A Batch Algorithm For Implicit Non-Rigid Shape and Motion Recovery

Adrien Bartoli, Søren Olsen

WDV'05 – WORKSHOP ON DYNAMICAL VISION AT ICCV'05, BEIJING, CHINA, OCTOBER 2005

A Batch Algorithm For Implicit Non-Rigid Shape and Motion Recovery

Adrien Bartoli
CNRS - LASMEA, France — Adrien.Bartoli@gmail.com

Søren I. Olsen
DIKU, Danemark — ingvor@diku.dk

Abstract

The recovery of 3D shape and camera motion for non-rigid scenes from single-camera video footage is a very important problem in computer vision. The low-rank shape model consists in regarding the deformations as linear combinations of basis shapes. Most algorithms for reconstructing the parameters of this model along with camera motion are based on three main steps. Given point tracks and the rank, or equivalently the number of basis shapes, they factorize a measurement matrix containing all point tracks, from which the camera motion and basis shapes are extracted and refined in a bundle adjustment manner. There are several issues that have not been addressed yet, among which, choosing the rank automatically and dealing with erroneous point tracks and missing data.

We introduce theoretical and practical contributions that address these issues. We propose an implicit imaging model for non-rigid scenes from which we derive non-rigid matching tensors and closure constraints. We give a non-rigid Structure-From-Motion algorithm based on computing matching tensors over subsequences, from which the implicit cameras are extrated. Each non-rigid matching tensor is computed, along with the rank of the subsequence, using a robust estimator incorporating a model selection criterion that detects erroneous image points.

Preliminary experimental results on real and simulated data show that our algorithm deals with challenging video sequences.

1. Introduction

Structure-From-Motion – the recovery of 3D shape and camera motion from images – is one of the most studied problems in computer vision. The decades of work has led to significant successes, especially when the observed environment is static. However, the assumption of rigidity is violated in many cases of interest, for example expressive faces, moving cars, etc. For that reason, dealing with non-rigid scenes coming from single-camera footage has received an increasing attention over the last few years. The problem is highly challenging since both the camera motion and the non-rigid 3D shape have to be recovered. A major step forwards for such cases was made by Bregler *et al.* [5, 9], Brand [4] and Aanæs *et al.* [1]. Building on the work of [2, 7], they developed and demonstrated factorization of images of non-rigid scenes, where the non-rigidity was represented as a linear combination of *basis shapes*. Xiao *et al.* [14] studied the degenerate deformations that may defeat the reconstruction algorithms.

This paper tackles the two following open problems. *(i)* the factorization of a measurement matrix containing all point tracks in the presence of missing and erroneous image points. This must be done to recover the parameters of the implicit imaging model. Most previous work do not deal with missing data [1, 4, 5, 9, 13]. *(ii)* the automatic choice of the rank r of the measurement matrix, characterising the degree of non-rigidity in the sequence. Most previous work rely on a user-defined rank [4, 5, 9, 10, 13].

More precisely, we build on the low-rank shape model to derive an *implicit imaging model* projecting points affinely from $\mathbb{R}^r$ – the implicit shape points – onto the images using *implicit camera matrices*. The rank r reflects the degree of non-rigidity of the model and is thus a very important parameter. This implicit model is simpler than the *explicit model* used in *e.g.* [5, 10], in the sense that it ignores the replicated block structure of the camera matrices. The implicit model gives weaker constraints on point tracks than the explicit model. It is the model used for non-rigid factorization in *e.g.* [5, 9, 13]. Based on this model, we derive *non-rigid matching tensors* that constrain point tracks and encapsulate information about the implicit camera matrices. We define non-rigid closure constraints relating the matching tensors to the implicit camera matrices. These theoretical concepts are based on the fact that implicit reconstruction is performed in $\mathbb{R}^r$. They lead to a batch algorithm for computing the motion and structure matrices in the presence of erroneous and missing data. The idea is to robustly compute a set of matching tensors over several subsequences using MAPSAC and the GRIC criterion to choose the associated rank [8]. From these matching tensors, we solve for the implicit camera matrices using the closure constraints. The next step consists in computing the basis shapes by non-rigid triangulation. We refine both the implicit cameras and implicit shape in a bundle adjustment manner. Finally, each image point is classified as an inlier or an outlier. Almost all steps in this algorithm are done robustly, meaning that blunders are detected and thus do not corrupt the computation.

