Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Image and Vision Computing 25 (2007) 297–310

image AND vision COMPUTING

www.elsevier.com/locate/imavis

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

A. Del Bue *, F. Smeraldi, L. Agapito

Department of Computer Science, Queen Mary University of London, London E1 4NS, UK

Received 19 October 2004; received in revised form 4 August 2005; accepted 11 October 2005

Abstract

In this paper, we address the problem of estimating the 3D structure and motion of a deformable object given a set of image features tracked automatically throughout a video sequence. Our contributions are twofold: firstly, we propose a new approach to improve motion and structure estimates using a non-linear optimization scheme and secondly we propose a tracking algorithm based on ranklets, a recently developed family of orientation selective rank features.

It has been shown that if the 3D deformations of an object can be modeled as a linear combination of shape bases then both its motion and shape may be recovered using an extension of Tomasi and Kanade's factorization algorithm for affine cameras. Crucially, these new factorization methods are model free and 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 methods is that they do not provide correct structure and motion estimates: the motion matrix has a repetitive structure which is not respected by the factorization algorithm. In this paper, we present a non-linear optimization method to refine the motion and shape estimates which minimizes the image reprojection error and imposes the correct structure onto the motion matrix by choosing an appropriate parameterization.

Factorization algorithms require as input a set of feature tracks or correspondences found throughout the image sequence. The challenge here is to track the features while the object is deforming and the appearance of the image therefore changing. We propose a model free tracking algorithm based on ranklets, a multi-scale family of rank features that present an orientation selectivity pattern similar to Haar wavelets. A vector of ranklets is used to encode an appearance based description of a neighborhood of each tracked point. Robustness is enhanced by adapting, for each point, the shape of the filters to the structure of the particular neighborhood. A stack of models is maintained for each tracked point in order to manage large appearance variations with limited drift. Our experiments on sequences of a human subject performing different facial expressions show that this tracker provides a good set of feature correspondences for the non-rigid 3D reconstruction algorithm.

Keywords: Non-rigid structure for motion; Rank features; Ranklets; Point tracking

1. Introduction

Recent work in non-rigid factorization [4,6,23] 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 [22] developed in the early 1990'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 measurement matrix—which contains the image coordinates of a set of features matched throughout an image sequence—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.

Bregler et al. [6] were the first to use a factorization-based method for the recovery of non-rigid structure and motion. The decomposition between motion and shape parameters is not unique however, and the motion matrix is only obtained up to a post-multiplication by a transformation matrix. While this matrix can be easily computed in the case of rigid structure by enforcing orthonormality constraints on the camera motion, its computation in the non-rigid case is not trivial since the motion matrix has a replicated block structure which must be imposed.

* Corresponding author.
E-mail address: lourdes@dcs.qmul.ac.uk (L. Agapito).
URL: http://www.dcs.qmul.ac.uk/~lourdes/.

0262-8856/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2005.10.004

Several methods have been proposed so far to compute the transformation matrix. Bregler et al. [6] enforced orthonormality constraints on the camera rotations in a similar way to the rigid factorization scheme. Later, Brand [4] proposed an improvement to Bregler et al.'s method using numerically well-behaved heuristics to compute the transformation matrix and adding a final minimization to regularize the shape. Torresani et al. [23] also extended the method by Bregler et al. by introducing a final trilinear optimization on the motion and structure parameters. However, none of these methods is completely satisfactory at recovering the 3D structure since they do not impose the full block structure on the motion matrix.

Recently, Xiao et al. [25], proved that the orthonormality constraints on the camera rotations are not sufficient to compute the transformation matrix and they proposed a new set of constraints on the shape bases. Their work proves that when both sets of constraints are imposed, a closed-form solution to the problem of non-rigid structure from motion exists. However, their solution requires that there be K frames (where K is the number of basis shapes) in which the shapes are known to be independent.

In this paper, we propose an alternative solution to the computation of the transformation matrix which uses a bundle adjustment step to refine an initial estimate by minimizing the 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 in the parameterization of the problem. The effectiveness of our solution is supported by comparative results with existing non-rigid factorization methods using real image sequences with points tracked automatically with the algorithm outlined below.

The rank constraint can also be used to improve the estimate of optical flow in areas of the image with low texture [4,23], an approach inspired by its rigid equivalent [11]. However, optical flow estimation, being a differential operation, is inherently sensitive to noise. For this reason we decided to base our reconstruction on a point tracking algorithm, which according to the image descriptors used can afford a greater robustness.

In our case, the choice of image descriptors is dictated by the nature of the structure from motion problem, which requires substantial pose variations in order to achieve accurate reconstructions. The problem is aggravated by the deformations of the subject, making the use of highly invariant descriptors mandatory. A natural choice is represented by rank features that have often been applied to the matching problem because of their invariance under a wide range of transformations [2,27].

In this paper, we introduce a tracking algorithm based on ranklets, a recently developed family of multi-scale rank features that present an orientation selectivity pattern similar to Haar wavelets [20]. The usefulness of orientation selectivity in appearance-based features is supported both by classic Gabor filter approaches [15] and by its ubiquitous presence in biological vision systems [8]. In the case of ranklets, orientation selectivity is supplemented by the inherent robustness of rank based descriptors. Ranklets have been shown to be effective in a challenging pattern recognition task over deformable objects, namely face detection [21].

Similarly to classic multi-scale algorithms [15,18], our approach uses a vector of ranklets to encode an appearance based description of the neighborhood of each of a sparse set of tracked points. The aspect ratio of the filters is adapted for each local neighborhood to maximize filter response; this is expected to make the representation locally more discriminative, thus minimizing drift. Large variations in appearance are handled by implementing a stack of models for each tracked point. This allows the tracker to follow the points through a wide range of deformations with limited drift, and eventually recalibrate whenever the object reverts to its original appearance. The use of filter adaptation and dynamic model updating makes this algorithm particularly suitable for tracking deformable structures.

The paper is organized as follows. In Section 2, we review 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. In Section 5, we present two sets 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. Finally, we present our conclusions and an Appendix A with a description of the ranklet feature family.

2. Non-rigid factorization: overview

Tomasi and Kanade's factorization algorithm for rigid structure [22] has been recently extended to the case of non-rigid deformable 3D structure [4,6,23]. Here, the deformations of the 3D shape are modeled linearly so that the 3D shape of any specific configuration of a non-rigid object is approximated by a linear combination of a set of K shape bases which represent the K principal modes of deformation of the object. A perfectly rigid object would correspond to the situation where $K=1$. Each basis-shape $(S_1, S_2, \ldots S_k)$ is a $3\times P$ matrix which contains the 3D locations of the P points describing the object for that particular mode of deformation. The 3D shape of any configuration can be expressed in terms of the shape bases $\mathbf{S}_i$ and the deformation weights l_i in the following way:

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

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

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

where

$$\mathbf{R}_f = \begin{bmatrix} r_{f,1} & r_{f,2} & r_{f,3} \\ r_{f,4} & r_{f,5} & r_{f,6} \end{bmatrix} \tag{2}$$

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 origin of the 3D coordinate system will be located at the centroid of the shape $\mathbf{S}$. If all P points can be tracked throughout an image sequence we may stack all the point tracks from frame 1 to F into a $2F \times P$ measurement matrix $\mathbf{W}$ and we may write:

$$\mathbf{W} = \begin{bmatrix} u_{1,1} & \dots & u_{1,P} \\ v_{1,1} & \dots & v_{1,P} \\ \vdots & & \vdots \\ u_{F,1} & \dots & u_{F,P} \\ v_{F,1} & \dots & v_{F,P} \end{bmatrix} = \begin{bmatrix} l_{11}\mathbf{R}_1 & \dots & l_{1K}\mathbf{R}_1 \\ \vdots & & \vdots \\ l_{F1}\mathbf{R}_F & \dots & l_{FK}\mathbf{R}_F \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 \\ \vdots \\ \mathbf{S}_K \end{bmatrix} = \mathbf{MS} \tag{3}$$

Since $\mathbf{M}$ is a $2F \times 3K$ matrix and $\mathbf{S}$ is a $3K \times P$ matrix, in the noiseless case, the rank of $\mathbf{W}$ is $r \leq 3K$ (with $r = 3K$ when no degeneracies are present). Note that, in relation to rigid factorization, in the non-rigid case the rank is incremented by three with every new mode of deformation. The goal of factorization algorithms is to exploit this rank constraint to recover the 3D pose and shape (shape bases and deformation coefficients) of the object from the point correspondences stored in $\mathbf{W}$.

