Using Priors for Improving Generalization in Non-Rigid Structure-from-Motion

S. I. Olsen[1] **A. Bartoli**[2,1]
[1] DIKU, Copenhagen, Denmark [2] LASMEA, Clermont-Fd, France

Abstract

This paper describes how the generalization ability of methods for non-rigid Structure-from-Motion can be improved by using priors. Most point tracks are often visible only in some of the images; predicting the missing data can be important. Previous Maximum-Likelihood (ML)-approaches on implicit non-rigid Structure-from-Motion generalize badly. Although the estimated model fits well to the visible training data, it often predicts the missing data badly. To improve generalization we propose to add a temporal smoothness prior and a continuous surface shape prior to an ML-approach. 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 a 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 ML-solution. The proposed method is fully automatic: it handles a substantial amount of missing data as well as outlier contaminated data, and automatically estimates the rank of the measurement matrix.

1 Introduction

Non-rigid Structure-from-Motion concerns the simultaneous recovery of the deforming world structure and camera motion from image features. Such analysis extends the classical rigid setup [10] to situations with deforming scenes such as expressive faces, moving cars, *etc*. In [1, 3, 5, 13, 18] methods where the non-rigidity was represented as a linear combination of *basis shapes* were developed and analyzed.

Many previous methods cannot handle situations with missing data [1, 3, 5, 13, 16, 17], but see also [6, 9, 14]. The amount of non-rigidity – the number of basis shapes – often is assumed known [3, 5, 13, 14, 17]. These assumptions seriously limit the applicability of the methods. Recently an implicit low-rank model solving both problems has been proposed [2]. The present paper reviews and extends this approach. One major difference is the use of a MAP-estimation where priors are added to the ML cost function.

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. In general, the model minimizing the reprojection error – the ML-estimate – does not generalize well. We derive an alternative approach where the optimization function is augmented with a temporal smoothness prior and a surface smoothness prior. The priors we use are different from the ones favoring rigidity in [7, 14].

The proposed MAP-estimator is based on four main steps. First, we compute an initial solution with an existing ML-estimator. Second, we change the implicit coordinate frame such that the temporal smoothness prior is minimized. This ensures that the prior is globally satisfied since we derive a closed-form, optimal solution to this problem. Third, we re-estimate the implicit structure by minimizing a combination of the reprojection error and the surface shape prior. Finally, we jointly refine the motion and structure estimates by nonlinear optimization. Experimental results on simulated and real data show that the generalization ability is greatly improved compared to the standard ML-estimation.

Section 2 reviews the implicit low-rank imaging model, its matching tensors and closure constraints. In section 3 and 4 the rank and 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, section 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 $\mathscr{A}$. 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. The operator $\mathrm{vect}(\mathsf{X})$ 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 $\mathbf{S}_j$, represented by a $3 \times M$ shape matrix S onto N images represented by a $2N \times 3$ motion matrix J of stacked 2×3 affine camera projection matrices J_i:

$$\mathbf{x}_{ij} = \mathsf{J}_i \mathbf{S}_j + \mathbf{t}_i \tag{1}$$

where $\mathbf{t}_i$ is the position of the i'th camera. The $2N \times M$ matrix X of time varying coordinates is called *the measurement matrix* and has rank $r = 3$.

In the non-rigid case $r > 3$. The low-rank assumption is $r \ll \min\{2N, M\}$. The implicit low-rank non-rigid model extends (1) by letting the camera and shape matrices have dimensions $2N \times r$ and $r \times M$. The model is implicit because no assumptions are made on the replicated block structure of the camera matrices that often is used in explicit approaches *e.g.* [4, 5, 14]. Thus the implicit model is simpler than the explicit one and 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 camera matrix J_i (comprising camera pose and configuration weights) does no longer directly relate to the camera orientation.

The factorization of the centered measurement matrix $\bar{\mathsf{X}} = \mathsf{J}\mathsf{S} = (\mathsf{J}\mathscr{A})(\mathscr{A}^{-1}\mathsf{S})$ is ambiguous since the equation holds for any full rank $r \times r$ mixing matrix $\mathscr{A}$ defining the coordinate frame in which the cameras and shapes are represented. If X is filled (no missing data), one factorization can be found using SVD as $\bar{\mathsf{X}} = \mathsf{U}\Sigma\mathsf{V}^\mathsf{T}$. The joint implicit camera 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.