Roadmap. In §2, we derive the non-rigid shape and imaging models. We examine previous work in §3. We derive the non-rigid matching tensors and closure constraints in §§4 and 5 respectively. Our Structure-From-Motion algorithm is derived in §6 while the robust estimation of matching tensors and associated ranks is given in §7. Experimental results are reported in §8 and our conclusions in §9.

Notation. Vectors are denoted using bold fonts, *e.g.* $\mathbf{x}$ and matrices using sans-serif or calligraphic characters, *e.g.* M or $\mathcal{X}$. Index $i = 1, \ldots, n$ is used for the images, $j = 1, \ldots, m$ for the points and $k = 1, \ldots, l$ for the basis shapes, *e.g.* $\mathbf{x}_{ij}$ is the position of the j-th point track in the i-th image and $\mathbf{B}_{kj}$ is the k-th basis shape for the j-th point. Visibility indicators modeling occlusions are denoted v_{ij}. The Hadamard (element-wise) product is written $\odot$. The zero and one vectors are respectively $\mathbf{0}$ and $\mathbf{1}$, $\mathsf{0}$ is the zero matrix and $^\mathsf{T}$ is vector and matrix transpose. Bars indicate centred data, as in *e.g.* $\bar{\mathcal{X}}$. Notation $[i, i']$ refers to a subsequence between image i and image i', *e.g.* $\mathcal{X}_{[i,i']}$ is the measurement matrix for this subsequence. $\{\}$ is a set over some variable. We use the Singular Value Decomposition, denoted SVD, *e.g.* $\mathcal{X} = \mathsf{U}\Sigma\mathsf{V}^\mathsf{T}$ where U and V are orthonormal matrices, and Σ is diagonal, containing the singular values of $\mathcal{X}$ in decreasing order.

Noise distribution. The noise on image point positions is supposed to be centred Gaussian i.i.d. Under this hypothesis, minimizing the $\mathcal{L}_2$-norm between measured and predicted point positions, often dubbed the reprojection error, yields Maximum Likelihood Estimates.

2. Non-Rigid Imaging Model

We review the low-rank shape model, dubbed the explicit model and derive our implicit model.

2.1. Explicit Model

The low-rank shape assumption consists in writing the coordinates of a time-varying set of points $\mathbf{Q}_{ij}$ as linear combinations over l *basis shapes* $\mathbf{B}_{kj}$ with the *configuration weights* α_{ik}: $\mathbf{Q}_{ij} = \sum_{k=1}^{l} \alpha_{ik}\mathbf{B}_{kj}$. Points $\mathbf{Q}_{ij}$ are projected onto the images by affine cameras: $\mathbf{x}_{ij} = \mathsf{P}_i\mathbf{Q}_{ij} + \mathbf{t}_i$, from which the explicit imaging model is obtained:

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

This trilinear equation is the most explicit form of the low-rank shape imaging model. Only rank-3 basis shapes are considered for simplicity, but rank-2 and rank-1 basis shapes can be modeled as well [14].

2.2. Implicit Model

Rewriting (1), one obtains:

$$\begin{aligned} \mathbf{x}_{ij} &= \begin{pmatrix} \alpha_{i1}\mathsf{P}_i & \cdots & \alpha_{il}\mathsf{P}_i \end{pmatrix} \begin{pmatrix} \mathbf{B}_{1j} \\ \vdots \\ \mathbf{B}_{lj} \end{pmatrix} + \mathbf{t}_i \\ &= \mathsf{M}_i\mathbf{S}_j + \mathbf{t}_i \qquad \text{with} \\ \mathsf{M}_i &= \begin{pmatrix} \alpha_{i1}\mathsf{P}_i & \cdots & \alpha_{il}\mathsf{P}_i \end{pmatrix}. \end{aligned} \tag{2}$$

We call M_i a $(2 \times 3l)$ *explicit camera matrix* and $\mathbf{S}_j^\mathsf{T} = \begin{pmatrix} \mathbf{B}_{1j}^\mathsf{T} & \cdots & \mathbf{B}_{lj}^\mathsf{T} \end{pmatrix}$ a $(3l \times 1)$ *shape vector*. Introduce $r = 3l$, the rank of the model, a $(r \times r)$ full-rank matrix $\mathcal{A}$ and relaxing the replicated structure yields the bilinear *implicit model*. From (2), $\mathbf{x}_{ij} = \mathsf{M}_i\mathbf{S}_j + \mathbf{t}_i = \left(\mathsf{M}_i\mathcal{A}^{-1}\right)\left(\mathcal{A}\mathbf{S}_j\right) + \mathbf{t}_i$, giving:

$$\mathbf{x}_{ij} \quad = \quad \mathsf{J}_i\mathbf{K}_j + \mathbf{t}_i. \tag{3}$$