2.1. Previous work on non-rigid factorization

The rank constraint on the measurement matrix $\mathbf{W}$ can be easily imposed by truncating the SVD of $\mathbf{W}$ to rank $3K$. This will factor $\mathbf{W}$ into a motion matrix $\tilde{\mathbf{M}}$ and a shape matrix $\tilde{\mathbf{S}}$. Note that two issues have to be solved to obtain a successful decomposition into the correct motion and shape structure.

Firstly, the factorization of $\mathbf{W}$ into $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{S}}$ is not unique since any invertible $3K \times 3K$ matrix $\mathbf{Q}$ can be inserted in the decomposition leading to the alternative factorization: $\mathbf{W} = (\tilde{\mathbf{M}}\mathbf{Q})(\mathbf{Q}^{-1}\tilde{\mathbf{S}})$. The problem is to find a transformation matrix $\mathbf{Q}$ that renders the appropriate replicated block structure of the motion matrix shown in Eq. (3) and that removes the affine ambiguity, upgrading the reconstruction to a metric one. When 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 $\mathbf{M}$ after factorization becomes crucial. If this is not achieved, there follows an ambiguity between the motion parameters and the estimated 3D structure.

Secondly, in the non-rigid case the matrix $\mathbf{M}$ needs to be further decomposed into the 3D pose matrices $\mathbf{R}_f$ and the deformation weights l_{fk}, since their values are mixed inside the motion matrix (see Eq. (3)).

2.1.1. Computing the transformation matrix Q

In the rigid case (where the number of bases is $K = 1$) the problem of computing the transformation matrix $\mathbf{Q}$ that upgrades the reconstruction to a metric one can be solved linearly [22]. However, in the non-rigid case imposing the appropriate repetitive structure to the motion matrix $\tilde{\mathbf{M}}$ results in a more complex problem. The approach proposed by Brand [4] consists of correcting each column triple independently applying the rigid metric constraint to each $\tilde{\mathbf{M}}^k_{2F\times 3}$ vertical block in $\tilde{\mathbf{M}}$ shown here:

$$\tilde{\mathbf{M}} = \begin{bmatrix} \tilde{\mathbf{M}}^1 & \dots & \tilde{\mathbf{M}}^K \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{M}}_{11} & \dots & \tilde{\mathbf{M}}_{1K} \\ \vdots & & \vdots \\ \tilde{\mathbf{M}}_{F1} & \dots & \tilde{\mathbf{M}}_{FK} \end{bmatrix} = \begin{bmatrix} l_{1,1}\mathbf{R}_1 & \dots & l_{1,K}\mathbf{R}_1 \\ \vdots & & \vdots \\ l_{F,1}\mathbf{R}_F & \dots & l_{F,K}\mathbf{R}_F \end{bmatrix}$$

Since each 2×3 $\tilde{\mathbf{M}}_{fk}$ sub-block for a generic frame f and basis k is a scaled rotation (truncated to dimension 2 for weak perspective projection) a 3×3 matrix $\mathbf{Q}_k$ (with $k = 1 \dots K$) can be computed to correct each vertical block $\tilde{\mathbf{M}}^k$ by imposing orthogonality and equal norm constraints on the rows of each $\tilde{\mathbf{M}}_{fk}$. Each $\tilde{\mathbf{M}}_{fk}$ block will contribute with 1 orthogonality and 1 equal norm constraint to solve for the elements in $\mathbf{Q}_k$.

Each vertical block is then corrected in the following way: $\left(\hat{\mathbf{M}}^k \leftarrow \tilde{\mathbf{M}}^k \mathbf{Q}_k\right)$. The overall $3K \times 3K$ correction matrix $\mathbf{Q}$ will therefore be a block diagonal matrix with the following structure:

$$\begin{bmatrix} \mathbf{Q}_1 & 0 & \dots & 0 \\ 0 & \mathbf{Q}_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{Q}_K \end{bmatrix}. \tag{4}$$

Unlike the method proposed by Bregler [6]—where the metric constraint was imposed only to the rigid component so that $Q_i = Q_{\text{rigid}}$ for $i = 1 \dots K$—this provides a corrective transform for each column-triple of $\tilde{\mathbf{M}}$. The 3D structure matrix is corrected appropriately using the inverse transformation: $\hat{\mathbf{S}} \leftarrow \mathbf{Q}^{-1}\tilde{\mathbf{S}}$.

A block diagonal matrix is in any case only an approximation of the true $\mathbf{Q}$ that is usually dense in the off-diagonal elements. It has been recently proved by Xiao et al. [25] that the above mentioned metric constraints do not suffice for the estimation of the full corrective transform. A new set of *basis constraints* is introduced, and their work proves that when both sets of constraints are imposed a closed-form solution exists to the problem of non-rigid structure from motion.

2.1.2. Factorization of the motion matrix M

The final step in the non-rigid factorization algorithm deals with the factorization of the motion matrix $\tilde{\mathbf{M}}$ into the 2×3 rotation matrices $\mathbf{R}_f$ and the deformation weights $l_{f,k}$. Bregler et al. [6] proposed a second factorization round where each motion matrix 2-row sub-block $\tilde{\mathbf{M}}_f$ is rearranged as an outer product of rotation parameters and deformation coefficients and then decomposed using a series of rank-1 SVD's. However, in the presence of noise the second and higher singular values of the sub-blocks do not vanish and this results in bad estimates for the rotation matrices and the deformation weights.

Brand proposed an alternative method to factorize each motion matrix 2-row sub-block $\tilde{\mathbf{M}}_f$ using orthonormal decomposition, which factors a matrix directly into a rotation matrix and a vector [4].

Each motion matrix sub-block $\tilde{\mathbf{M}}_f$ (see [5] for details) is rearranged such that

$$\tilde{\mathbf{M}}_f \rightarrow \hat{\mathbf{M}}_f = \begin{bmatrix} l_{f,1}\mathbf{r}_f^{\mathrm{T}} & l_{f,2}\mathbf{r}_f^{\mathrm{T}} & \ldots & l_{f,k}\mathbf{r}_f^{\mathrm{T}} \end{bmatrix} \tag{5}$$

where $\mathbf{r}_f=[r_{f1},\ldots r_{f6}]$are the coefficients of the rotation matrix $\mathbf{R}_f$. The motion matrix $\hat{\mathbf{M}}_f$ of size $6\times K$ is then post-multiplied by the $K\times 1$ unity vector $\mathbf{c}=[1\ldots1]$ thus obtaining

$$\mathbf{a}_f = k\mathbf{r}_f^{\mathrm{T}} = \hat{\mathbf{M}}_f\mathbf{c} \tag{6}$$

where $k = l_{f,1} + l_{f,2} + \ldots + l_{f,K}$ (the sum of all the deformation weights for that particular frame f). A matrix $\mathbf{A}_f$ of size 2×3 is built by re-arranging the coefficients of the column vector $\mathbf{a}_f$. The analytic form of $\mathbf{A}_f$ is:

$$\mathbf{A}_f = \begin{bmatrix} kr_1 & kr_2 & kr_3 \\ kr_4 & kr_5 & kr_6 \end{bmatrix}. \tag{7}$$

Since $\mathbf{R}_f$ is an orthonormal matrix, the equation $\mathbf{A}_f\mathbf{R}_f^{\mathrm{T}} = \sqrt{\mathbf{A}_f\mathbf{A}_f^{\mathrm{T}}}$ is satisfied, leading to $\mathbf{R}_f^{\mathrm{T}} = \sqrt{\mathbf{A}_f\mathbf{A}_f^{\mathrm{T}}}/\mathbf{A}_f$. This allows one to find a linear least-squares fit for the rotation matrix $\mathbf{R}_f$.