Matching tensors [15] relate corresponding points over multiple images. In the non-rigid affine case the matching tensor is a matrix $\mathcal{N}$ whose columns span the d dimensional left nullspace of the centered measurement matrix $\bar{\mathsf{X}}$:

$$\mathcal{N}^{\mathsf{T}}\bar{\mathsf{X}} \quad = \quad 0. \tag{2}$$

The size of $\mathcal{N}$ is $(2N \times d)$ where the tensor dimension is $d = 2N - r$. $\mathcal{N}$ constrains each point track $\bar{\mathbf{x}}_j$ – the j-th column of $\bar{\mathsf{X}}$ – by d linear homogeneous equations $\mathcal{N}^{\mathsf{T}}\bar{\mathbf{x}}_j = \mathbf{0}$. The closure constraints introduced by Triggs in [15] for rigid scenes relate matching tensors to projection matrices. From (1) and (2) 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 the *$\mathcal{N}$-closure constraint*:

$$\mathcal{N}^{\mathsf{T}}\mathsf{J} \quad = \quad 0. \tag{3}$$

The joint implicit camera matrix J consequently lies in the right nullspace of $\mathcal{N}^{\mathsf{T}}$. From J, $\mathbf{S}_j$ is 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 robust so that blunders do not corrupt the computation. We use a RANSAC-based approach called MSAC [12]. Finally, we need r to compute $\mathcal{N}$. As described later we apply the GRIC model selection criterion [11] in conjunction with MSAC to estimate the optimal model size, *i.e.* r.

3 Handling Partial Data

As a number of previous methods [1, 3, 5, 13, 17] we factorize the measurement matrix X using SVD. Since X often is banded because of occlusions and imperfect tracking, handling of missing data is important. As [8, 9] we use a blockwise approach where the measurement matrix is partitioned into a set of highly overlapping blocks. Given r, a d-dimensional matching tensor $\mathcal{N}_b$ can be computed robustly for each block b. For each matching tensor, equation (3) gives a closure constraint on the joint camera matrix J:

$$\begin{pmatrix} 0_{(d\times 2(i_b-1))} & \mathcal{N}_b^{\mathsf{T}} & 0_{(d\times 2(N-i'_b))} \end{pmatrix} \mathsf{J} \quad = \quad 0 \tag{4}$$

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 $||\mathsf{A}\mathsf{J}||^2$ 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 a 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 $\mathbf{t}$: $\forall\gamma \in \mathbb{R}^r$, $\mathbf{x}_j = \mathsf{J}\mathbf{S}_j + \mathbf{t} = \mathsf{J}(\mathbf{S}_j - \gamma) + \mathsf{J}\gamma + \mathbf{t} = \mathsf{J}\mathbf{S}'_j + \mathbf{t}'$.

Given the estimates of J and $\mathbf{t}$, the shape vectors S_j could now be computed by a robust minimization of the reprojection error. However, as described in section 6.2, we prefer to postpone this computation until the prior is included.

4 Estimating the rank

With the exception of [1] most of the previous work assumes that the rank of X is given. We propose to use the robust estimator MSAC in conjunction with the GRIC model selection criterion proposed in [11]. Letting k be the number of parameters of the model and $\mathscr{L}$ the log-likelihood of the error distribution obtained by marginalizing a mixture of a Gaussian inlier part and a uniform outlier part, GRIC is defined by: $\text{GRIC} = -2\mathscr{L} + k\log(M)$. Expanding and removing constants the measure becomes:

$$\text{GRIC} = \sum_{j=1}^{M} \rho\left(\frac{e_j^2}{\sigma^2}\right) + Mr\lambda - \frac{1}{2}r(r-1)\log(M) \tag{5}$$

where e_j is the prediction error for the j-th point track, σ^2 is the variance of the point tracker localization error, where $\lambda = 2\log(U) - \log(2\pi\sigma^2)$, and where the function ρ is $\rho(x) = x$ for $x < T$ and $\rho(x) = T$ otherwise. T is the point of intersection of the Gaussian inlier distribution and the uniform outlier distribution and defined by: $T = 2\log\left(\frac{\gamma}{1-\gamma}\right) + (2N-r)\lambda$ where γ is the percentage of inliers. The value of U is determined by the relative weighting of the inlier and outlier distribution and have a major influence on the rank estimation. To estimate U we notice that an alternative approach to the estimation of T is by the value of inverse cumulative χ^2 distribution with $2N-r$ degrees of freedom. For relevant values of $2N-r$ this is approximately linear with a slope of λ. More details are given in [2]. To estimate the rank robustly we must sample the GRIC value 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.

