J Math Imaging Vis
DOI 10.1007/s10851-007-0060-3

Implicit Non-Rigid Structure-from-Motion with Priors

S.I. Olsen · A. Bartoli

Abstract This paper describes an approach to implicit Non-Rigid Structure-from-Motion based on the low-rank shape model. The main contributions are the use of an implicit model, of matching tensors, a rank estimation procedure, and the theory and implementation of two smoothness priors. Contrarily to most previous methods, the proposed method is fully automatic: it handles a substantial amount of missing data as well as outlier contaminated data, and it automatically estimates the degree of deformation. A major problem in many previous methods is that they generalize badly. Although the estimated model fits the visible training data well, it often predicts the missing data badly. To improve generalization a temporal smoothness prior and a surface shape prior are developed. The temporal smoothness prior constrains the camera trajectory and the configuration weights to behave smoothly. The surface shape prior constrains consistently close image point tracks to have similar implicit structure. We propose an algorithm for achieving a Maximum A Posteriori (MAP) solution and show experimentally that the MAP-solution generalizes far better than the prior-free Maximum Likelihood (ML) solution.

Keywords Computer vision · Structure-from-motion · Non-rigid · Low-rank shape

1 Introduction

Non-rigid Structure-from-Motion concerns the simultaneous recovery of the deforming world structure and camera motion from image features. This extends the classical rigid Structure-from-Motion [13] to situations with deforming scenes such as expressive faces, deforming bodies *etc*. A major step forward was made by Bregler *et al.* [5], Brand [3] and Aanæs *et al.* [1]. They represent non-rigidity as a linear combination of a small number of 3D *basis shapes*. The *low-rank shape* model is generic: it does not prescribe any particular type of 3D shape or deformation. Using only limited assumptions it allows a simultaneous recovery of 3D deformable shape and camera motion from monocular videos. Xiao *et al.* [23] studied the deformations that may defeat the reconstruction algorithms. Apart from [1, 2], most methods assume that the amount of non-rigidity—the number l of basis shapes—is known. If l is underestimated the deformation cannot be well modeled, and if overestimated the model will contain too many parameters. In the latter case the model will fit the noise in the data, and will not generalize well.

The data used in this and most previous methods consist of point coordinates obtained by tracking image interest points through a sequence. Because of occlusions and imperfect tracking the registered data often is partial: some or all points are visible only in a subset of the frames. Many early methods could not handle situations with missing data

This paper combines and extends two conference papers; the first one appeared in the Workshop on Dynamical Vision held at ICCV'05 [2], and the second one appeared at the 2007 British Machine Vision Conference [11]. This paper integrates the two publications into a comprehensive presentation of our approach to Non-Rigid Structure-from-Motion.

S.I. Olsen (✉)
Department of Computer Science, University of Copenhagen, Universitetsparken 1, 2100 Copenhagen Ø, Denmark
e-mail: ingvor@diku.dk

A. Bartoli
LASMEA (CNRS / UBP), Clermont-Ferrand, France
e-mail: adrien.bartoli@gmail.com

[1, 3, 5, 16, 21, 22]. Recently a number of methods [6, 10] have been proposed to handle the missing data problem for the rigid Structure-from-Motion problem.

Estimating a model from partial data allows one to predict the projection of all world points on all images. The model generalizes well if the predicted points, on frames where the point is not registered, are accurate. If the degree of deformation is overestimated the model is unlikely to generalize well. Priors have been shown to improve generalization. In [7] a prior for rigidity was presented. In [18] a probabilistic PCA model using hierarchical priors is applied to avoid overfitting.

The present paper draws on and extends our previous work [2, 11]. It gives an in-depth description of the implicit Maximum-Likelihood (ML) approach to non-rigid Structure-from-Motion [2], and its extension with temporal and shape smoothness priors [11], by which a Maximum A Posteriori (MAP) solution is formulated. The proposed MAP-estimator is based on four main steps: an initial solution is computed by using an ML-estimator minimizing the reprojection error. Second, the implicit coordinate frame is changed to maximize a temporal smoothness prior. Third, the implicit structure is re-estimated by minimizing a combination of the reprojection error and a surface shape prior. Finally, the motion and structure estimates are jointly refined by nonlinear optimization. The paper reports results on simulated and real data. It shows that the generalization ability of the MAP solution is greatly improved compared to the standard ML-estimation. Experiments show that tracks split by imperfect tracking can be glued correctly together.

Contributions Among the various methods using the low-rank shape model for non-rigid Structure-from-Motion, our framework is the first one to bring all of the following features:

- *Missing image points.* Most of the other methods are based on SVD to factorize a measurement matrix, *e.g.* [1, 3–5, 7, 17], and thus do not deal with missing image points. Our framework handles cases for which only a few percents of the image points are observed (for instance, we successfully handle a sequence with only about 13% of the image points being observed), meaning that extreme occlusions and highly dynamical scenes can be handled. This is achieved thanks to the low-rank matching tensors and closure constraints we propose.
- *Robustness.* Most of the other methods assume that the noise on the image point positions follows a centered Gaussian *i.i.d.* distribution, *e.g.* [1, 4, 5, 7, 18]. While this might be a sensible assumption if the points are manually clicked or at least checked by the user, this certainly is not true if the output of an automatic KLT-like point tracker is directly used as input. The points may drift from the ideal track, and may also be totally mismatched. Our algorithm outputs, for each image point, a binary variable indicating if it is an inlier or an outlier with respect to the low-rank shape model.
- *Rank selection.* Most of the other methods assume that the rank, *i.e.* the degree of deformation, is known, *e.g.* [4, 5, 7, 18]. This is definitely not a realistic assumption, since the rank is highly dependent on the scene content. Inspired by the GRIC model selection criterion, a robustified BIC, our algorithm computes the rank from the available data automatically.
- *Generic prior knowledge.* Most other methods assume the low-rank shape model as the only generic prior, *i.e.* the scene shape deforms according to a finite set of 'few' deformation modes, *e.g.* [1, 5, 7]. This clearly is not enough to obtain a model that will generalize well to the entire sequence when only a small fraction of the data is available. Natural generic priors such as smooth camera motion, shape deformation and continuous surface shapes, are easily included in our framework. We show that the high generalization ability of the recovered model allows us to glue point tracks split during tracking, due to *e.g.* an occlusion or a tracking failure.

Our framework is entirely automatic, as it takes as input the point tracks produced by some point tracker, computes the rank, classify each image point as valid or erroneous, and outputs the sought after implicit reconstruction. A further step is to upgrade the implicit reconstruction to an explicit, *i.e.* metric one, which has been described in details in several recent papers [4, 22].

Organization of the Paper Section 2 reviews the implicit low-rank imaging model, its matching tensors and closure constraints. In Sect. 3 we derive a method for estimating the degree of deformation. In Sect. 4 model estimation on partial data is described. Sections 5 and 6 describe the proposed priors and their implementation. Section 7 reports the experimental results. Finally, Sect. 8 concludes the paper.

Notation Vectors are denoted using bold fonts, *e.g.* $\mathbf{x}$ and matrices using sans-serif or calligraphic characters, *e.g.* M or $\mathcal{A}$. Index $i = 1, \ldots, N$ is used for the images, $j = 1, \ldots, M$ for the points. The Hadamard (element-wise) product is written $\odot$. Bars indicate 'centered' data, as in $\bar{\mathsf{X}}$. We use the Singular Value Decomposition, denoted SVD, *e.g.* $\mathsf{X} = \mathsf{U}\Sigma\mathsf{V}^{\mathsf{T}}$ where U and V are orthonormal matrices, and Σ is diagonal, containing the singular values of X in decreasing order. Operator vect(X) performs column-wise matrix vectorization.