In order to estimate the configuration weights the sub-block matrix $\tilde{\mathbf{M}}_f$ is then rearranged in a different way from Eq. (5):

$$\tilde{\mathbf{M}}_f \rightarrow \bar{\mathbf{M}}_f = \left[l_{f,1}\mathbf{r}_f \ldots l_{f,k}\mathbf{r}_f\right]^{\mathrm{T}} \tag{8}$$

The configuration weights for each frame f are then derived exploiting the orthonormality of $\mathbf{R}_f$ since:

$$\bar{\mathbf{M}}_f\mathbf{r}_f^{\mathrm{T}} = \left[l_{f,1}\mathbf{r}_f\mathbf{r}_f^{\mathrm{T}} \ldots l_{f,k}\mathbf{r}_f\mathbf{r}_f^{\mathrm{T}}\right]^{\mathrm{T}} = 2\left[l_{f,1} \ldots l_{f,k}\right]^{\mathrm{T}}. \tag{9}$$

Brand included a final minimization scheme [4] in his *flexible factorization* algorithm: the deformations in $\tilde{\mathbf{S}}$ should be as small as possible relative to the mean shape. The idea here is that most of the image point motion should be explained by the rigid component. This is equivalent to the shape regularization used by other authors [1,23].

Although the methods described so far provide an estimate for the camera motion and the non-rigid shape, they fail to render the appropriate replicated structure of the motion matrix. In Section 3, we will describe a non-linear optimization scheme which allows to disambiguate between the motion and shape parameters.

3. Non-linear optimization for non-rigid structure from motion

Our approach is to reformulate the problem of estimating the non-rigid model parameters in terms of a non-linear minimization scheme of a geometrically meaningful cost function. The goal is to estimate the camera matrices $\mathbf{R}_i$ 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:

$$\min_{\mathbf{R}_i\mathbf{S}_k l_{i,k}} \sum_{i,j} \|\mathbf{x}_{ij} - \hat{\mathbf{x}}_{ij}\|^2$$

$$= \min_{\mathbf{R}_i\mathbf{S}_k l_{i,k}} \sum_{i,j} \|\mathbf{x}_{ij} - \left(\mathbf{R}_i \sum_k l_{i,k}\mathbf{S}_k\right)\|^2 \tag{10}$$

The non-linear optimization of the cost function is achieved using a Levenberg-Marquardt minimization scheme modified to take advantage of the sparse block structure of the matrices involved [24]. This method is generically termed bundle-adjustment in the computer vision and photogrammetry communities and it provides a maximum likelihood estimate assuming that the noise can be modeled with a Gaussian distribution. Levenberg-Marquardt [17] uses a mixture of Gauss–Newton and gradient descent minimization schemes switching from the first to the second when the estimated Hessian is close to being singular. Most of the computational burden is represented by the Gauss–Newton descent step, each iteration of which requires the calculation of the inverse of the Hessian of the cost function. Assuming local linearities the Hessian matrix $\mathbf{H}$ can be approximated as $\mathbf{H}=\mathbf{J}\mathbf{J}^{\mathrm{T}}$ (Gauss–Newton approximation) where Levenberg represents the Jacobian matrix in the model parameters. The size of $\mathbf{J}$ increases with the dimensionality of the model. This will render any implementation of a Levenberg-Marquardt minimization scheme too computationally expensive for the non-rigid factorization scenario, where the number of parameters in the model is particularly high. To alleviate this effect we reduce the computational cost by exploiting the sparse nature of the Jacobian matrix which is graphically represented in Fig. 1. The implementation details are described in Section 3.1.

3.1. Implementation

We have chosen to parameterize the camera matrices $\mathbf{R}_f$ using unit quaternions [10] giving a total of $4\times F$ rotation parameters, where F is the total number of frames. Quaternions ensure that there are no strong singularities and that the orthonormality of the rotation matrices is preserved

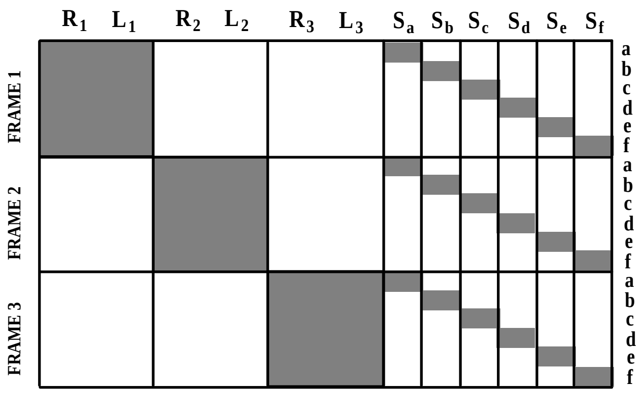


Fig. 1. Sparse structure of the Jacobian matrix for three frames, six points (*a*,*b*,*c*,*d*,*e*,*f*).

by merely enforcing the unitary norm of the 4-vector. This would not be the case with the Euler angle or the rotation matrix parameterizations, where orthonormality of the rotations is more complex to preserve. Indeed, in our initial implementation we parameterized the 3D pose using the six entries of the rotation matrices $\mathbf{R}_f$, 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 $3K \times P$ coordinates of the $\mathbf{S}_k$ shape bases and the $K \times F$ deformation weights l_{ik}.

A further critical factor is the choice of an initialization for the parameters of the model. It is crucial, for bundle adjustment techniques to work, that the initial estimate be close to the global minimum to increase the speed of convergence and reduce the chance of being trapped in local minima, particularly when the cost function has a large number of parameters as in this case.

We have chosen to use a similar initialization to the one used by Torresani et al. in their final tri-linear optimization scheme [23]. The idea is to initialize the camera matrices with the motion corresponding to the rigid component that is likely to encode the most significant part of the overall motion. This assumption is appropriate in the scenario of human facial motion analysis, but would evidently be inadequate for highly deformable objects such as a hand or the human body. The basis shapes were initialized using the values obtained from Brand's non-rigid factorization method, 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 more robust estimate of the camera rotation parameters and thus to a better convergence.

An alternative initialization which we have found to give a good starting point is to use the estimates given by Brand's algorithm for both motion and structure. Occasionally, however, we have observed problems with the convergence of this minimization and generally when the motion associated to the rigid component is used as the initial estimate the minimization reaches the minimum of the cost function in fewer iterations.

Note that a significant component of rigid motion is required to estimate the 3D structure. For a scenario with a nearly static subject, we would suggest a stereo factorization approach like [7] followed by an analogous non-linear refinement of the motion and shape components.

Occasionally, the non-linear optimization leads to a solution corresponding to a local minimum. In particular, at times we have found that 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 cannot change significantly from one frame to the next since the images are closely spaced in time. This is implemented by 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 the relief present in the 3D data is preserved. Similar regularization terms have also been reported in [1,23].

The work presented here is most closely related to the work by Aanæs and Kahl, who also proposed a bundle adjustment solution for the non-rigid scenario [1]. However, their approach differs in some fundamental aspects. Firstly, their initial estimate of the non-rigid shape was obtained by estimating the mean and variance of the 3D data obtained directly from image measurements. The approach assumes that the cameras are calibrated, and although the authors state that their algorithm would work in the uncalibrated case 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. We are using quaternions instead which, as will be shown in Section 5, leads to better behaved results for the motion estimates.

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 actually possible to decouple motion and deformation parameters (see Section 5 for a detailed description).

4. Tracking points with ranklets

We generate 2D point tracks for our reconstruction algorithm by means of an appearance-based tracking algorithm built on ranklets [20]. Ranklets appear to be particularly suited for this task because, being rank features, they are invariant to monotonic intensity transformations. This gives some robustness to the illumination changes caused by the 3D rotations and deformations of the tracked object. Also, ranklets display an orientation selectivity pattern similar to Haar wavelets, which has been shown to improve their discriminative power as compared to other rank features. The definition of ranklets is given in the Appendix A; for a more detailed description, we refer the reader to [20]. For the purpose of understanding the tracking algorithm, it will be sufficient to think of ranklets as a family of orientation selective, multi-scale (non-linear) filters.

The fiducial points used for 3D reconstruction are automatically selected based on a saliency criterion (specified below), and subsequently tracked throughout the image sequence.

4.1. Feature selection with adaptive appearance-based modeling

In analogy with classic approaches involving multi-scale features [15,18], we choose to encode the local image neighborhood of each point by means of a vector of ranklets consisting of a total of nine filters arranged in three frequency channels and three 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 Fig. 11, we decided to track 110 points).

For each tracked point, an optimization step is performed to adapt the shape of the filters to the specific appearance of the neighborhood of the point. 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 purpose of adaptation step is maximizing the saliency of the tracked location across the local neighborhood, thus facilitating tracking. The support of a few adapted filters is shown in Fig. 2.

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.