5 The Priors

Below we motivate and formulate the temporal smoothness prior and the surface shape prior. In the following section the implementation of the priors is described.

5.1 Temporal Smoothness

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 coordinate axes are encapsulated in the J_i-matrices, these should vary smoothly from frame to frame giving the smoothness measure:

$$\mathscr{E}_{\mathsf{J}}(\mathsf{J}) = \sum_{i=1}^{N-1} ||\mathsf{J}_i - \mathsf{J}_{i+1}||^2 = ||\mathsf{L}||^2 \tag{6}$$

where L is the $2(N-1) \times r$ matrix of stacked projection difference matrices. The previously described factorization is ambiguous up to a $r \times r$ full rank mixing matrix $\mathscr{A}$. From (6) we see that $\mathscr{E}_{\mathrm{J}}(\mathsf{J}) \neq \mathscr{E}_{\mathrm{J}}(\mathsf{J}\mathscr{A})$.

5.2 Surface Shape

Points which are close in space also are close in the images. In case of points on a deforming continuous surface the opposite is true as well. Solutions obtained by the method described above 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. The shape similarity $\alpha(j_1, j_2)$ of two point tracks $j_1 \neq j_2$ is measured by a Gaussian function $e^{\lambda d^2(j_1,j_2)}$ of the maximal distance $d(j_1, j_2) = \max_i \left\{||\mathbf{x}_{ij_1} - \mathbf{x}_{ij_2}||_2\right\}$ in the images in which both tracks are visible. The surface shape prior then is:

$$\mathscr{E}_{\mathrm{S}}(\mathsf{S}) = \sum_{(j_1,j_2)} \alpha(j_1, j_2) \cdot ||\mathbf{S}_{j_1} - \mathbf{S}_{j_2}||^2. \tag{7}$$

As for the smoothness prior we see that $\mathscr{E}_{\mathrm{S}}(\mathsf{S}) \neq \mathscr{E}_{\mathrm{S}}(\mathscr{A}^{-1}\mathsf{S})$.

6 Non-Rigid SfM With Priors

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

$$\mathscr{E}_{\mathrm{RE}} + \gamma\mathscr{E}_{\mathrm{J}} + \beta\mathscr{E}_{\mathrm{S}} \tag{8}$$

must be obtained by nonlinear optimization. To ensure a good starting point, and because the coordinate frame in which the shapes are represented influences the solution, we choose (initially) this frame by minimizing the temporal smoothness prior. As shown below this fixes the mixing matrix up to an orthogonal matrix, to which the surface shape prior is invariant. Next, by using the surface shape prior an initial guess for S is estimated. Finally J and S are jointly refined by nonlinear least-squares optimization. The constants γ and β in (8) 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.

6.1 The Coordinate Frame

The prior measure (6) obviously depends on the mixing matrix. Consequently we (partially) determine this as the $r \times r$ full rank matrix $\mathscr{A}$ minimizing $\mathscr{E}_{\mathrm{J}}(\mathsf{J}\mathscr{A}) = ||\mathsf{L}\mathscr{A}||^2$. 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 $\det(\mathscr{A}) = 1$. Let $\mathsf{L} = \mathsf{U}\Sigma\mathsf{V}^\mathsf{T}$ be a (reduced) SVD of L. Below we sketch a proof for a closed-form solution for $\mathscr{A}$:

$$\mathscr{A} = \left(\sqrt[r]{\prod_{k=1}^{r} \sigma_k}\right) \mathsf{V}\Sigma^{-1}. \tag{9}$$

Given $\mathscr{A}$ we change the coordinate frame by $\mathsf{J} \leftarrow \mathsf{J}\mathscr{A}$ and $\mathsf{S} \leftarrow \mathscr{A}^{-1}\mathsf{S}$ without changing the reprojection error. However the value of the prior $\mathscr{E}_{\mathrm{J}}(\mathsf{J})$ is significantly reduced. It should be noted that (9) only fixes the mixing matrix up to a $r \times r$ orthogonal matrix.