2 The Implicit Low-Rank Non-Rigid Model

The standard rigid model describes the affine projection $\mathbf{x}_{ij}$ of a set of M 3D world points $\mathbf{W}_j$, represented by a $3 \times M$ shape matrix W onto N images represented by a $2N \times 3$ motion matrix P of stacked 2×3 affine camera projection matrices P_i:

$$\mathbf{x}_{ij} = \mathsf{P}_i \mathbf{W}_j + \mathbf{t}_i + \eta_{ij}, \tag{1}$$

where $\mathbf{t}_i$ is the position of the i-th camera and η_{ij} is a noise term. The $2N \times M$ matrix X of time varying coordinates $\mathbf{x}_{ij}$ is called the *measurement matrix* and has rank $r = 3$ [13]. In the non-rigid case, $r > 3$. The low-rank assumption is $r \ll \min\{2N, M\}$. In *explicit* non-rigid models it is assumed that the shape points $\mathbf{W}_{ij}$ can be written as linear combinations over l *basis shapes* $\mathbf{B}_{kj}$ with the *configuration weights* α_{ik}: $\mathbf{W}_{ij} = \sum_{k=1}^{l} \alpha_{ik} \mathbf{B}_{kj}$. The explicit imaging model is:

$$\mathbf{x}_{ij} = \mathsf{P}_i \left(\sum_{k=1}^{l} \alpha_{ik} \mathbf{B}_{kj} \right) + \mathbf{t}_i + \eta_{ij}. \tag{2}$$

Defining $\mathbf{K}_j^{\mathsf{T}} = (\mathbf{B}_{1j}^{\mathsf{T}} \ \cdots \ \mathbf{B}_{lj}^{\mathsf{T}})$, and $\mathsf{M}_i = (\alpha_{i1}\mathsf{P}_i \ \cdots \ \alpha_{il}\mathsf{P}_i)$ the model writes $\mathbf{x}_{ij} = \mathsf{M}_i \mathbf{K}_{ij} + \mathbf{t}_i + \eta_{ij}$ and thus $\mathsf{X} = \mathsf{MK} + \mathbf{t}\mathbf{1}^{\mathsf{T}} + \eta$ where M and K are $2N \times 3l$ and $3l \times M$ matrices and $\mathbf{1}$ is a vector of ones.

Replacing the model rank assumption $3l$ with the more general assumption $r = 3l$ being a positive integer, introducing a $(r \times r)$ full-rank matrix $\mathcal{A}$ and relaxing the replicated structure of the explicit motion matrix M_i gives the *implicit model*:

$$\begin{aligned} \mathbf{x}_{ij} &= \mathsf{M}_i \mathbf{K}_j + \mathbf{t}_i + \eta_{ij} \\ &= (\mathsf{M}_i \mathcal{A})\,(\mathcal{A}^{-1} \mathbf{K}_j) + \mathbf{t}_i + \eta_{ij} = \mathsf{J}_i \mathbf{S}_j + \mathbf{t}_i + \eta_{ij} \end{aligned} \tag{3}$$

and thus:

$$\mathsf{X} = \mathsf{JS} + \mathbf{t}\mathbf{1}^{\mathsf{T}} + \eta,$$

where the $2N \times r$ matrix $\mathsf{J} = \mathsf{M}\mathcal{A}$ is called the *implicit motion matrix*, and where the $r \times M$ matrix $\mathsf{S} = \mathcal{A}^{-1}\mathsf{K}$ is named the *implicit shape matrix*. The matrix $\mathcal{A}$ in (3) often is called *the mixing matrix* and represents a *corrective transformation* by which an implicit model can be upgraded to an explicit one. $\mathcal{A}$ defines the implicit coordinate frame in which the motion and the shapes are represented.

The model (3) is called implicit because no assumption is made on the replicated block structure of the motion matrices that often is used in explicit approaches *e.g.* [4, 5, 17]. Thus the implicit model is simpler than the explicit one but gives weaker constraints on point tracks. Note that the implicit (basis) shape vectors $\mathbf{S}_j$ are more difficult to interpret in terms of world coordinates. Similarly, the implicit motion matrices J_i (comprising camera pose and configuration weights) do no longer directly relate to the camera orientation. Here we assume that r is known. In Sect. 3 we describe how r is estimated.

Because the factorization of X is ambiguous, due to the freedom of choosing $\mathcal{A}$, an upgrading from implicit to explicit representation is important. Xiao *et al.* [22] show that constraints on both the explicit motion and shape matrices must be considered to achieve a unique solution, namely the 'rotation' and the 'basis' constraints. They give a closed-form solution based on these constraints. In [4] Brand presents an alternative less noise sensitive method without the 'basis' constraints. We consider the upgrading as a postprocessing step that is not further dealt with in this paper.

Our goal is thus, given X with missing and erroneous elements, to recover J, S, and $\mathbf{t}$ while detecting the erroneous elements, and predicting the missing ones. If X is complete (no missing data), one approximate factorization can be found using SVD as $\bar{\mathsf{X}} = \mathsf{U}\Sigma\mathsf{V}^{\mathsf{T}}$, where $\bar{\mathsf{X}}$ is the centered measurement matrix, *i.e.* with the translational part being canceled. The implicit motion and shape matrices J and S, are recovered as the r leading columns of *e.g.* U and the rows of $\Sigma\mathsf{V}^{\mathsf{T}}$ respectively. The assumption is that the information in the $d = 2N - r$ dimensional discarded subspace corresponds to the noise η. This method however has limited interest in practice since real data almost always contain errors and missing points. We propose a method that deals with this kind of measurements. It is based on extending the rigid matching tensors [20] to the low-rank shape model—we call them *low-rank matching tensors*. Matching tensors relate corresponding points over multiple images. Examples are the fundamental matrix and the trifocal tensor. In the non-rigid affine case the matching tensor is a $2N \times d$ matrix $\mathcal{N}$ whose columns span the d dimensional left nullspace of the centered measurement matrix $\bar{\mathsf{X}}$:

$$\mathcal{N}^{\mathsf{T}} \bar{\mathsf{X}} = 0. \tag{4}$$

As before $\mathcal{N}$ can be estimated using SVD. The closure constraints relate matching tensors to projection matrices. From (1) and (4) and for all implicit shape points $\mathbf{S}_j \in \mathbb{R}^r$ we have $\mathcal{N}^{\mathsf{T}}\mathsf{J}\mathbf{S}_j = \mathbf{0}$, which gives our $\mathcal{N}$*-closure constraint*:

$$\mathcal{N}^{\mathsf{T}} \mathsf{J} = 0. \tag{5}$$

The implicit motion matrix J consequently lies in the right nullspace of $\mathcal{N}^{\mathsf{T}}$ and may be estimated using an SVD. From J, $\mathbf{S}_j$ can be retrieved point-wise by triangulation. From $\mathbf{x}_j = \mathsf{J}\mathbf{S}_j$ we get $\mathbf{S}_j = \mathsf{J}^{\dagger}\mathbf{x}_j$, where $\mathsf{J}^{\dagger}$ is the pseudoinverse of J. In case of outlier contaminated data the computation of $\mathcal{N}$ as well as the triangulation must be done robustly so that blunders do not corrupt the computation. We use a RANSAC-based approach called MSAC [15].