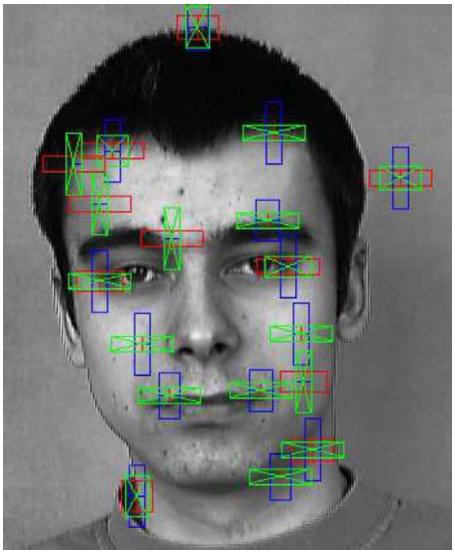


Fig. 2. Results of filter adaptation for a few tracked points (lowest frequency channel). The rectangles represent the support of the filters. The orientation selectivity of each filter is indicated by a horizontal, vertical or diagonal line drawn across the corresponding rectangle.

4.2. Updating the models

Due to the deformations and pose changes 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. The model update problem has been studied extensively in the case of image domain algorithms, in which all appearance variations directly affect the quality of the match (see [3,9] and related work). In our case, the feature space representation affords a degree of invariance that makes this problem less critical, so that a simplified but flexible approach (in some ways related to the ideas in [16]) is sufficient.

The solution we adopted involves maintaining a stack model for each tracked point. A new model is acquired from the current best estimate of the position of a point whenever the residual distance from the original model (after matching) exceeds a threshold τ. The filter adaptation step is repeated and the new model is stored on the stack above the previous one. This procedure is repeated when necessary up to a given maximum number of models, after which the particular point is considered lost.

While tracking each point the most recently acquired model, which is on top of the stack, is used first. A further gradient descent is then initiated in an attempt to match the previous model on the stack. If the resulting discrepancy is now below τ the last model is discarded by popping the stack, and the point is assumed to have recovered the appearance it had at an earlier time. The algorithm then attempts to work its way further down the stack by iterating the procedure. In this way, the active model is always the oldest acquired model for which the matching distance does not exceed τ.

The stack model provides a mechanism for tracking a point across a range of deformations and pose variations, during which the point may occasionally revert to its original appearance. 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 (if this is different from the active model). The contents of the stack above the grafting position are then discarded. 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 carried out with two main objectives in mind:

- To compare the performance of each of the two new proposed algorithms (the tracker and the bundle adjustment based 3D reconstruction) with existing techniques.

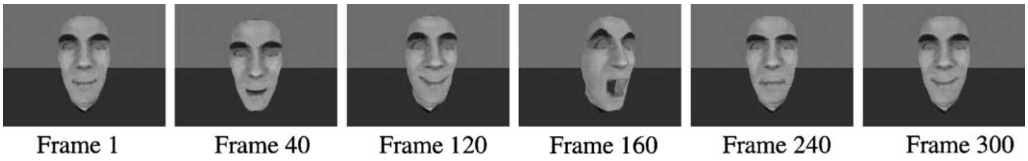


Fig. 3. The key frames show a subset of the deformations present in the synthetic face sequence. The generated shape combines simultaneous rotation and deformations throughout the 300 frame long sequence.

- To show the results obtained with the complete system—automatic ranklet-based tracking followed by the bundle adjustment based 3D reconstruction.

First, we have compared the performance of the ranklet-based tracking algorithm with the classic Kanade, Lucas and Tomasi (KLT) tracker. Secondly, we have compared the 3D reconstructions obtained with our new bundle-adjustment algorithm with those obtained with an algorithm similar to Brand's. Finally, we have demonstrated the complete system on a real sequence of a human subject performing different facial expressions.

5.1. Evaluation of the ranklet-based tracker on synthetic data

Firstly, we test the performance of the tracker compared to the well-known Kanade, Lukas and Tomasi (KLT) algorithm [14] that is the standard approach for image point registration (an implementation is freely available online[1]). The KLT tracker computes the location of selected image points in each consecutive frame by estimating an affine warping of the image patch surrounding each point from the residual of the image intensities, in a Newton–Raphson style minimization (see [19] for more details).

We compare both approaches using a 300 frame sequence (see Fig. 3) of a deforming synthetic face created with an open source software for ray-tracing[2]. The facial expressions and the rotation/translation components are computed using a 3D model that is subsequently projected with an affine camera. The sequence is created in such a way that occlusions are completely avoided over the 194 points selected for tracking. This selection must allow a fair comparison of both tracking algorithms. Indeed, each algorithm has a different saliency criterion and would thus select a different subset of points to track. To make our experiment as unbiased as possible, rather than allowing each algorithm to select the tracked points according to its saliency criteria, we manually selected an arbitrary set of points from the projected 3D model for which ground truth information is available.

The tests are carried out under different image noise levels in order to compare the robustness of the two methods. For each frame, white Gaussian noise of a given variance is added to the image grayscale values (normalized between 0 and 1). The effect of the different noise levels on the original images is shown in Fig. 4 along with the ground truth for the position of the selected points. The configuration parameters for each algorithm were set to prevent the occurrence of lost tracks to allow the comparison of trackers with different point rejection criteria.

Tracking accuracy is then measured using the root mean square (RMS) value of the discrepancy between the estimated position of the points and the ground truth. The performance of our ranklet-based tracker is equivalent to that of the KLT in the ideal noise-free and for almost noise-free conditions (Fig. 5), with the RMS error for both trackers being below 2 pixels at the end of the sequence. However, as the noise level increases the accuracy of the KLT algorithm degrades much more quickly than that of our algorithm, leading to a final RMS error of 10 pixels for KLT, versus 4 pixels for our algorithm.

It must be noted that our ranklet-based tracker is in many ways less sophisticated than the implementation of the KLT algorithm we used. The latter features a matching stage using 2D affine transformations and subpixel matching, which are all absent in our algorithm. Thus, the better performance of our algorithm is remarkable, and must be ascribed to the robustness of ranklets as image descriptors. On the other hand, it would be entirely feasible to extend our algorithm with subpixel matching and an approximation to affine matching, such as the one used in [13] in the case of Haar wavelets. This would likely lead to further improvements.

5.2. Bundle adjustment with manually tracked points

In this section, we compare the results obtained with our bundle-adjustment based 3D reconstruction algorithm with those obtained using Brand's non-rigid factorization method. We used a real video test sequence which shows the face of a subject performing an almost rigid motion for the first 200 frames, moving his head up and down. The subject then changed facial expression with his head facing front for the next 309 frames (see Fig. 6). The point features which appear in Fig. 6 were manually marked throughout the sequence.

The results of the 3D reconstructions for some key frames in the sequence obtained using Brand's factorization method are shown in Fig. 7. The front views of the 3D reconstruction show that the recovered 3D shape does not reproduce the facial expressions very accurately. Besides, depth estimation is not very accurate, which is evident by inspection of the top views of the reconstruction. Notice the asymmetry of the left and right side of the face.

[1] http://www.ces.clemson.edu/~stb/klt/
[2] http://www.povray.org/

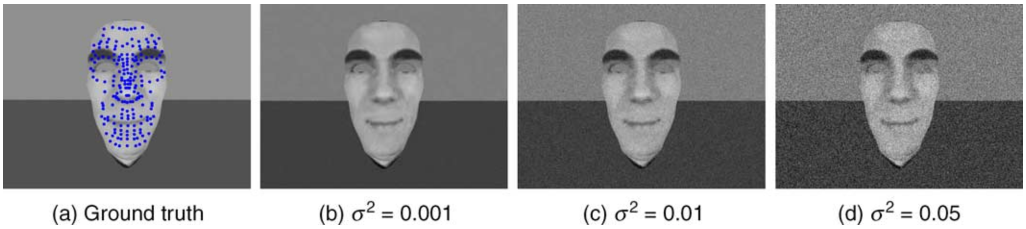

(a) Ground truth (b) $\sigma^2 = 0.001$ (c) $\sigma^2 = 0.01$ (d) $\sigma^2 = 0.05$