We call $\mathsf{J}_i = \mathsf{M}_i\mathcal{A}^{-1}$ and $\mathbf{K}_j = \mathcal{A}\mathbf{S}_j$ the *implicit camera matrix* and the *implicit shape matrix* respectively. Matrix $\mathcal{A}$ represents a *corrective transformation*. As shown in the next section, this is the model used for non-rigid factorization. The model generalizes, in some sense, the $\mathbb{P}^k \to \mathbb{P}^2$ projection matrices introduced by Wolf *et al.* [12].

3. Previous Work

Most of the previous work [1, 4, 5, 9, 13] is based on factorizing a measurement matrix using SVD and hence do not cope with missing data. We note that Torresani *et al.* [10] propose an approach where the likelihood of the explicit model is maximized over the entire image sequence using a generalized EM (Expectation Maximization) algorithm which finds the nearest local optimum. The important rank selection problem is neglected in most papers, besides [1]. Below, we describe the three main steps involved in most algorithms. The inputs are the complete measurement matrix $\mathcal{X}$ and the rank r. The outputs are the camera pose, the configuration weights and the basis shapes.

Step 1: Factorizing. A $(2n \times m)$ measurement matrix $\mathcal{X}$ is built by gathering all point coordinates. The translation part of the imaging model, *i.e.* the $\mathbf{t}_i$, is estimated as the mean of the point coordinates in each image. A $(2n \times 1)$ joint translation vector $\mathbf{t}^\mathsf{T} = \begin{pmatrix} \mathbf{t}_1^\mathsf{T} & \cdots & \mathbf{t}_n^\mathsf{T} \end{pmatrix}$ is built and used to centre the measurement matrix: $\bar{\mathcal{X}} \leftarrow \mathcal{X} - \mathbf{t} \cdot \mathbf{1}^\mathsf{T}$, from which we get:

$$\underbrace{\begin{pmatrix} \mathbf{x}_{11} & \cdots & \mathbf{x}_{1m} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_{n1} & \cdots & \mathbf{x}_{nm} \end{pmatrix}}_{\bar{\mathcal{X}}_{(2n \times m)}} = \underbrace{\begin{pmatrix} \mathsf{J}_1 \\ \vdots \\ \mathsf{J}_n \end{pmatrix}}_{\mathcal{J}_{(2n \times r)}} \underbrace{\begin{pmatrix} \mathbf{K}_1 & \cdots & \mathbf{K}_m \end{pmatrix}}_{\mathcal{K}_{(r \times m)}},$$

where $\mathcal{J}$ and $\mathcal{K}$ are the joint implicit camera and shape matrices. The centred measurement matrix is factorized using SVD as $\bar{\mathcal{X}} = \mathsf{U}\Sigma\mathsf{V}^\mathsf{T}$. The joint implicit camera and shape matrices $\mathcal{J}$ and $\mathcal{K}$, are recovered as the r leading columns of *e.g.* U and $\Sigma\mathsf{V}^\mathsf{T}$ respectively.

Step 2: Upgrading. The implicit model is upgraded to the explicit one by computing a corrective transformation. Xiao *et al.* [13] show that constraints on both the explicit camera 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. Previous work [4, 5, 9] use only the rotation constraints, leading to ambiguous solutions. For instance, Brand [4] shows that a block-diagonal corrective transformation is a good practical approximation. Once the replicated structure has been approximately enforced, the rotation matrices are extracted using orthonormal decomposition. The configuration weights are then recovered using the orthonormality of the rotation matrices. Bregler *et al.* [5] assume that the information about each basis shape is distributed in the appropriate column triple in the shape matrix by the initial SVD, in other words that the entries off the block-diagonal of the corrective transformation matrix are negligible. Experiments show that this assumption restricts the cases that can be dealt with since only limited non-rigidity can be handled. A second factorization round on the reordered weighted motion matrix elements enforces the replicated block structure, yielding the weight factors and the P_i, which are upgraded to Euclidean by computing a linear transformation as in the rigid factorization case. Aanæs *et al.* [1] assume that the structure resulting from rigid factorization gives the mean non-rigid structure and camera motion. Given the camera motion, recovering the structure is done by examining the principal components of the estimated variance.