3 Estimating the Rank

Estimating the rank r of the measurement matrix is of utmost importance. If r is chosen too small the model will not be able to express the deformations; if chosen too large the model will fit the noise. For many real sequences the transition between the singular value subspaces containing deformation information and those containing noise is blurred. This makes a guessing of r difficult. For explicit models (using $l = \lceil \frac{r}{3} \rceil$) an upgrade to a metric model may be difficult if l is selected too large [4]. In [18] it is argued that if appropriate priors are used an overestimation of r is not severe. We have made similar observations using the priors described in Sect. 5. However still a good guess of r is needed.

Most previous work assumes that the rank of X is given. A simple rank estimation by thresholding the singular value spectrum is used in [24]. If outliers corrupt the data or if the energy of the weakest non-rigid components is comparable in magnitude to the noise then such methods are unlikely to work. In [12] a deformation index is based on the correlation matrix of the in-frame position information. In [1] the Bayes Information Criteria (BIC) is used for rank selection. We propose to use the GRIC model selection criterion proposed in [14]. GRIC is a robustified version of BIC. Let k be the number of parameters of the model and $\mathcal{L}$ the log-likelihood of the error distribution, both functions of r. Then we aim at selecting the r minimizing $-2\mathcal{L} + k\log(M)$. In GRIC the error distribution is obtained from a mixture between a Gaussian inlier part and a uniform outlier part:

$$\mathrm{P} = \gamma \mathrm{P}_{\mathrm{in}} + (1-\gamma)\mathrm{P}_{\mathrm{out}} \tag{6}$$

$$= \frac{\gamma}{c}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{2N-r} \exp\left(-\frac{e^2}{2\sigma^2}\right) + \frac{1-\gamma}{v}, \tag{7}$$

where the error of the fit for the inliers can be modeled by an isotropic zero mean Gaussian. Here e is the $2N-r$ dimensional error perpendicular to the fitting manifold and $\frac{1}{c}$ is a prior of the point track, assuming a uniform distribution on the volume (with size c) on which an observation may occur. v similarly is the volume of space in which an outlier can occur. When, for a point track, $\frac{e^2}{\sigma^2}$ exceeds a value T, the track can be classified as being more probable to belong to the outlier distribution than to the inlier distribution. It is easy to show that:

$$T = 2\log\left(\frac{\gamma}{1-\gamma}\right) + (2N-r)\lambda, \tag{8}$$

where

$$\lambda = 2\log(U) - \log(2\pi\sigma^2), \tag{9}$$

$$U = \left(\frac{v}{c}\right)^{\frac{1}{2N-r}}. \tag{10}$$

Replacing the mixture model with a maximization approach, and using:

$$\rho\left(\frac{e2}{\sigma^2}\right) = \begin{cases} \frac{e2}{\sigma^2} & \text{if } \frac{e2}{\sigma^2} \le T, \\ T & \text{if } \frac{e2}{\sigma^2} > T \end{cases} \tag{11}$$

the log-likelihood term $-2\mathcal{L} = -2\log(P)$ can be shown to equate:

$$\sum_i^M \rho\left(\frac{e2}{\sigma^2}\right) - 2M\log\left(\frac{\gamma}{c}\right) - M(2N-r)\lambda, \tag{12}$$

where we have assumed independence of the M observations. Because the matching tensor has $d = 2N-r$ equations in $2N$ coordinates, and because the equations are homogeneous and orthogonal we have:

$$k = 2Nd - d - \frac{d(d-1)}{2} = (2N^2 - N) - \frac{1}{2}r(r-1). \tag{13}$$

Ignoring the constant terms $2M\log(\frac{\gamma}{c})$, $(2N^2-N)\log(M)$, and $2MN\lambda$ the GRIC measure becomes:

$$\mathrm{GRIC} = \sum_{j=1}^{M} \rho\left(\frac{e_j^2}{\sigma^2}\right) + Mr\lambda - \frac{1}{2}r(r-1)\log(M). \tag{14}$$

It is clear that the r minimizing this equation depends on the value of U defined in (10). To avoid estimating U we notice that an often used alternative approach to the estimation of T is by the value of the inverse cumulative χ^2 distribution with $2N-r$ degrees of freedom [8]. For relevant values of $2N-r$ this is approximately linear with a slope of λ. Thus we estimate λ directly.

Because the data may contain outliers we use the robust estimator MSAC [15] in conjunction with the GRIC. Thus to find the rank value minimizing GRIC we must sample this repeatedly for all relevant values of r. To limit the computational cost the sequence of trials is divided into groups using gradually narrower intervals of possible rank values.

As described in the next section a limitation to partial data forces the rank-value analysis to be made in (overlapping) frame intervals. From such block-rank values a global one may then be estimated, *e.g.* by the maximum. In general the block-rank values will be smaller than the global rank value, because no block may contain the complexity of deformation present in the full sequence. If the frame span for each block is very small the underestimate may be severe. In the next section a heuristic for selecting the block size is discussed. Here the deformation within each block is assumed sufficiently representative for the deformation within the compound sequence. In this case the maximum may be a good rank estimate, because it is the largest underestimate that there is evidence for. One alternative would be to apply a robust maximum operator.

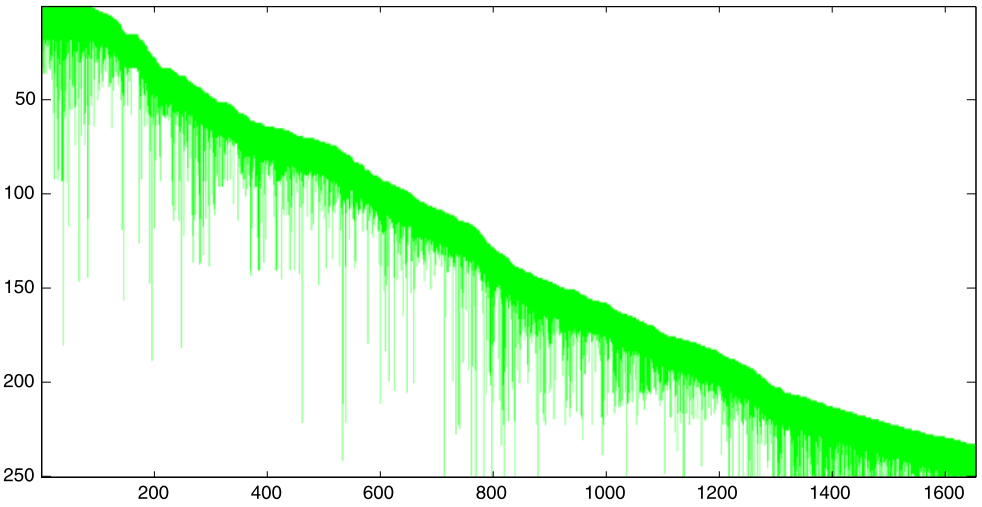

Fig. 1 1654 tracks in 250 frames. The degree of visibility is 13.26%

4 Handling Partial Data

Because of occlusions and imperfect tracking the measurement matrix X often will be defined only in a diagonal band. Figure 1 shows an example of the visibility matrix $\mathcal{V}$ for a sequence with 250 frames. $\mathcal{V}$ is a $N \times M$ binary matrix associated to X specifying if $\mathbf{x}_{ij}$ is defined or is missing. For practical use, handling of partial data is of great importance. A direct application of SVD, as described in the previous sections and used in a large number of previous methods [1, 3, 5, 16, 22], is not possible. One method to proceed is to extract blocks of complete data from the diagonal band of X [9, 10]. A survey of this and other methods for partial data handling can be found in [6]. As discussed in a following section we need an initial factorization of X which we then can improve by an iterative non-linear refinement step. For this purpose an application of the matching tensor approach on a block partitioning is ideal.