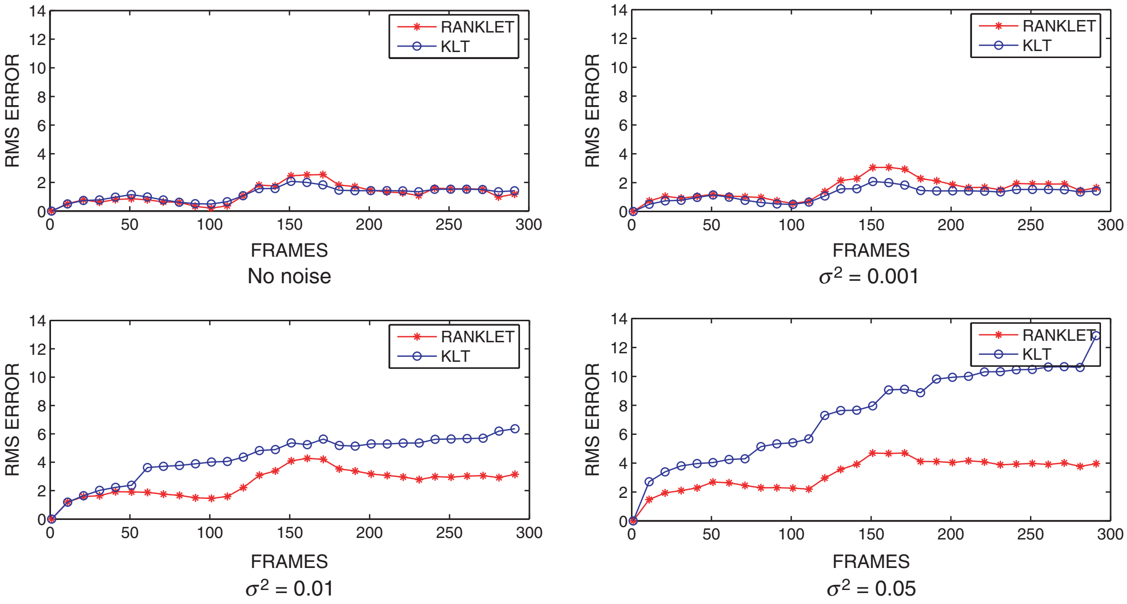
Fig. 4. (a) Ground thruth for the set of points selected for tracking. Figures (b–d) Original image with added white Gaussian noise of variance σ^2. Image greyscale values are normalized between 0 and 1.

In Fig. 8, we show the reconstructed 3D shape recovered 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 (frame 467) or the opening of the mouth in surprise (frame 358). Finally, the top views show that the overall relief appears to be well preserved, as is the symmetry of the face.

The evolution of the weights $l_{f,i}$ of the deformation modes introduced in Eq. (1) can be traced throughout the sequence. In Fig. 9, we show the value of the weight associated with the rigid component (top) and of those associated with the four deformation modes (middle). Results are given for both Brand’s flexible factorization (left) and for the bundle adjustment scheme (right). Notice how Brand’s flexible factorization has a tendency to suppress weak deformations occurring in the subject—the weights associated with the deformation modes have a small value. This results in the recovered 3D shape not reproducing the facial expressions accurately. The weights associated with the deformation modes have higher values in the bundle-adjusted solution. Interestingly, around frame 360 the first non-rigid mode of deformation experiences a large peak, which corresponds to the opening of the mouth in surprise as shown in Fig. 6. This indicates some tendency in the configuration weights to reflect the underlying facial expressions. Although this peak is present also in Brand’s solution, the corresponding 3D reconstruction in Fig. 7 is not very accurate.

The results obtained for the motion parameters are shown in the bottom graph of Fig. 9. The rotation angles around the *X*-, *Y*- and *Z*-axes (up to an overall rotation) are recovered for each of the 509 frames in the sequence. In particular, the tilt angle varied smoothly throughout the first 200 frames capturing the up and down tilt of the head of about 50° in total while the rotation angles around the other 2 axes did not vary significantly throughout the sequence. Notice that both solutions capture this motion correctly. However, the results obtained with the bundle-adjusted solution (right) are smoother than those obtained using Brand’s algorithm (left).

The non-linear refinement step is initialized using the values of the weight associated with the rigid component—scale—and the corresponding rotation angles shown in Fig. 10. It can be observed from the plot that the rigid component of the motion is a good description of the object’s rotation, and in fact the bundle-adjustment step does not optimize these parameters



No noise
$\sigma^2 = 0.001$
$\sigma^2 = 0.01$
$\sigma^2 = 0.05$

Fig. 5. compare the performance of the ranklet tracking method with the KLT tracker for different levels of noise.

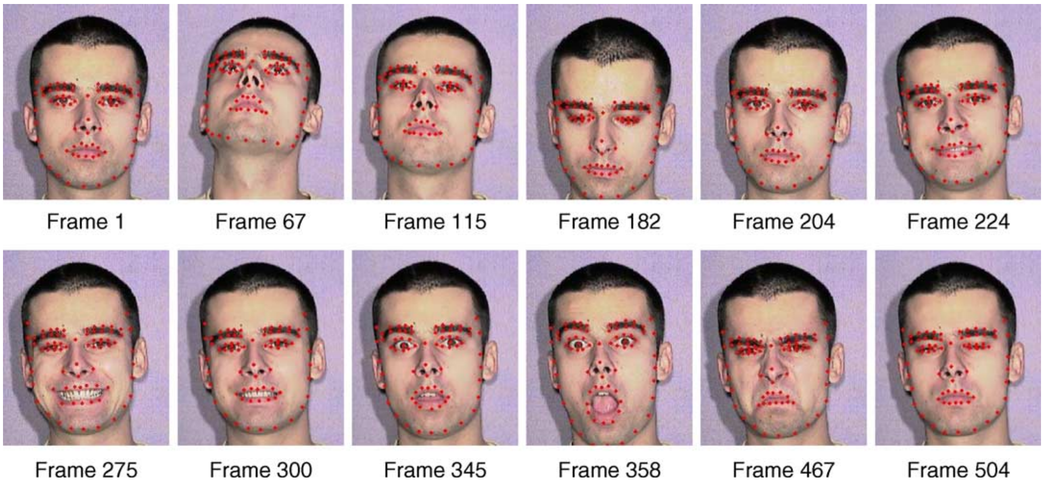


Fig. 6. Key frames of the sequence used in the experiments in Section 5.2, with manual points superposed. The subject performed an almost rigid motion for the first 200 frames moving the head sideways and then changed facial expression for the next 309 frames.

much further. The additional bundle adjustment step simultaneously refines the initial values and modifies the morph weights to model the motion and deformation of the object.

5.3. Testing the combined system

In this section, we describe a combined test of the integrated system in which the ranklet-based tracker automatically generates the tracks that are input into the non-linear optimization algorithm, to yield a 3D reconstruction. The system has to cope with a complex 960 frame sequence in which the subject is performing at the same time 3D motion and different facial expressions.

A total of 91 points were initialized automatically according to the saliency criterion described in Section 1. As can be seen in Fig. 2, the tracker was able to follow a good number of feature points reliably throughout the sequence, even in relatively poorly textured areas such as the subject's cheekbones. Throughout the 960 frame sequence, only eight points out of the initial 91 were lost, showing that the tracker can cope with significant deformations and pose changes. However, a certain number of points initialized on homogenous texture turned out to be unreliable, and they evidently affect the 3D shape estimation (Fig. 11).

Fig. 12 describes the operation of the model update mechanism described in Section 2. For each key frame, a histogram of the depth of the model stack across all tracked points is presented. As can be seen, more models are acquired for a large number of points in the presence of large deformations. The points then revert to the original models as the subject recovers its original appearance. The combined effect of the model updating technique and backtracking phases is to allow the tracker to follow the points through a wide range of deformations, while at the same time limiting drift.

We present in Fig. 13 the front, top and side views of the 3D reconstruction of six key frames with different expressions.