A proof of equation (9). Let $\mathscr{A} = \mathsf{QDW}$ be an SVD of $\mathscr{A}$. We parameterize $\mathscr{A}$ as $\mathscr{A} = \mathsf{QD}$ since $\mathscr{E}_{\mathrm{J}}(\mathsf{J}\mathscr{A}) = \mathscr{E}_{\mathrm{J}}(\mathsf{JQD})$. Let $\mathsf{Y} = \mathsf{V}^{\mathsf{T}}\mathsf{Q} \in \mathscr{O}(r)$. We can rewrite $\mathscr{E}_{\mathrm{J}}(\mathsf{J}\mathscr{A})$ as:

$$||\mathsf{L}\mathscr{A}||^2 \;=\; ||\mathsf{U}\Sigma\mathsf{V}^{\mathsf{T}}\mathsf{QD}||^2 \;=\; ||\Sigma\mathsf{YD}||^2 \;=\; d_1^2||\Sigma\mathbf{y}_1||^2 + \cdots + d_r^2||\Sigma\mathbf{y}_r||^2 \qquad (10)$$

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 $\mathsf{Y} = \mathrm{I}$, *i.e.* $\mathsf{Q} = \mathsf{V}$. The minimization problem then is reduced to:

$$\min_{\{d_k\},\prod d_k=1, d_r\geq\cdots\geq d_1\geq 0} \sum_{k=1}^{r} (\sigma_k d_k)^2. \qquad (11)$$

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

$$\min_{\{d_k\}} \sum_{k=1}^{r} (\sigma_k d_k)^2 + \lambda \left(\prod_{z=1}^{r} d_z - 1 \right) + \sum_{j=1}^{r} \mu_j (d_j - d_{j-1}). \qquad (12)$$

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

$$2\sigma_k^2 d_k = \lambda \left(\prod_{z=1, z\neq k}^{r} d_z \right) = \frac{\lambda}{d_k}. \qquad (13)$$

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

$$\alpha = \sqrt[r]{\prod_{k=1}^{r} \sigma_k}. \qquad (14)$$

Putting things together we reach expression (9).

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 (12) 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 $\prod_{z=1}^{r} d_z$ can easily be show to be positive semi-definite.

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:

$$\mathscr{E}_{\mathrm{RE}} + \beta\mathscr{E}_{\mathrm{S}} \;=\; \|\mathscr{V} \odot (\mathsf{X} - \mathsf{JS} - \mathbf{t}\cdot\mathbf{1}^{\mathsf{T}})\|^2 + \beta \sum_{(j_1,j_2)\in\Omega} \alpha(j_1,j_2)\cdot||\mathbf{S}_{j_1} - \mathbf{S}_{j_2}||^2 \qquad (15)$$

where $\mathscr{V}$ 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 increase in reprojection error is limited by a factor of 0.1 to 0.5. This is done using an iterative approach. Equation (15) can be rewritten as:

$$\mathscr{E}_{\text{RE}} + \beta \mathscr{E}_{\text{S}} \; = \; ||\mathbf{v} \cdot (\bar{\mathbf{x}} - \mathscr{M}\mathbf{s})||^2 + \beta ||\mathscr{L}\mathbf{s}||^2 \tag{16}$$

where $\bar{\mathbf{x}} = \text{vect}(\bar{\mathsf{X}})$ and $\mathbf{s} = \text{vect}(\mathsf{S})$. $\mathscr{M} = \text{diag}_M(\mathsf{J})$ 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 $\mathscr{L}$ has p row blocks $\mathscr{L}_{(j_1,j_2)}$ of the form:

$$\mathscr{L}_{(j_1,j_2)} = \alpha(j_1, j_2) \cdot (0...0, \mathrm{I}, 0, ...0, -\mathrm{I}, 0...0) \tag{17}$$

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

$$\mathbf{s} = [\mathscr{M}^\top \mathscr{M} + \beta \mathscr{L}^\top \mathscr{L}]^{-1} \mathscr{M}^\top \mathbf{x}\,. \tag{18}$$

7 Experimental Results

In the experiments reported below we concentrate on the improvement with respect to generalization by using the camera smoothness and surface shape priors.