Given r, a d dimensional matching tensor $\mathcal{N}_b$ can be computed robustly for each block b. For each matching tensor, (5) gives a closure constraint on the joint motion matrix J:

$$\left(\mathsf{0}_{(d\times 2(i_b-1))} \quad \mathcal{N}_b^{\mathsf{T}} \quad \mathsf{0}_{(d\times 2(N-i'_b))}\right)\mathsf{J} = \mathsf{0}, \tag{15}$$

where i_b and i'_b are indexes of the first and last frame in block b. Stacking the constraints for all blocks yields an homogeneous linear least squares problem $\|\mathsf{AJ}\|^2$ which must be solved such that J has full column rank. Without loss of generality the full column rank constraint can be replaced by constraining J to be column orthonormal. A solution is given by the r last columns of V in the SVD $\mathsf{A} = \mathsf{U}\Sigma\mathsf{V}^{\mathsf{T}}$.

For each block the translation vector $\mathbf{t}^b$ is computed prior to $\mathcal{N}_b$. The joint translation vector $\mathbf{t}$ can be found by minimizing the reprojection error $\sum_b \|\mathbf{t}^b - \mathsf{J}_b\mathbf{T}_b - \mathbf{t}_b\|^2$, where $\mathbf{T}$ is the reconstructed centroid, and where the subscript b in J_b, $\mathbf{T}_b$, and $\mathbf{t}_b$ denotes the restriction of the joint matrices and vectors to the frames within block b. The reprojection error is rewritten $\|\mathsf{B}\mathbf{w} - \mathbf{b}\|^2$, where the unknown vector $\mathbf{w}$ contains $\mathbf{T}$ and $\mathbf{t}$. The solution is given by using the pseudo-inverse since there is an r dimensional ambiguity, making B rank deficient with a left nullspace of dimension r. This correspond to the translational ambiguity between the basis shapes and the joint translation $\mathbf{t}$: $\forall \boldsymbol{\gamma} \in \mathbb{R}^r$, $\mathbf{x}_j = \mathsf{J}\mathbf{S}_j + \mathbf{t} = \mathsf{J}(\mathbf{S}_j - \boldsymbol{\gamma}) + \mathsf{J}\boldsymbol{\gamma} + \mathbf{t} = \mathsf{J}\mathbf{S}'_j + \mathbf{t}'$.

Given the estimates of J and $\mathbf{t}$, the shape vectors $\mathbf{S}_j$ now can be computed by a robust minimization of the reprojection error. An advantage is that this makes possible a detection of outliers. Alternatively, as described in Sect. 6.2, computation of $\mathbf{S}_j$ may be postponed until the prior information is included.

As discussed in the previous section it is an advantage for the rank estimation if the data partitioning is made to maximize the block frame span. However, increasing this span will decrease the number of tracks visible in all of the frames within the block. As a minimum M_b must be larger than $2N_b$, where N_b and M_b are the block frame span and the number of tracks within the block. Because a RANSAC- based estimation is used $M_b \gg 2N_b$ is preferred. One heuristic is to choose N_b by the maximal value such that $M_b > 4N_b$. In practice this may not be possible if $N \gg M$, if only very few data is visible, or if some tracks are very short. Often is an advantage to eliminate tracks shorter than 10–20 frames. To guide the block partitioning one heuristic is to start with the previously mentioned choice of $M_b = 4N_b$, and then decrease or increase N_b, still requiring $M_b > 2N_b$, towards a situation where the block shape becomes similar to the shape of the measurement matrix itself.

5 The Priors

Prior knowledge on both motion and shape can be very useful in Non-Rigid Structure-from-Motion. In [7, 19] the analysis is bootstrapped from an assumption of a rigid scene. More basis shapes are incrementally added if necessary. In [24] a prior matrix is built from observed trajectories. Using a spectral clustering method a RANSAC-based motion segmentation is then derived. In [18] a probabilistic principal component analysis is used as an hierarchical Bayesian prior. The method makes possible a simultaneous estimation of 3D shape and motion, and of the deformation model. A Gaussian prior is put on the configuration weights α_{ik} in (2). Thus, the shapes are sought as similar as possible to each other. To a large degree the purpose of this approach and our surface shape prior (see below) are similar.

It is generally recognized [4, 23] that upgrading an initial factorization to a metric one, *i.e.* estimating the mixing matrix, is a hard problem. The methods in [4, 23] both rely on a non-trivial optimization step. The global optimum is rarely found unless the optimization is initialized at a point close to it. Thus an initial choice of coordinate frame according to a prior may show crucial for successful upgrading.

For nonlinear models with many parameters often many completely different parameter settings may result in fits which are approximately equally good when measured on the training data. However, when measured on data held back for test usage some solutions may predict these much better that others. Such solutions are said to generalize better. Often it is an advantage to select a solution that generalizes well but have a slightly worse reprojection error than the other way around. Below we motivate and formulate a temporal smoothness prior and a surface shape prior, both intended to improve generalization.

5.1 Temporal Smoothness Prior

For most image sequences, the camera motion is smooth. For points on a smoothly deforming surface the configuration weights smoothly vary as well which means that the surface does not 'jump' between poses but rather smoothly interpolates them. Since both the configuration weights and the camera parameters are encapsulated in the J_i matrices, these should vary smoothly from frame to frame giving the smoothness measure:

$$\mathcal{E}_{\mathrm{J}}(\mathsf{J}) = \sum_{i=1}^{N-1} \|\mathsf{J}_i - \mathsf{J}_{i+1}\|^2 = \|\mathsf{L}\|^2, \tag{16}$$

where L is the $2(N-1) \times r$ matrix of stacked projection difference matrices. The previously described factorization is ambiguous up to an $r \times r$ full rank mixing matrix $\mathcal{A}$. From (16) we see that $\mathcal{E}_{\mathrm{J}}(\mathsf{J}) \neq \mathcal{E}_{\mathrm{J}}(\mathsf{J}\mathcal{A})$. This suggests to select the $\mathcal{A}$ minimizing (16). Note that since $\mathcal{E}_{\mathrm{J}}(\mathsf{J}\mathcal{R})$ is invariant to any orthonormal matrix $\mathcal{R}$ we will not totally fix the mixing matrix but leave freedom for any orthogonal transform.

5.2 Surface Shape Prior

Points which are close in space also project closely on the images. In case of points on a deforming continuous surface the opposite is true as well. Solutions obtained by the previously described method does not encourage such behavior. As a consequence the projected trajectories for such close tracks may deviate significantly outside the estimation area. Often the ability to generalize acceptably disappears just 2–5 frames away from the images in which the points are visible. To improve generalization a surface shape prior is imposed.

First notice that without fixing the coordinate frame in which the shapes in S are represented the usual norm distance between two shapes is meaningless. However having fixed the mixing matrix (up to an orthogonal matrix) it becomes meaningful.

The shape similarity $\alpha(j_1, j_2)$ of two point tracks $j_1 \neq j_2$ is measured by a decreasing function of a distance measure $d(j_1, j_2)$ between the point tracks. The surface shape prior then is:

$$\mathcal{E}_{\mathrm{S}}(\mathsf{S}) = \sum_{(j_1,j_2)\in\Omega} \alpha(j_1, j_2) \cdot \|\mathbf{S}_{j_1} - \mathbf{S}_{j_2}\|^2, \tag{17}$$

where Ω is the set of track tuples simultaneously visible for a minimum number of, say 10, frames. As for the shape similarity a Gaussian $\alpha(j_1, j_2) = \exp(-\frac{d(j_1,j_2)^2}{2\sigma^2})$ is appropriate. In the experiments we computed σ as 0.03 times the image width. One measure of track distance is the maximum: $d(j_1, j_2) = \max_i\{\|\mathbf{x}_{ij_1} - \mathbf{x}_{ij_2}\|_2\}$. Alternative measures include robust estimates of the average or maximum point track distance. As for the temporal smoothness prior, $\mathcal{E}_{\mathrm{S}}$ is invariant to any orthonormal matrix $\mathcal{R}$.

