vision, Stockholm*, volume 2, pages 97–108, 1994.

[9] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *Proc. 4th European Conference on Computer Vision, Cambridge*, pages 709–720, April 1996.

[10] W. K. Tang and Y. S. Hung. A factorization-based method for projective reconstruction with minimization of 2-d reprojection errors. In *Proc. of the 24th DAGM Symposium on Pattern Recognition*, pages 387–394, London, UK, 2002. Springer-Verlag.

[11] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[12] 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.

[13] B. Triggs. Factorization methods for projective structure and motion. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 845–851, 1996.

[14] B. Triggs. Some notes on factorization methods for projective structure and motion. Unpublished, 1998.

[15] T. Ueshiba and F. Tomita. A factorization method for projective and euclidean reconstruction from multiple perspective views via iterative depth estimation. In *Proc. 5th European Conference on Computer Vision, Freiburg, Germany*, volume 1, pages 296–310, 1998.

[16] J. Xiao, J. Chai and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. In *Proc. 8th European Conference on Computer Vision, Prague, Czech Republic*, May 2004.

A Solving the WLS Minimizations

In this appendix we show different ways to rearrange equation (6) in order to solve the 4 WLS minimization problems described in section 3.1.

- We solve for each $\mathbf{S}_j$ (the shape bases associated with each 3D point j) using:

$$\begin{bmatrix} \lambda_{1j}\mathbf{x}_{1j} \\ \vdots \\ \lambda_{mj}\mathbf{x}_{mj} \end{bmatrix} = \begin{bmatrix} l_{11}\mathtt{P}_1 & \cdots & l_{1d}\mathtt{P}_1 \\ \vdots & \ddots & \vdots \\ l_{m1}\mathtt{P}_m & \cdots & l_{md}\mathtt{P}_m \end{bmatrix} \begin{bmatrix} \mathbf{S}_j \end{bmatrix} \tag{8}$$

- Solving for $\mathtt{P}_i$ is straightforward rewriting equation (6) as:

$$\lambda_{ij}\mathbf{x}_{ij} = \mathtt{P}_i \sum_{k=1}^{d} l_{ik}\mathtt{S}_{kj} \tag{9}$$

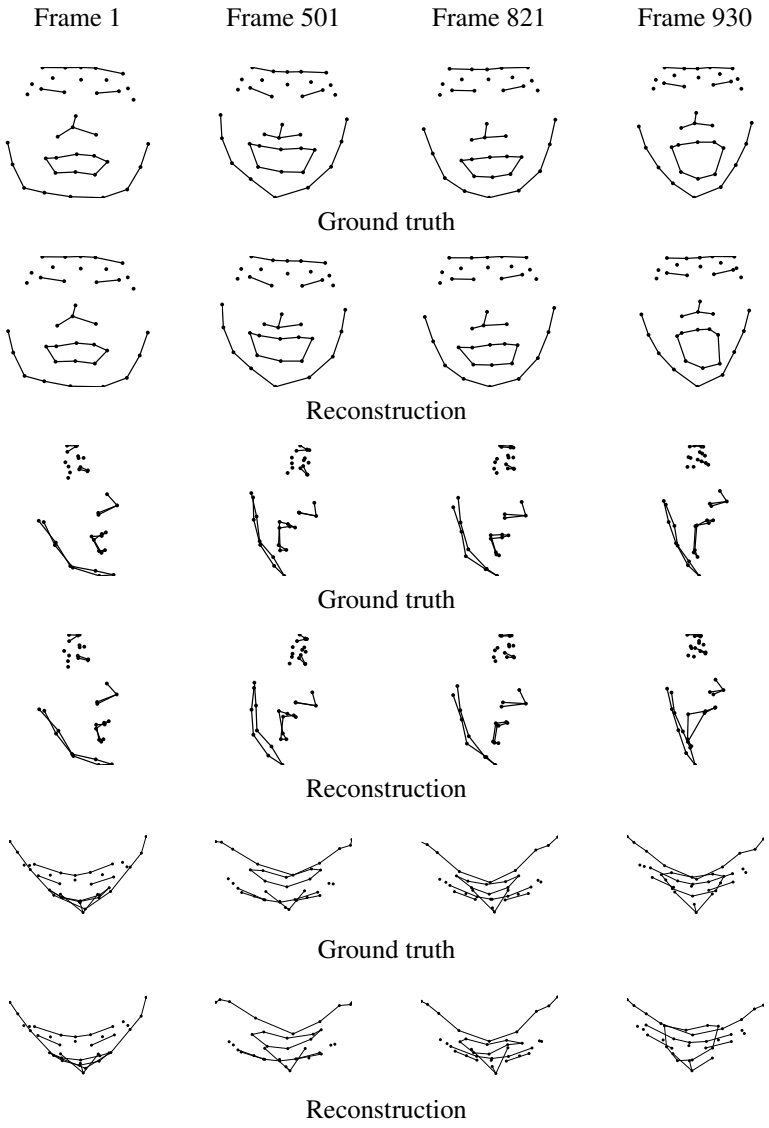


Figure 4: Front, side and top views of the ground truth and reconstructed face. Reconstructions are shown for frames 1, 501, 821 and 930.

- The following rearrangement of equation (6) allows the recovery of the configuration weights in a least-squares sense:

$$\begin{bmatrix} \lambda_{i1}x_{i1} \\ \vdots \\ \lambda_{in}x_{in} \end{bmatrix} = \begin{bmatrix} \mathrm{P}_i\mathbf{S}_{11} & \cdots & \mathrm{P}_i\mathbf{S}_{d1} \\ \vdots & & \vdots \\ \mathrm{P}_i\mathbf{S}_{1n} & \cdots & \mathrm{P}_i\mathbf{S}_{dn} \end{bmatrix} \begin{bmatrix} l_{i1} \\ \vdots \\ l_{id} \end{bmatrix} \tag{10}$$

- Solving for λ_{ij} is a straightforward minimization solved in a least-squares sense directly from equation (6).