7.1 Synthetic Data

In the first test we generated synthetic data with 100 frames and 100 point tracks and with true rank varying from 3 to 18. For each data set, models with and without use of the two priors were estimated from the diagonal 60% entries. The estimation error is measured as a function of the generalization distance in frames. For medium to large distances the error distribution were very long tailed. Therefore for each distance we measured the improvement in generalization by the ratio of medians without and with prior use. The generalization improvement measure increased with the rank as well as with the generalization distance. Figure 1 shows to the left the average (over all data sets) of the improvement. In more absolute terms we relate the error in the generalization area to the error in the training area by the percentage of points with generalization error exceeding a value $\mu + k\sigma$, where μ and σ are the mean and spread of the reconstruction error and $k = 2.5$, 5, and 10. An example is shown to the right in Figure 1. The smoothness measure (6) decreased by a factor between 80 and 500. The results on synthetic data showed that at the expense of a small increase of the reprojection error, the generalization error can be significantly reduced. In particular the number of very large errors is reduced. Experiments that are not reported here showed that the generalization improvement increased with the difficulty of the data, *e.g.* with the amount of measurement noise and with r.

7.2 Real Data

We applied the same testing procedure on data from two real sequences called *Bears* and *Groundhog day*. Figure 2 shows single frames from the two sequences. From the two

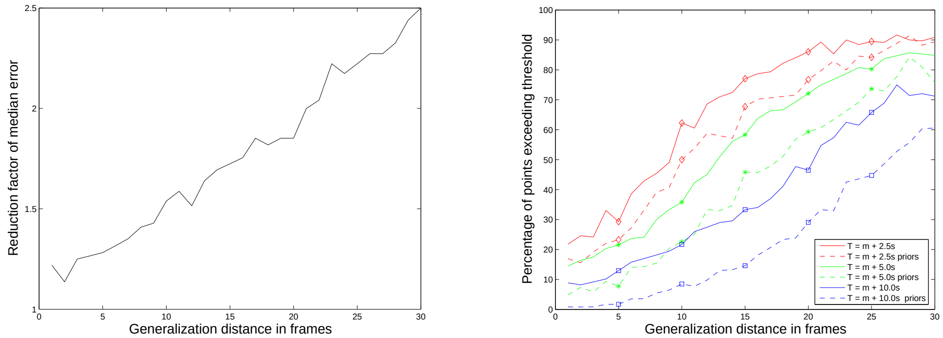

Figure 1: Results on synthetic data. Left: Average generalization improvement factor as a function of the generalization distance. Right: Percentage of point tracks with reprojection error exceeding three thresholds (see text), with and without prior use, as functions of the generalization distance.

Figure 2: Images from the 94-frames *Bears* sequence (left) and the 75-frames *Groundhog day* sequence (right) with marked points.

(originally banded) measurement matrices filled sub-matrices were extracted and a diagonal band with 50 % entries selected for training. The measurement matrices showed 94 and 75 frames with 94 and 117 point tracks. On the *Bears* sequence the camera smoothness measure was reduced by a factor of 108.7. The rank was estimated to 5. After initial estimation $\mathscr{E}_{\mathrm{RE}} = 0.82$ pixels. Applying the priors increased this to 1.20 pixels, a small payment for the improved generalization. Figure 3 shows plots of the percentage of point tracks as function of the generalization distance in frames, with and without use of the priors, and with reprojection error exceeding the previously described thresholds $\mu + k\sigma$, using $k = 2.5$, 5 and 10. Figure 3 shows that without prior use the generalization becomes bad even for short generalization distances. With prior use the error is significantly reduced. For the sequence *Bears* the generalization becomes possible at least up to a distance of 30 frames. On the more difficult sequence *Groundhog day* the camera smoothness measure was reduced by a factor of 5660.3. The rank was estimated to 14.