Step 3: Nonlinear refinement. The solution obtained so far is finely tuned in a bundle adjustment manner by minimizing *e.g.* the reprojection error. The algorithms proposed in [4, 9] differ by the prior they are using to regularize the solution. These priors state that the reconstructed shapes should not vary too much between consecutive images.

4. Non-Rigid Matching Tensors

Matching tensors are known for the rigid case. Examples are the fundamental matrix and the trifocal tensor. They relate the image position of corresponding points over multiple images. The implicit imaging model allows us to derive matching tensors for non-rigid scenes. These tensors are briefly mentioned in [6, §18.3.1].

A non-rigid matching tensor is a matrix $\mathcal{N}$ whose columns span the d dimensional nullspace of the $(2n \times m)$ centred measurement matrix $\bar{\mathcal{X}}$:

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

The size of matrix $\mathcal{N}$ is $(2n \times d)$ where the tensor dimension is $d = 2n - r$. Loosely speaking, $\mathcal{N}$ constrain each point track $\bar{\mathbf{x}}_j$ – the j-th column of $\bar{\mathcal{X}}$ – by $\mathcal{N}^\mathsf{T}\bar{\mathbf{x}}_j = \mathbf{0}$. These constraints easily extend to the non centred measurement matrix $\mathcal{X}$ by substituting $\bar{\mathcal{X}} = \mathcal{X} - \mathbf{t} \cdot \mathbf{1}^\mathsf{T}$ into equation (4):

$$\begin{pmatrix}\mathcal{N}^\mathsf{T} & -\mathcal{N}^\mathsf{T}\mathbf{t}\end{pmatrix}\begin{pmatrix}\mathcal{X} \\ \mathbf{1}^\mathsf{T}\end{pmatrix} \quad = \quad \mathsf{0}.$$

Minimal number of points and views. The three following parameters are characteristic of an image sequence: the number of images n, the number of point tracks m and the rank r. They can be related to each other, in particular for, given r, deriving what the minimal number of point tracks and views are for computing the matching tensor. The computation is possible if the $(2n \times m)$ centred measurement matrix $\bar{\mathcal{X}}$ is at least of size $(r \times r)$. Counting the point track needed to compute the translations for centring the measurement matrix, we directly get the minimal number of point tracks as $m \geq r + 1$. From $2n \geq r$, we obtain the minimal number of views as $n \geq \lfloor \frac{r}{2} \rfloor + 1$. These numbers can also be derived by counting the number of degrees of freedom in the tensor and the number of independent constraints given by equation (4).

Example: 2D rigid scene. In this case, $r = 2$ and pairs of points are related by a 2D affine transformation that can be estimated from 3 point correspondences. With centred coordinates, the relationship is $\bar{\mathbf{x}}_{2j} = \mathsf{A}\bar{\mathbf{x}}_{1j}$, *i.e.* :

$$\underbrace{\begin{pmatrix}\mathsf{A} & -\mathrm{I}\end{pmatrix}}_{\mathcal{N}^\mathsf{T}}\begin{pmatrix}\bar{\mathbf{x}}_{1j} \\ \bar{\mathbf{x}}_{2j}\end{pmatrix} \quad = \quad \mathbf{0},$$

from which we observe that the matching tensor has size (4×2). More generally, even-rank matching tensors predict an image point given all other $n - 1$ image points.

Example: 3D rigid scene. In this case, $r = 3$ and pairs of points are related by the affine fundamental matrix that can estimated from 4 point correspondences. With centred coordinates, the relationship is $(\bar{\mathbf{x}}_2^\mathsf{T} \;\; 1)\bar{\mathsf{F}}_A(\bar{\mathbf{x}}_1^\mathsf{T} \;\; 1)^\mathsf{T} = 0$ with $\bar{\mathsf{F}}_A = \left(\begin{smallmatrix}0 & 0 & a \\ 0 & 0 & b \\ c & d & 0\end{smallmatrix}\right)$ the centred affine fundamental matrix:

$$\underbrace{\begin{pmatrix}c & d & a & b\end{pmatrix}}_{\mathcal{N}^\mathsf{T}}\begin{pmatrix}\bar{\mathbf{x}}_{1j} \\ \bar{\mathbf{x}}_{2j}\end{pmatrix} \quad = \quad \mathbf{0}.$$

More generally, odd-rank matching tensors predict the equivalent of an epipolar line in an image given all other $n - 1$ image points.