6 Non-Rigid Structure-from-Motion with Priors

The model simultaneously minimizing the reprojection error, the smoothness prior and the surface shape prior, *i.e.* the cost:

$$\mathcal{E}_{\mathrm{RE}}(\mathsf{J}, \mathsf{S}) + \gamma\mathcal{E}_{\mathrm{J}}(\mathsf{J}) + \beta\mathcal{E}_{\mathrm{S}}(\mathsf{S}) \tag{18}$$

must be minimized by nonlinear optimization. To ensure a good starting point, and because the coordinate frame in which the shapes are represented influences the solution through the priors, we choose (initially) this frame by minimizing the temporal smoothness prior. This fixes the mixing matrix up to an orthogonal matrix, to which the surface

Table 1 Summary of our non-rigid low-rank implicit structure-from-motion algorithm

OBJECTIVE

Given M point tracks over N images as a possibly incomplete $(2N \times M)$ measurement matrix X, compute the implicit non-rigid motion J_i, the implicit non-rigid shape points $\mathbf{S}_j$, and an estimate of the rank r. Classify each image point as an inlier or an outlier.

ALGORITHM

1. Partition the sequence into overlapping blocks with complete data (Sect. 4). For each block, robustly estimate the block rank and the associated matching tensor (Sects. 2 and 3).
2. Estimate the global rank r: apply the closure constraints to solve for the joint implicit motion matrix J and the joint translation vector $\mathbf{t}$ (Sect. 4).
3. Detect the outliers by robustly fitting the model to each point track using RANSAC.
4. Use the temporal smoothness prior to select the coordinate transformation $\mathcal{A}$ and apply this to J (Sects. 5.1 and 6.1).
5. Estimate the shape vectors $\mathbf{S}_j$ minimizing a weighted sum of the reprojection error and the shape smoothness prior measure (Sects. 5.2 and 6.2).
6. Nonlinearly refine the implicit motion and shape points by minimizing a combination of the reprojection error, the temporal smoothness measure and the shape smoothness measure.
7. Estimate the missing data and glue tracks if they comply with the estimation.

shape prior is invariant. Next, by using the surface shape prior an initial guess for S is obtained. Finally J and S are jointly refined by nonlinear least-squares optimization. The constants γ and β in (18) are chosen *ad hoc* such that the two priors initially contribute relative to the reprojection error with certain amounts, say 0.2 and 0.02. Below, the initial application of the two priors is described. The algorithm is summarized in Table 1.

6.1 The Coordinate Frame

The temporal smoothness prior measure (16) obviously depends on the mixing matrix. Consequently we (partially) determine this as the $r \times r$ full rank matrix $\mathcal{A}$ minimizing $\mathcal{E}_{\mathsf{J}}(\mathsf{J}\mathcal{A}) = \|\mathsf{L}\mathcal{A}\|^2$. The motivation is that determining the mixing matrix ensures that the camera motion is 'close' to the optimal one. To avoid the shrinking effect of reducing the prior value by simply scaling down J we require $\det(\mathcal{A}) = 1$. Let $\mathsf{L} = \mathsf{U}\Sigma\mathsf{V}^{\mathsf{T}}$ be an SVD of L. A closed-form solution for $\mathcal{A}$ is provided in the Appendix.

$$\mathcal{A} = \left(\sqrt[r]{\prod_{k=1}^{r} \sigma_k} \right) \mathsf{V}\Sigma^{-1}. \tag{19}$$

Given $\mathcal{A}$ we change the coordinate frame by $\mathsf{J} \leftarrow \mathsf{J}\mathcal{A}$ and $\mathsf{S} \leftarrow \mathcal{A}^{-1}\mathsf{S}$ without changing the reprojection error. However the value of the prior $\mathcal{E}_{\mathsf{J}}(\mathsf{J})$ is significantly reduced.

6.2 Surface Shape Prior Implementation

Having fixed the non-rotational part of the mixing matrix it becomes meaningful to compute an estimate of the structure S. Given the modified joint motion matrix J, S is sought to minimize a weighted sum of the reprojection error and the surface shape prior:

$$\begin{aligned} &\mathcal{E}_{\mathrm{RE}}(\mathsf{J},\mathsf{S}) + \beta\mathcal{E}_{\mathsf{S}}(\mathsf{S}) \\ &\quad = \|\mathcal{V} \odot (\mathsf{X} - \mathsf{J}\mathsf{S} - \mathbf{t}\mathbf{1}^{\mathsf{T}})\|^2 \\ &\qquad + \beta \sum_{(j_1,j_2)\in\Omega} \alpha(j_1,j_2) \cdot \|\mathbf{S}_{j_1} - \mathbf{S}_{j_2}\|^2, \end{aligned} \tag{20}$$

where $\mathcal{V}$ is the combined inlier and visibility matrix and Ω is the set of 'close' point tracks. The S minimizing this expression leads to a larger reprojection error compared to the initial solution. The reprojection error increases with β. We choose a value of β such that the reprojection error either remains below say 2 pixels or is increased by a factor smaller than say 0.5. Since the result is not sensitive to an accurate value of β an approximate value is found using an iterative approach with only few iterations. Equation (20) can be rewritten:

$$\mathcal{E}_{\mathrm{RE}}(\mathsf{J},\mathsf{S}) + \beta\mathcal{E}_{\mathsf{S}}(\mathsf{S}) = \|\mathbf{v} \cdot (\bar{\mathbf{x}} - \mathcal{M}\mathbf{s})\|^2 + \beta\|\mathcal{L}\mathbf{s}\|^2, \tag{21}$$

where $\bar{\mathbf{x}} = \mathrm{vect}(\bar{\mathsf{X}})$ and $\mathbf{s} = \mathrm{vect}(\mathsf{S})$. $\mathcal{M} = \mathrm{diag}_M(\mathsf{J})$ is a $(2NM) \times (rM)$ block diagonal matrix with M repetitions of J. If $p = |\Omega|$ is the number of 'close' pairs of tracks then $\mathcal{L}$ has p row blocks $\mathcal{L}_{(j_1,j_2)}$ of the form:

$$\mathcal{L}_{(j_1,j_2)} = \alpha(j_1,j_2) \cdot (\mathsf{0}\ldots\mathsf{0},\mathsf{I},\mathsf{0}\ldots\mathsf{0},-\mathsf{I},\mathsf{0}\ldots\mathsf{0}), \tag{22}$$

where I and $\mathsf{0}$ are the $r \times r$ identity and zero matrices, and where the position of the two identity matrices correspond to the indexes j_1 and j_2. Thus $\mathcal{L}$ has size $(rp) \times (rM)$. With this rewriting we can directly see that the least squares solution is:

$$\mathbf{s} = [\mathcal{M}^{\top}\mathcal{M} + \beta\mathcal{L}^{\top}\mathcal{L}]^{-1}\mathcal{M}^{\top}\mathbf{x}. \tag{23}$$

Due to the sparseness of the matrices an implementation using a sparse matrix representation is advantageous.