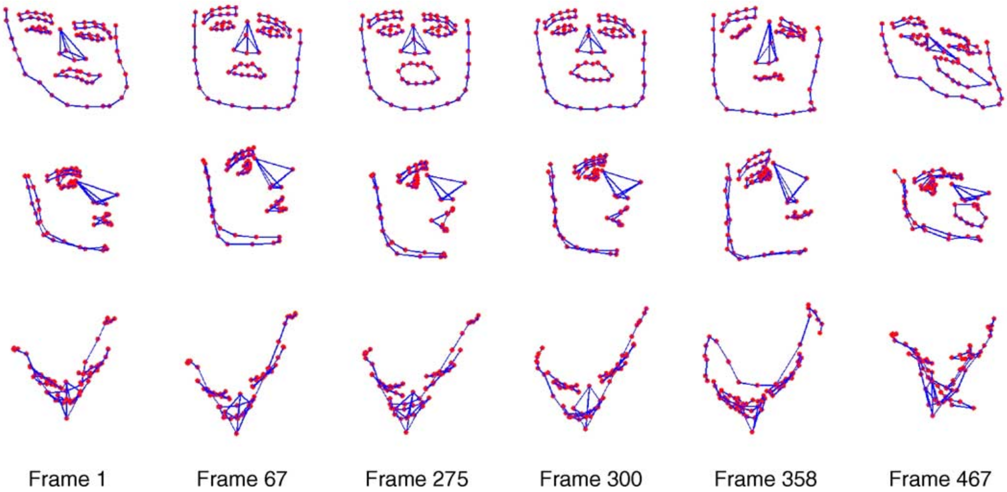


Fig. 7. Front, side and top views of the 3D reconstructions obtained from the non-rigid factorization algorithm without bundle adjustment for some of the key frames in the sequence.

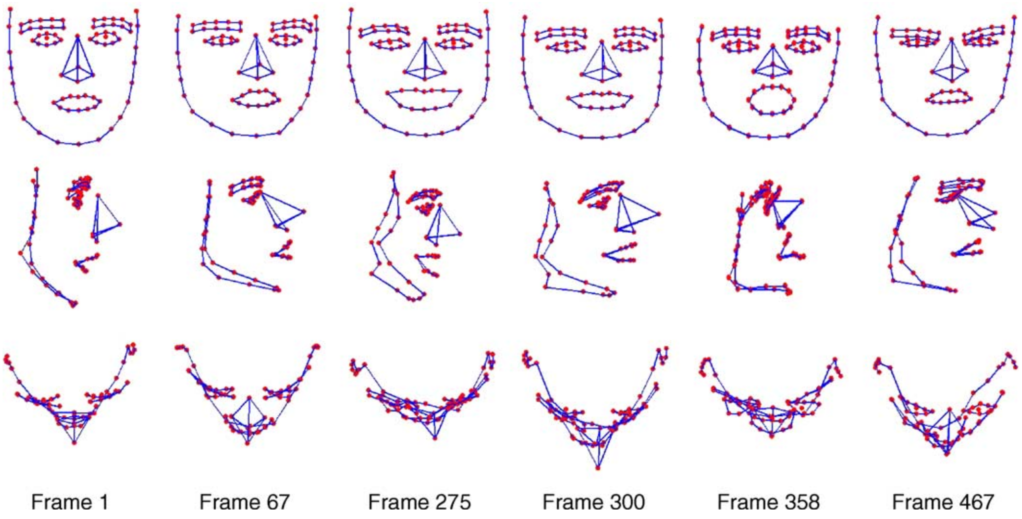
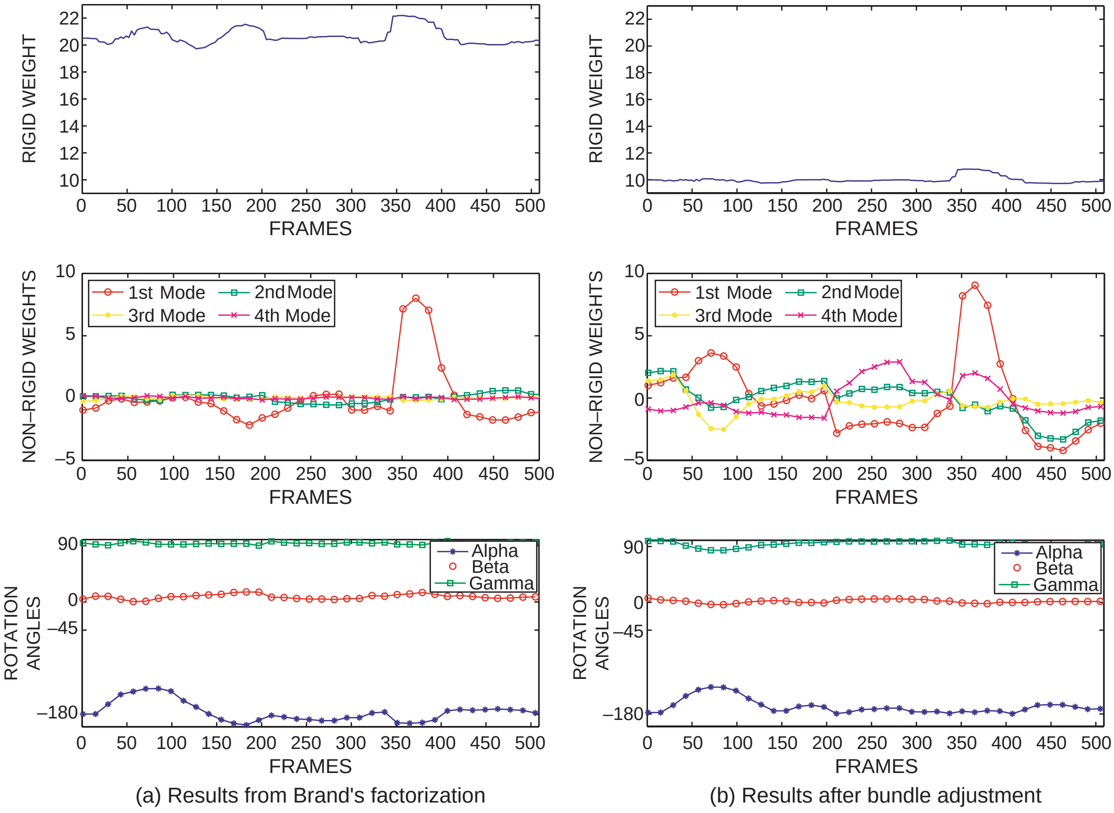


Fig. 8. Front, side and top views of the 3D reconstructions after bundle adjustment for some of the key frames in the sequence.

The number of basis shapes is fixed to $K = 8$ and the initialization of the non-linear optimization is identical to the one described in Section 3.1. The overall depth is generally correct: notice the points belonging to the neck relative to the position of the face, and the nose pointing out from the face plane. Face symmetry is generally well preserved and can be seen in the top views of the reconstruction. Outliers are evident in frame 710 in the eyebrow region and generally on the neck area where the tracker performs poorly; such feature points are wrongly reconstructed by our non-rigid model.

Finally, the reconstructed motion and deformation parameters are displayed in Fig. 14. The estimated angles reproduce the rotation of the subject's head reasonably well, with values between 10° and 15° for the 'beta' angle, while 'alpha' and 'gamma' show tiny variations. The rigid weight is nearly constant for the whole sequence in accordance with the subject's



(a) Results from Brand's factorization

(b) Results after bundle adjustment

Fig. 9. Values obtained for the rigid component (top), deformation weights (middle) and rotation angles (bottom) using Brand's approach (A) and bundle adjustment (B) for the sequence in Fig. 6. Bundle adjustment provides smoother and better behaved solutions for all the parameters.

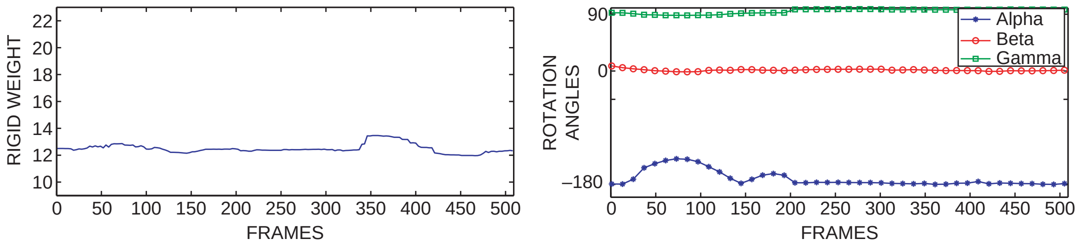


Fig. 10. Values for the initialization of the non-linear minimization algorithm. The values obtained for the rigid component (left) and the rotation angles (right) are computed with the motion corresponding to the rigid component.