OBJECTIVE

Given m point tracks over n images as a an incomplete $(2n \times m)$ measurement matrix $\mathcal{X}$ and a $(n \times m)$ visibility matrix $\mathcal{V}$, compute the implicit non-rigid cameras J_i, the non-rigid shape points $\mathbf{K}_j$ and the rank r.

ALGORITHM

1. Partition the sequence, see §6.1 while robustly computing the matching tensors $\{\mathcal{N}_{[i_b,i'_b]}\}$ and associated ranks, see §7.2.
2. Solve for the implicit cameras $(\mathsf{J}_i, \mathbf{t}_i)$ using the closure constraints, see §6.2.
3. Triangulate the point tracks to get the implicit shape points $\mathbf{K}_j$, see §6.3.
4. Nonlinearly refine the implicit cameras and shape points by minimizing the reprojection error, see §6.4.
5. Classify each image point track as an inlier or an outlier.

Table 1: Summary of our non-rigid implicit Structure-From-Motion algorithm.

5. Non-Rigid Closure Constraints

The closure constraints introducted by Triggs in [11] relate matching tensors to projection matrices. These constraints are used to derive a batch Structure-From-Motion algorithm dealing with high amounts of missing data.

In this section, we derive new types of closure constraints for the non-rigid case, based on the above-derived matching tensors, namely the $\mathcal{N}$-closure. Our derivation is valid for any rank r.

Let $\mathbf{K} \in \mathbb{R}^r$ be an implicit shape point. We project $\mathbf{K}$ in the images using the joint implicit camera matrix $\mathcal{J}$: $\bar{\mathbf{x}} = \mathcal{J}\mathbf{K}, \forall \mathbf{K} \in \mathbb{R}^r$. From the definition (4) of the matching tensors, $\mathcal{N}^\mathsf{T}\bar{\mathbf{x}} = \mathbf{0}$. Substituting the joint projection equation yields $\mathcal{N}^\mathsf{T}\mathcal{J}\mathbf{K} = \mathbf{0}, \forall \mathbf{K} \in \mathbb{R}^r$, which gives the *$\mathcal{N}$-closure constraint*:

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

This constraint means that the joint implicit camera matrix lies in the right nullspace of $\mathcal{N}^\mathsf{T}$.

6. Non-Rigid Structure-From-Motion

Our batch algorithm for implicit non-rigid Structure-From-Motion is based on the above-derived non-rigid matching tensors and closure constraints. It is summarized in table 1. We consider only sets of consecutive images for simplicity. It begins by selecting a set of s subsequences $\{[i_b, i'_b]\}_{b=1}^{b=s}$ and by computing a set of matching tensors $\{\mathcal{N}_{[i_b,i'_b]}\}$, one for each subsequence, and the associated rank estimates $\{r_{[i_b,i'_b]}\}$. Our joint tensor and rank estimation algorithm is presented in §7. The full sequence rank r is the maximum over all subsequence ranks: $r = \max_b(r_{[i_b,i_b]})$.

6.1. Partitioning the Sequence

The measurement matrix is partitioned into overlapping blocks with points visible in all of the selected images. Before going into further details, we must figure out what the minimal tensor dimension is, and how many views each tensor should operate on. Let $[i_b, i'_b]$ and $[i_{b+1}, i'_{b+1}]$ be two consecutive subsequences and let $\delta_{b,b+1} = i_{b+1} - i_b$ be the offset between them. We need to determine what the maximum value of $\delta_{b,b+1}$ is. The b-th matching tensor, with dimension $d_b = 2n_b - r_b$, gives d_b constraints. The number of unknowns constrained by the first matching tensor only is $\delta_{1,2}$, from which we get $\delta_{1,2} \leq n_1 - \lfloor \frac{r_1+1}{2} \rfloor$. Making the same reasoning for the b-th tensor, *i.e.* ignoring the constraints coming from previous overlapping sets, gives a bound on $\delta_{b,b+1}$:

$$\delta_{b,b+1} \quad \leq \quad n_b - \lfloor \frac{r_b + 1}{2} \rfloor. \tag{6}$$

Taking into account the other constraints lead to a tighter bound on $\delta_{b,b+1}$, but requires a cumbersome formalism to count the number of constraints and unknowns. Requiring $\delta_{b,b+1} > 0$ gives the minimal size of each image set as:

$$n_b \quad \geq \quad \lfloor \frac{r_b + 1}{2} \rfloor + 1. \tag{7}$$

For instance, for a 2D rigid scene, *i.e.* $r = 2$, the minimal n_b is 2 from equation (7) and the maximal $\delta_{b,b+1}$ is 1 from equation (6), *i.e.* using the affine transformations over pairs of consecutive views is fine. For a 3D rigid scene, *i.e.* $r = 3$, the minimal n_b is 3 and the maximal $\delta_{b,b+1}$ is 1, meaning that using trifocal tensors over triplets of consecutive of views is fine[1].