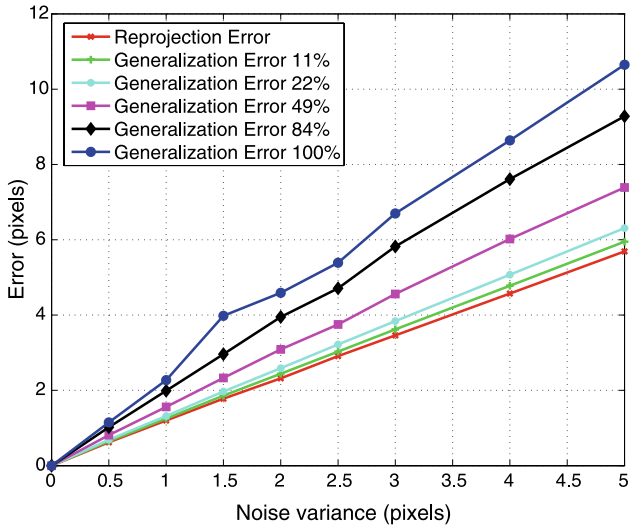

Fig. 2 Reprojection error and generalization error versus the variance of added noise σ for different percentages of hidden points to compute the generalization error

7 Experimental Results

In the experiments reported below we will first test the ability of the basic estimation method (without priors) to behave well on "easy" synthetic sequences with partial, noisy, and outlier corrupted data. Next we report experiments showing the advantages of the temporal smoothness prior and the shape smoothness prior on real images. Finally, an example of track gluing is reported.

7.1 Experiments on Easy Synthetic Data

We simulated $N = 180$ cameras observing a set of $M = 1000$ points generated from $l = 5$ basis shapes, hence with rank $r = 3l = 15$. The configuration weights were chosen in order to give a decaying energy to successive deformation modes, and such that the singular spectrum decayed smoothly to zero at the true rank value. The simulation setup produces a complete measurement matrix from which we extract a sparse, band-diagonal measurement matrix X with about 50% of the data, similar to what a real intensity-based point tracker produces. Gaussian distributed noise with zero mean and a variance of σ^2 was added to the image points. In the first experiment the true rank was assumed known. No prior information was used. Figure 2 shows plots of the reprojection error and the generalization error as functions of σ. The generalization measures are made in 5 bands including the training data and 11%, 22%, 49%, 84%, and all of the data held back for test. Figure 2 shows that the error is approximately proportional and slightly larger to the noise level. As expected the generalization error increases with the generalization distance. The reason that the generalization error is only slightly larger compared to the reprojection error is that the data was designed to be "easy".

Table 2 Average and standard deviation of estimated rank estimate as function of the true rank. In the test 300 "easy" sequences with a varying amount of up to 50% outliers was used

True rank	3	6	9	12	15	18
Average	3.67	5.98	8.47	11.06	13.68	16.32
Std.	0.44	0.41	0.57	0.60	0.62	0.76

In the second experiment the rank value estimation was tested using the same data. Table 2 show the average and the standard deviation of the estimated rank value as function of the true rank value. In the test 50 sequences for each of 6 different amounts of outlier contamination (0%, 10%, 20%, 30%, 40% and 50%) was used. No matter the degree of outlier contamination the results were, as expected, very similar. In Table 2 these numbers are averaged. As seen GRIC slightly over/under-estimates the rank when this is small/large.

The results show that the basic implicit non-rigid structure-from-motion modeling works well on "easy" synthetic data. When the difficulty of the data increases, *e.g.* when the singular value spectrum becomes flatter, the modeling error increases, and the rank estimate gets more uncertain with a tendency to be overestimated. In case of real data, overestimation is less meaningful because the model is empirical, *i.e.* no real data are fully explained by the model. In any case such "overestimation" does not seem serious if priors are used.

7.2 Experiments with Priors Using Real Videos

In the following experiments we measure the improvement in generalization by applying the two priors. The generalization is measured by the average point prediction accuracy as a function of the generalization distance, *i.e.* the column-wise distance in frames to the closest data point used for training. To make the experiment realistic only real sequences are used. Figure 3 shows single frames from the two sequences called *Bears* and *Groundhog day*. The sequence *Bears* shows a limited amount of deformation of a continuous surface. In total 94 points were visible in 94 images. The *Groundhog day* sequence is more difficult showing several independent deformations. Originally the measurement matrix was partial, so a complete sub-matrix of 75 frames and 117 point tracks was extracted. From the two complete matrices diagonal bands with 50% entries were selected for training. A third sequence was constructed from the (complete) sequence *Bears* by splitting each track in three sub-tracks. To make the test more realistic the data related to the last 1–7 frames of each track was randomly

Fig. 3 Images from the *Bears* sequence (*top*) and the *Groundhog day* sequence (*bottom*) with marked points

deleted. This resulted in a new measurement matrix with 282 tracks. The visibility matrix is shown to the left in Figure 6.

On the sequence *Bears* with partial data the rank was estimated to 5. After initial estimation $\mathcal{E}_{\mathrm{RE}} = 0.82$ pixels. Applying the priors increased this to 1.20 pixels. The temporal smoothness measure was reduced by a factor of 108.7. Fig. 4 shows on the top a plot of the average generalization error as function of the generalization distance. Without prior use the generalization becomes bad even for short generalization distances. With prior use the error is significantly reduced. For this, relatively easy, sequence the estimated model seems reliable up to a distance of about 15–20 frames. To illustrate the effect of the prior usage Fig. 5 show a close-up of 4 tracks from the *Bears* sequence. The positions computed by using the two priors (squares) are much closer to the true positions (stars) than the ones obtained by not using the priors (diamonds).

On the sequence constructed from *Bears* by track splitting the rank was, as before, estimated to 5. The reprojection error was increased from 0.53 pixel to 1.74 pixels by application of the priors. The temporal smoothness measure was reduced by a factor of 124.2. Figure 4 shows on the bottom that the average generalization error with prior usage is almost constant about 2–3 pixels independently of the generalization distance. This is much less than without prior