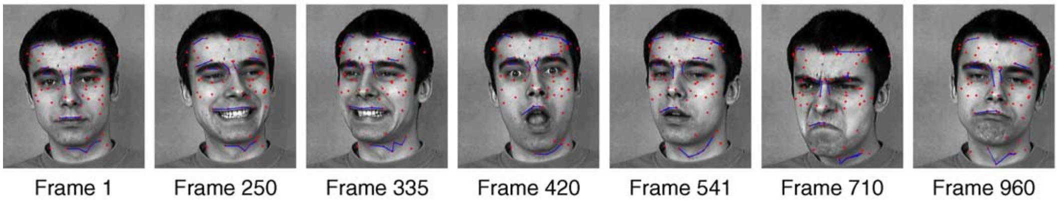


Fig. 11. Key frames in the sequence used to test the complete system. The subject performed simultaneous rigid and non-rigid motion. Automatically tracked points are superimposed. A set of wireframes outlines the face structure.

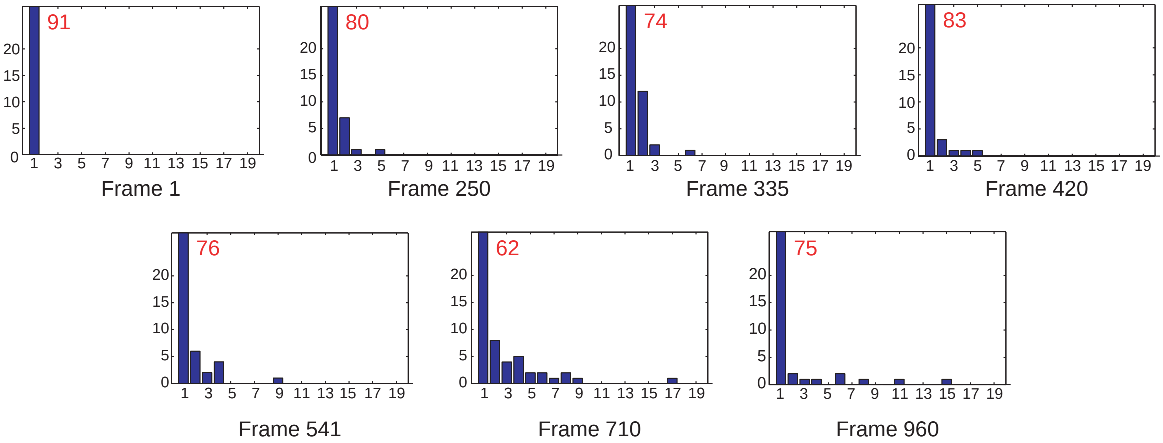


Fig. 12. Operation of the model update mechanism across the sequence of Fig. 11. The histograms show how many points (y-axis) have how many models (x-axis) on their stack. Notice how new models are added to accommodate large deformations (frame 710); by frame 960 most points have reverted to their original model, and so has the appearance of the subject.

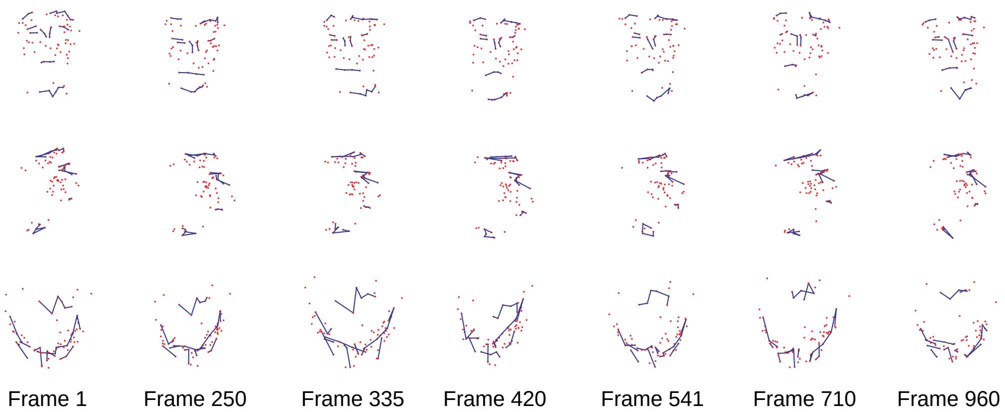


Fig. 13. Front, side and top views of the 3D reconstructions obtained by the combined system for some of the key frames in the sequence.

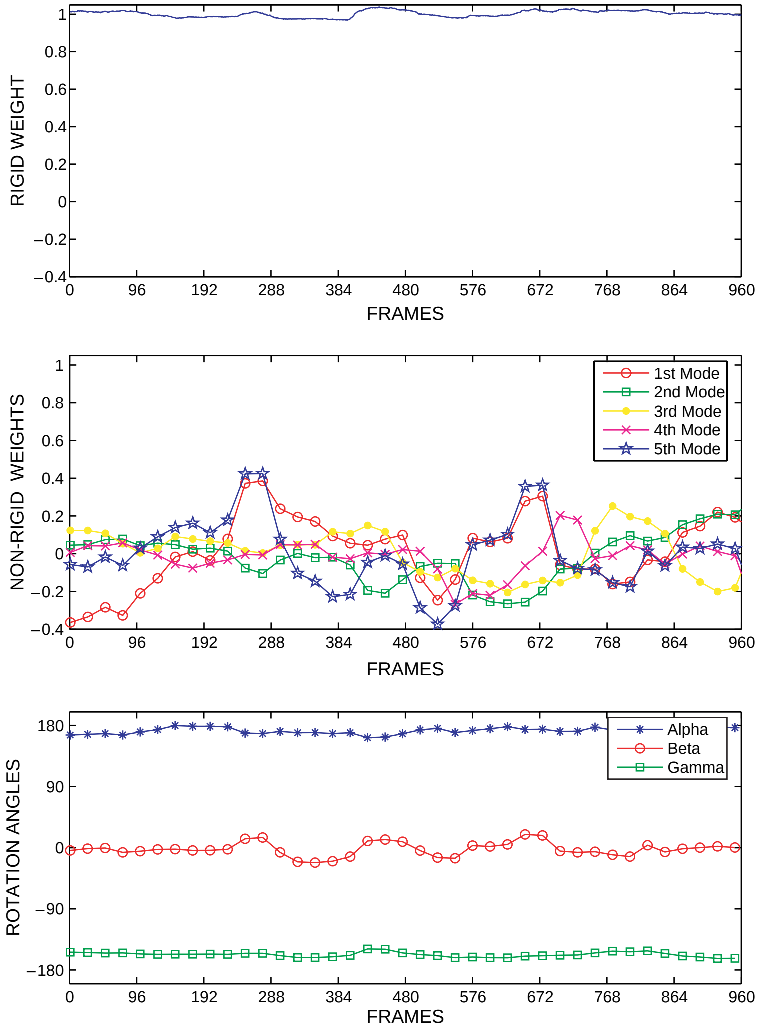


Fig. 14. Evolution of the rigid weight ($K=1$), the first five non-rigid weights ($K=2,\ldots,6$) and the rotation angles (in degrees) throughout the sequence of Fig. 11.

head being at a constant distance from the camera. The non-rigid configuration weights present more erratic behavior; the two evident spikes around frames 280 and 670 correspond, respectively, to the grin and anger facial expressions.

6. Summary and conclusions

In this work, we introduced a novel complete system for non-rigid structure from motion analysis, starting from unannotated video sequences. The system comprises an adaptive point-tracking algorithm based on ranklets, that feeds into a novel non-linear optimization method for structure from motion.

The tracking algorithm we have introduced uses ranklets to build a feature space representation of a neighborhood of each tracked point. As with all other rank features, the representation has a high degree of invariance to the illumination and appearance changes resulting from 3D rotations and deformations of the object. In addition, the orientation selectivity of ranklets contributes to the discriminative power of the representation, thus minimizing point drift. This effect is enhanced by locally adapting the aspect ratio of the filters to the appearance of the tracked points, thus making the feature extraction process point-specific.

Large variations in appearance are handled by a model stack mechanism, which effectively keeps track of the feature space trajectory of the single points. By adaptively storing new models on the stack when deformations make it necessary and reverting to older models when the quality of the match allows it, the algorithm is able to cope with significant appearance variations without loosing track. Precision is gradually recovered as the points revert to their original appearance.

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.

In our experiments, the tracking algorithm has showed its ability to handle the deformations and pose changes of a non-rigid object effectively. By integrating it with our factorization 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 the 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. Enrique Muñoz, José Miguel Buenaposada and Luis Baumela provided the code for the synthetic 3D face model. Lukas Zalewski provided the sequences and the manual tracks used for the reconstruction. This work was partially supported by EPSRC Grant GR/S61539/01. Alessio Del Bue holds a Queen Mary Studentship Award.