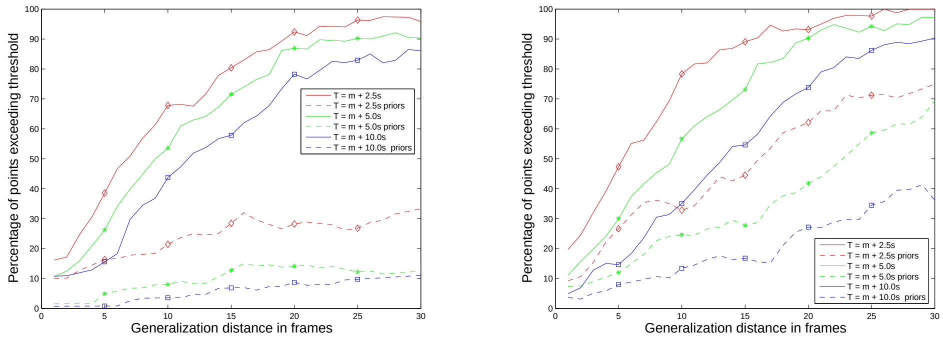

Figure 3: Percentages of point tracks in the sequences *Bears* (left) and *Groundhog day* (right) with reprojection error exceeding three thresholds (see text), with and without prior use, as functions of the generalization distance.

Figure 3 shows to the right that the generalization distance is increased by a factor of 2 to 4. This is still significant, but less impressive compared to the sequence *Bears*. A main reason is that a continuous surface is seen on the *Bears* sequence giving strength to the surface shape prior. This is not the case for the *Groundhog day* sequence.

In figure 4 a close-up of 4 tracks from the *Bears* sequence is shown. The positions

Figure 4: Close-up sequence of 4 point tracks which visible parts (use for training) all ended 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.

computed by using the two are much closer to the true positions than the ones obtained by not using the priors.

8 Conclusions

We proposed 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 in frames where data is missing, *i.e.* the generalization ability. Building on previous work the approach automatically estimates the rank of the measurement matrix, handles outliers and a substantial amount of missing data. Future work will show if the improved generalization allows detecting and gluing point tracks split because of imperfect tracking. We expect the temporal smoothness prior to drive the estimated model closer to an explicit configuration. Further work how much this will help in such ‘self-calibration’.

References

[1] H. Aanæs and F. Kahl. Estimation of deformable structure and motion. *The Vision and Modelling of Dynamic Scenes Workshop*, 2002.

[2] A. Bartoli and S. Olsen. A Batch Algorithm For Implicit Non-Rigid Shape and Motion Recovery. *Workshop on Dynamical Vision at ICCV'05*, 2005.

[3] M. Brand. Morphable 3D models from video. *Conf. on Computer Vision and Pattern Recognition*, 2001.

[4] M. Brand. A Direct Method for 3D Factorization of Nonrigid Motion Observed in 2D. *Conf. on Computer Vision and Pattern Recognition*, 2005.

[5] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. *Conf. on Computer Vision and Pattern Recognition*, 2000.

[6] A. M. Buchanan, A. W. Fitzgibbon Damped Newton Algorithms for Matrix Factorization with Missing Data. *Conf. on Computer Vision and Pattern Recognition*, 316–322, 2005.

[7] A. Del Bue, X. Lladó, L. de Agapito Non-Rigid Metric Shape and Motion Recovery from Uncalibrated Images Using Priors. *Conf. on Computer Vision and Pattern Recognition*, 1191–1198, 2006.

[8] D. W. Jacobs Linear Fitting with Missing Data for Structure-from-Motion. *Computer Vision and Image Understanding*, vol 82 no. 1, 57–81, 2001.

[9] D. Martinec and T. Pajdla 3D Reconstruction by Fitting Low-Rank Matrices with Missing Data. *Conf. on Computer Vision and Pattern Recognition*, pp. 198-205, 2005.

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

[11] P. H. S. Torr: Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *International Journal of Computer Vision*, 50(1):27–45, 2002.

[12] P. H. S. Torr, A. Zisserman: MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, vol. 78, 138–156, 2000.

[13] L. Torresani and C. Bregler. Space-time tracking. *European Conference on Computer Vision*, 2002.

[14] L. Torresani and A. Hertzmann. Automatic non-rigid 3D modeling from video. *European Conference on Computer Vision*, 2004.

[15] B. Triggs. Linear projective reconstruction from matching tensors. *Image and Vision Computing*, 15(8), 1997.

[16] R. Vidal and D. Abretske. Nonrigid Shape and Motion from Multiple Perspective Views. *European Conference on Computer Vision*, 2006.

[17] J. Xiao, J.-X. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. *European Conference on Computer Vision*, 2004.

[18] J. Xiao and T. Kanade. Non-rigid shape and motion recovery: Degenerate deformations. *International Conference on Computer Vision and Pattern Recognition*, 2004.