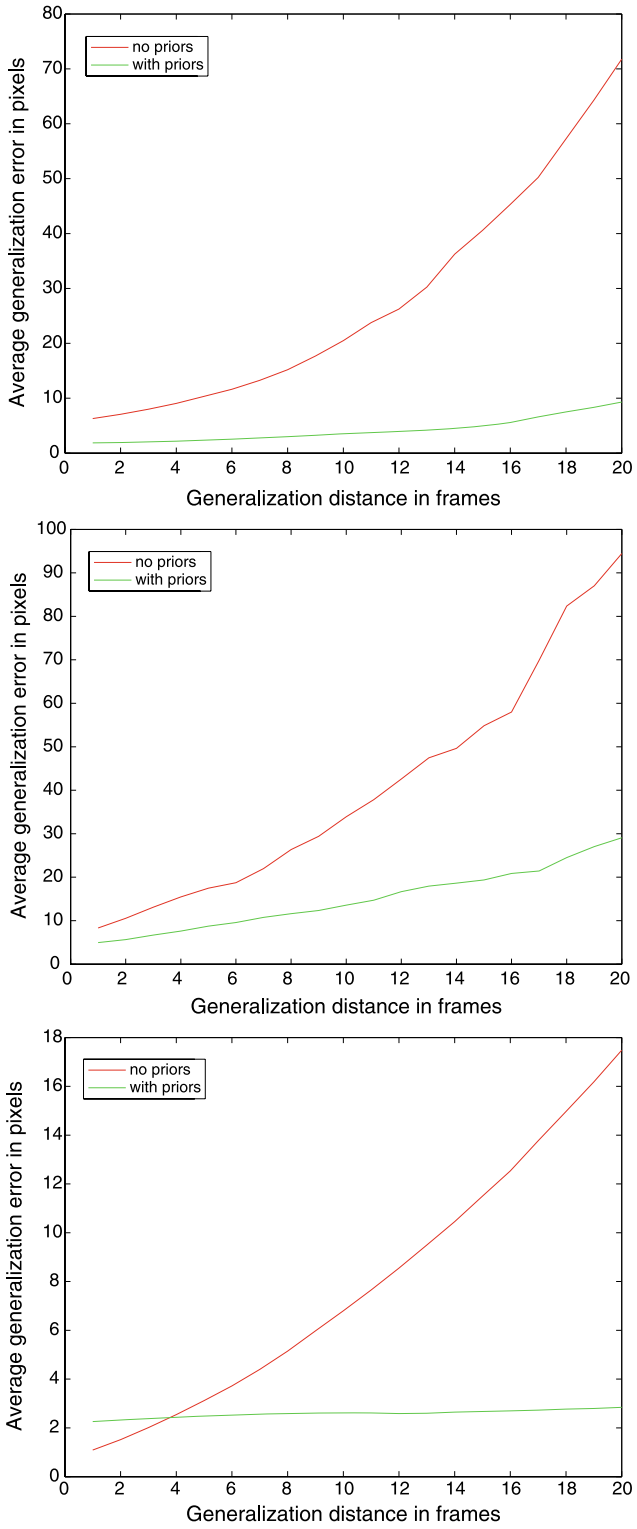

Fig. 4 Average generalization error as function of the generalization distance for the sequence *Bears* (*top*) and *Groundhog day* (*middle*), and the sequence obtained from *Bears* by track splitting (*bottom*)

usage. The reason the generalization errors here is 3–4 times smaller compared to the one showed on the top of Figure 4

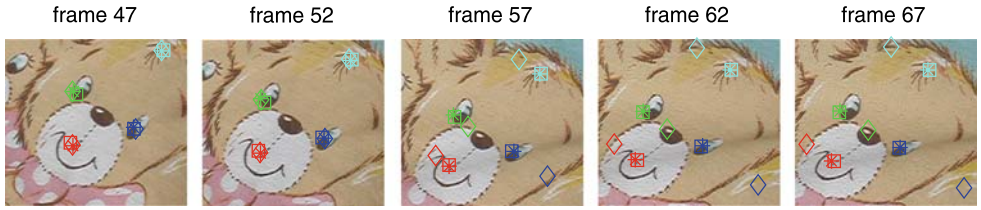

Fig. 5 Close-up sequence of 4 point tracks which visible parts (use for training) all ending close to frame number 47. 'True' positions, given by the tracker, are shown by *stars*. Predicted positions estimated without using the priors are shown by *diamonds*. Predicted positions estimated with use of the priors are shown by *squares*

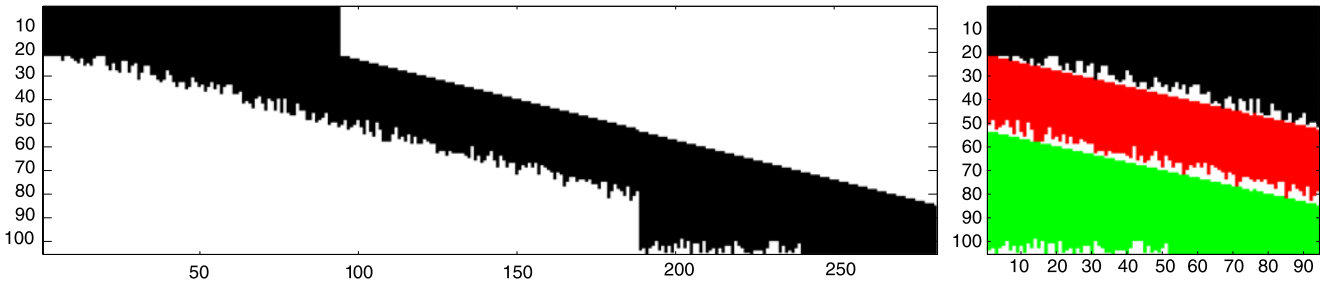

Fig. 6 *Left*: Visibility matrix for measurement matrix constructed by splitting the tracks of *Bears* in 3. *Right*: Glued tracks

is that here most data is maintained. The data is just split into more tracks.

On the sequence *Groundhog day* the rank was estimated to 14, indicating the difficulty of the sequence. The initial reprojection error was increased from 0.96 pixel to 1.62 pixels by application of the priors. The temporal smoothness measure was reduced by a factor of 5660.3. Figure 4 shows in the middle that the improvement in generalization still is significant, but less impressive compared to the other two sequences. A main reason is that here the assumption on scene smoothness made for the surface shape prior does not exactly match the physical scene behavior.

7.3 Experiments with Track Gluing

Often tracks are split due to imperfect tracking and it would be advantageous to glue together the parts. The experiments reported above indicate that without the priors the generalization error would be too large to allow a detection of split tracks. With use of the priors this however might be possible. Below we report a simple experiment using the previously described data obtained from *Bears* by splitting each track in three. Figure 6 show to the left the visibility matrix of the data.

After having estimated the model a gluing algorithm was run. This worked by iteratively gluing a point track with the best fitting track located up to 8 frames before or after the point track. A threshold on the fit was used to stop the gluing process. The resulting matrix of glued tracks showed to the right in Fig. 6 was identical to the original unsplit measurement matrix, *i.e.* the gluing was perfect. Probably this result is due to the simplicity of the deformation. In cases where tracks are split in more shorter sub-tracks a complete gluing cannot be made in a single pass because the reliable generalization distance still is limited. In such cases the estimation and gluing processes may be iterated.

8 Conclusions

We described an implicit non-rigid Structure-from-Motion approach with priors for temporal smoothness and surface shape coherency. We showed that the priors significantly improves the prediction of points projections in frames where data is missing, *i.e.* the generalization ability. Experiments have shown the improvement of the priors sufficient for gluing together point tracks split by imperfect tracking. To our knowledge our approach to Non-rigid Structure-from-Motion is the first that simultaneously can handle a substantial amount of missing data and outliers, can estimate the rank of the measurement matrix, and includes generic prior knowledge on temporal and surface smoothness. We expect the temporal smoothness prior to drive the estimated model closer to an explicit configuration. Further work will show how much this helps in upgrading to metric.

Appendix: Proof: Maximizing the Temporal Smoothness Prior

Below we sketch a proof that by choosing $\mathcal{A}$ as in (19) the temporal smoothness measure (16) is minimized. Thus

we find the implicit coordinates maximizing the temporal smoothness. Let $\mathsf{L} = \mathsf{U}\Sigma\mathsf{V}^\mathsf{T}$ be an SVD of L. Let $\mathcal{A} = \mathsf{QDW}$ be an SVD of $\mathcal{A}$. We parameterize $\mathcal{A}$ as $\mathcal{A} = \mathsf{QD}$ since $\mathcal{E}_\mathrm{J}(\mathsf{J}\mathcal{A}) = \mathcal{E}_\mathrm{J}(\mathsf{JQD})$. Let $\mathsf{Y} = \mathsf{V}^\mathsf{T}\mathsf{Q} \in \mathcal{O}(r)$. We can rewrite $\mathcal{E}_\mathrm{J}(\mathsf{J}\mathcal{A})$ as:

$$\begin{aligned}\|\mathsf{L}\mathcal{A}\|^2 &= \|\mathsf{U}\Sigma\mathsf{V}^\mathsf{T}\mathsf{QD}\|^2 \\ &= \|\Sigma\mathsf{YD}\|^2 = d_1^2\|\Sigma\mathbf{y}_1\|^2 + \cdots + d_r^2\|\Sigma\mathbf{y}_r\|^2, \end{aligned} \tag{24}$$

where $d_r \geq d_{r-1} \geq \cdots \geq d_1 \geq 0$ and with $\mathbf{y}_i$ the columns of Y. We want to find the $\mathbf{y}_i$ and the d_k minimizing the expression under the constraints that $\prod d_k = 1$, and that Y is orthonormal. Due to the ordering of the singular values we can split the minimization problem into r subproblems corresponding to the terms in the sum. From this we get $\mathsf{Y} = \mathsf{I}$, *i.e.* $\mathsf{Q} = \mathsf{V}$. The minimization problem then is reduced to:

$$\min_{\{d_k\}, \prod d_k = 1, d_r \geq \cdots \geq d_1 \geq 0} \sum_{k=1}^{r} (\sigma_k d_k)^2. \tag{25}$$