Appendix A. Ranklets

Ranklets are a family of orientation selective rank features designed in close analogy with Haar wavelets. However, whereas Haar wavelets are a set of filters that act linearly on the intensity values of the image, ranklets are defined in terms of the relative order of pixel intensities [20].

Ranklets are defined starting from the three Haar wavelets $h_i(x)$, $i = 1,2,3$ shown in Fig. 15, supported on a local window **S**. Given the counter images of $\{+1\}$ and $\{-1\}$ under $h_i(x)$, $\mathbf{T}_i = h_i^{-1}(\{+1\})$ and $\mathbf{C}_i = h_i^{-1}(\{-1\})$, our aim with ranklets is to perform a non-parametric comparison of the relative brightness of these two regions.

A straightforward non-parametric measure of the intensity of the pixels in $\mathbf{T}_i$ compared to those in $\mathbf{C}_i$ can be obtained as follows. Consider the set $\mathbf{T}_i \times \mathbf{C}_i$ of all pixel pairs $(\mathbf{x},\mathbf{y})$ with $\mathbf{x} \in \mathbf{T}_i$ and $\mathbf{y} \in \mathbf{C}_i$. Let $\mathcal{W}^i_{YX}$ be the number of such pairs in which the pixel from the set $\mathbf{T}_i$ is brighter than the one from $\mathbf{C}_i$, that is:

$$\mathcal{W}^i_{YX} = \#\{(\mathbf{x},\mathbf{y}) \in \mathbf{T}_i \times \mathbf{C}_i | I(\mathbf{y}) < I(\mathbf{x})\}. \tag{11}$$

Essentially, $\mathcal{W}^i_{YX}$ will be close to its maximum value, i.e. the number of pairs in $\mathbf{T}_i \times \mathbf{C}_i$, if the pixels in the $\mathbf{T}_i$ region are brighter than those in the $\mathbf{C}_i$ region; conversely, it will be close to its minimum value (i.e. 0) if the opposite is true. Remembering that the $\mathbf{T}_i$ and $\mathbf{C}_i$ sets coincide by definition with the '$+1$' and '-1' regions of the wavelets in Fig. 15, we see that each $\mathcal{W}^i_{YX}$ displays the same orientation selective response pattern as the corresponding Haar wavelet h_i.

The procedure outlined above would of course be very inefficient to carry out directly, as it would require processing $N^2/4$ pixel pairs, where N is the number of pixels in the support window **S**. Luckily a more efficient algorithm exists, with complexity of at most $O(N \log N)$. It will suffice to sort all the pixels in **S** according to their intensity. Indicate the rank of pixel **x** with $\pi(\mathbf{x})$; we then have

$$\mathcal{W}^i_{YX} = \sum_{\mathbf{x} \in T_i} \pi(\mathbf{x}) - (N/2 + 1)N/4. \tag{12}$$

(for a proof, see [12]). The quantity $\mathcal{W}^i_{YX}$ is known as the Mann–Whitney statistics for the observables (the pixels) in $\mathbf{T}_i$ and $\mathbf{C}_i$ (according to the standard terminology, these would be the 'Treatment' and 'Control' sets). The Mann–Whitney statistics is equivalent to the Wilcoxon statistics W_s [12].

For practical reasons, it is convenient to define ranklets as

$$\mathcal{R}^i = 2\frac{\mathcal{W}^i_{YX}}{N^2/4} - 1, \tag{13}$$

so that their value increases from -1 to $+1$ as the pixels in $\mathbf{T}_i$ become brighter than those in $\mathbf{C}_i$.

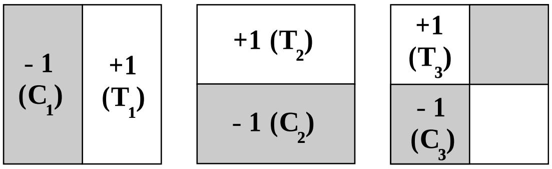


Fig. 15. The three 2D Haar wavelets $h_1(\mathbf{x})$, $h_2(\mathbf{x})$ and $h_3(\mathbf{x})$ (from left to right). Letters in parentheses refer to the **T** and **C** pixel sets defined in the text.

References

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

[2] D.N. Bhat, S.K. Nayar, Ordinal measures for visual correspondence, in: Proceedings of the CVPR, 1996, pp. 351–357.

[3] M. Black, A. Jepson, Eigen-tracking: robust matching and tracking of articulated objects using a view-based representation, International Journal of Computer Vision 36 (2) (1998) 63–84.

[4] M. Brand, Morphable models from video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, 2001 (2) 456–463.

[5] M. Brand, R. Bhotika, Flexible flow for 3d nonrigid tracking and shape recovery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, 2001, pp. 315–322.

[6] C. Bregler, A. Hertzmann, H. Biermann, Recovering non-rigid 3d shape from image streams, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina, 2000, pp. 690–696.

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

[8] J.G. Daugman, Two-dimensional spectral analysis of cortical receptive field profiles, Vision Research 20 (1980) 847–856.

[9] G.D. Hager, P.N. Belhumeur, Efficient region tracking with parametric models of geometry and illumination, IEEE Transactions on PAMI 20 (10) (1998) 1025–1039.

[10] B.K.P. Horn, Closed form solutions of absolute orientation using unit quaternions, Journal of the Optical Society of America A 4 (4) (1987) 629–642.

[11] M. Irani, Multi-frame optical flow estimation using subspace constraints, in: Proceedings of the Seventh International Conference on Computer Vision, Kerkyra, Greece, 1999 626–633.

[12] E.L. Lehmann, Nonparametrics: Statistical Methods Based on Ranks, Holden-Day, San Francisco, CA, 1975.

[13] A.P. Leung, S. Gong, An optimization framework for real-time appearance-based tracking under weak perspective, In Proc. British Machine Vision Conference, Oxford, UK 2005.

[14] Bruce D. Lucas, Takeo Kanade, An iterative image registration technique with an application to stereo vision, in: International Joint Conference on Artificial Intelligence, 1981, pp. 674–679.

[15] B.S. Manjunath, C. Shekhar, R. Chellappa, A new approach to image feature detection with applications, Pattern Recognition 31 (1996) 627–640.

[16] I. Matthews, T. Ishikawa, S. Baker, The template update problem in: R. Harvey, J.A. Bangham (Eds.), Proceedings of the British Machine Vision Conference, Norwich, UK, vol. II (2003), pp. 649–658.

[17] J.J. Moré, The Levenberg-Marquardt algorithm: implementation and theory in: G.A. Watson (Ed.), Numerical Analysis, Springer, Berlin, 1977, pp. 105–116. Lecture Notes in Mathematics 630.

[18] R.P.N. Rao, D.H. Ballard, An active vision architeture based on iconic representations, Artificial Intelligence Journal 78 (1995) 461–505.

[19] Jianbo Shi, Carlo Tomasi, Good features to track, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94), Seattle, 1994, pp. 593–600.

[20] F. Smeraldi, Ranklets: orientation selective non-parametric features applied to face detection, in: Proceedings of the 16th ICPR, Quebec, QC, vol. 3, 2002, pp. 379–382.

[21] F. Smeraldi, A nonparametric approach to face detection using ranklets, in: Proceedings of the 4th International Conference on Audio and Video-based Biometric Person Authentication, Guildford, UK June 2003 pp. 351–359

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

[23] L. Torresani, D. Yang, E. Alexander, C. Bregler, Tracking and modeling non-rigid objects with rank constraints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, 2001.

[24] Bill Triggs, Philip McLauchlan, Richard Hartley, Andrew Fitzgibbon, Bundle adjustment—a modern synthesis in: W. Triggs, A. Zisserman, R. Szeliski (Eds.), Vision Algorithms: Theory and Practice, LNCS, Springer, Berlin, 2000, pp. 298–375.

[25] Jing Xiao, Jinxiang Chai, Takeo Kanade, A closed-form solution to non-rigid shape and motion recovery, in: The Eigth European Conference on Computer Vision 2004 573–587.

[26] Jing Xiao, Takeo Kanade, Non-rigid shape and motion recovery: degenerate deformations, in: IEEE Conference on Computer Vision and Pattern Recognition, 2004 668–675.

[27] R. Zabih, J. Woodfill, Non-parametric local transforms for computing visual correspondence, in: Proceedings of the Third ECCV, 1994, pp. 151–158.