Introducing Lagrange multipliers λ and μ_j a compound objective function is formulated:

$$\min_{\{d_k\}} \sum_{k=1}^{r} (\sigma_k d_k)^2 + \lambda \left(\prod_{z=1}^{r} d_z - 1 \right) + \sum_{j=1}^{r} \mu_j (d_j - d_{j-1}). \tag{26}$$

It can easily be shown that this function has a minimum given by:

$$2\sigma_k^2 d_k = \lambda \left(\prod_{z=1, z\neq k}^{r} d_z \right) = \frac{\lambda}{d_k}. \tag{27}$$

Letting $\alpha = \sqrt{\lambda/2}$ and checking the unit determinant constraint it is seen that:

$$\alpha = \sqrt[r]{\prod_{k=1}^{r} \sigma_k}. \tag{28}$$

Putting things together we reach expression (19).

To show that the minimum is global the Karush-Kuhn-Tucker conditions can be applied. A sufficient condition for the minimum to be global is that the three terms in (26) are twice differentiable and that the Hessian matrix evaluated in $\mathbb{R}^{r+}$ is positive semi-definite. The Hessian for the first term is diagonal with elements $2\sigma_k^2$. The last term is linear so the Hessian is a positive semi-definite null matrix. The Hessian for the second term $\prod_{z=1}^{r} d_z$ is given by:

$$\mathsf{H} = \begin{pmatrix} 0 & \prod_{i\neq 1,2}^{r} d_i & \cdots & \prod_{i\neq 1,r}^{r} d_i \\ \prod_{i\neq 1,2}^{r} d_i & 0 & \cdots & \prod_{i\neq 2,r}^{r} d_i \\ \vdots & \vdots & \vdots & \vdots \\ \prod_{i\neq 1,r}^{r} d_i & \prod_{i\neq 2,r}^{r} d_i & \cdots & 0 \end{pmatrix}. \tag{29}$$

For $\mathbf{x} \in \mathbb{R}^{r+}$ it is clear that $\mathbf{x}^\top \mathsf{H}\mathbf{x} \geq 0$, so H is positive semi-definite in $\mathbb{R}^{r+}$.

References

1. Aanæs, H., Kahl, F.: Estimation of deformable structure and motion. In: The Vision and Modeling of Dynamic Scenes Workshop (2002)
2. Bartoli, A., Olsen, S.: A batch algorithm for implicit non-rigid shape and motion recovery. In: Workshop on Dynamical Vision at ICCV'05 (2005)
3. Brand, M.: Morphable 3D models from video. In: Conf. on Computer Vision and Pattern Recognition, pp. 456–463 (2001)
4. Brand, M.: A direct method for 3D factorization of nonrigid motion observed in 2D. In: Conf. on Computer Vision and Pattern Recognition, pp. 122–128 (2005)
5. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3D shape from image streams. In: Conf. on Computer Vision and Pattern Recognition, pp. 690–696 (2000)
6. Buchanan, A.M., Fitzgibbon, A.W.: Damped Newton algorithms for matrix factorization with missing data. In: Conf. on Computer Vision and Pattern Recognition, pp. 316–322 (2005)
7. Del Bue, A., Lladó, X., de Agapito, L.: Non-rigid metric shape and motion recovery from uncalibrated images using priors. In: Conf. on Computer Vision and Pattern Recognition, pp. 1191–1198 (2006)
8. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2003)
9. Jacobs, D.W.: Linear fitting with missing data for structure-from-motion. Comput. Vis. Image Underst. **82**(1), 57–81 (2001)
10. Martinec, D., Pajdla, T.: 3D reconstruction by fitting low-rank matrices with missing data. In: Conf. on Computer Vision and Pattern Recognition, pp. 198–205 (2005)
11. Olsen, S., Bartoli, A.: Using priors for improving generalization in non-rigid structure-from-motion. In: Proceedings of the British Machine Vision Conference, pp. 1050–1059 (2007)
12. Roy-Chowdhury, A.K.: A measure of deformability of shapes, with application to human motion analysis. In: Conf. on Computer Vision and Pattern Recognition, pp. 398–404 (2005)
13. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: A factorization method. Int. J. Comput. Vis. **9**(2), 137–154 (1992)
14. Torr, P.H.S.: Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. Int. J. Comput. Vis. **50**(1), 27–45 (2002)
15. Torr, P.H.S., Zisserman, A.: MLESAC: A new robust estimator with application to estimating image geometry. Comput. Vis. Image Underst. **78**, 138–156 (2000)
16. Torresani, L., Bregler, C.: Space-time tracking. In: European Conference on Computer Vision, pp. 801–812 (2002)
17. Torresani, L., Hertzmann, A.: Automatic non-rigid 3D modeling from video. In: European Conference on Computer Vision, pp. 299–312 (2004)
18. Torresani, L., Hertzmann, A., Bregler, C.: Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. In: IEEE PAMI (2007)
19. Torresani, L., Yang, D.B., Alexander, E.J., Bregler, C.: Tracking and modeling non-rigid objects with rank constraints. In: Conf. on Computer Vision and Pattern Recognition, pp. 493–500 (2001)
20. Triggs, B.: Linear projective reconstruction from matching tensors. Image Vis. Comput. **15**(8), 617–625 (1997)

21. Vidal, R., Abretske, D.: Nonrigid shape and motion from multiple perspective views. In: European Conference on Computer Vision, pp. 205–218 (2006)
22. Xiao, J., Chai, J.-X., Kanade, T.: A closed-form solution to non-rigid shape and motion recovery. In: European Conference on Computer Vision, pp. 573–587 (2004)
23. Xiao, J., Kanade, T.: Non-rigid shape and motion recovery: Degenerate deformations. In: International Conference on Computer Vision and Pattern Recognition, pp. 668–675 (2004)
24. Yan, J., Pollefeys, M.: Articulated motion segmentation using RANSAC with priors. In: Workshop on Dynamical Vision (2005)

S.I. Olsen received a M.Sc. in Computer Science in 1984 and a Ph.D. in Computer Science in 1988 both from University of Copenhagen. Since 1991 he has been associate professor at the Department of Computer Science at the University of Copenhagen. Currently he is with the E-Science center at the University of Copenhagen. His main research interest is computer vision, image processing, and pattern recognition.

A. Bartoli is a permanent CNRS research scientist at the LASMEA laboratory in Clermont-Ferrand, France, since October 2004 and a visiting professor at DIKU in Copenhagen, Denmark for 2006–2009. Before that, he was a post-doctoral researcher at the University of Oxford, UK, in the Visual Geometry Group, under the supervision of Prof. Andrew Zisserman. He did his Ph.D. in the Perception group, in Grenoble at INRIA, France, under the supervision of Prof. Peter Sturm and Prof. Radu Horaud. He received the 2004 INPG Ph.D. Thesis prize and the 2007 best paper award at CORESA. Since September 2006, he is co-leading the ComSee research team. His main research interests are in Structure-from-Motion in rigid and non-rigid environments and machine learning within the field of computer vision.