In practice, we do not know the ranks r_b at this step. We tune an initial guess while jointly partitioning the sequence and computing the matching tensors, as described in §7.2.

6.2. Solving For the Implicit Cameras

The leading part. We solve for the non-rigid cameras using the closure constraints. For each computed matching tensor, equation (5) gives the following constraints on the joint camera matrix $\mathcal{J}$:

$$\left(0_{(d_b \times 2(i_b-1))} \quad \mathcal{N}^\mathsf{T}_{[i_b,i'_b]} \quad 0_{(d_b \times 2(n-i'_b))}\right) \mathcal{J} \quad = \quad 0.$$

Stacking the constraints for all $\{[i_b, i'_b]\}_{b=1}^{b=s}$ yields an homogeneous system $\mathsf{A}\mathcal{J} = 0$. It must be solved, *e.g.* in the least-squares sense, while ensuring that matrix $\mathcal{J}$ has full

[1] Triggs [11] states this result and shows the equivalence of using pairs of fundamental matrices over triplets of consecutive views.

column rank: $\min_{\mathcal{J}} \|\mathsf{A}\mathcal{J}\|^2$ s.t. $\det(\mathcal{J}) \neq 0$. We replace the full column rank constraint by a column orthonormality constraint, *i.e.* $\mathcal{J}^\mathsf{T}\mathcal{J} = \mathrm{I}_{(r\times r)}$. Note that the latter implies the former. This is done without loss of generality since for any full column rank joint camera matrix $\mathcal{J}$, there exist several coordinate transformations, say $\mathsf{G}_{(r\times r)}$, such that $\mathcal{J}\mathsf{G}$ is column orthonormal. One such a transformation is given by the QR decomposition of $\mathcal{J} = \mathcal{J}'\mathsf{G}^{-1}$. The transformed problem is solved by using the SVD $\mathsf{A} = \mathsf{U}\Sigma\mathsf{V}^\mathsf{T}$. Matrix $\mathcal{J}$ is given by the r last columns of V. Note that matrix A typically has a band-diagonal shape that one might exploit to efficiently compute its singular vectors, see *e.g.* [3].

The translations. The implicit imaging model (3) is $\mathbf{x}_{ij} = \mathsf{J}_i\mathbf{K}_j + \mathbf{t}_i$. By minimizing a least-squares error over all image points, the translations $\mathbf{t}_i$ in the joint translation vector $\mathbf{t}$, along with the basis shape vectors $\mathbf{K}_j$ can be reconstructed. We prefer to postpone the basis shape vector reconstruction to the next step, for robustness purposes. Instead, we consider the translation estimate $\mathbf{y}_{[i,i']}$ for each subsequence $[i, i']$, giving the centroid with respect to the points visible in the subsequence. We reconstruct these centroids along with vector $\mathbf{t}$. Note that in the absence of missing data, these centroids coincide. We minimize the reprojection error $\sum_{b=1}^{s} \|\mathbf{y}_{[i_b,i'_b]} - \mathcal{J}_{[i_b,i'_b]}\mathbf{Y}_{[i_b,i'_b]} - \mathbf{t}_{[i_b,i'_b]}\|^2$, where $\mathcal{J}_{[i,i']}$ and $\mathbf{t}_{[i,i']}$ are respectively a partial joint projection matrix and a partial joint translation vector restricted to the subsequence $[i, i']$, and $\mathbf{Y}_{[i,i']}$ is the reconstructed centroid. By expanding the cost function, the reprojection error is rewritten $\|\mathsf{A}\mathbf{w} - \mathbf{b}\|^2$, where the unknown vector $\mathbf{w}$ contains the $\mathbf{Y}_{[i_b,i'_b]}$ and $\mathbf{t}$. The solution is given by using the pseudo-inverse of matrix A, as $\mathbf{w} = \mathsf{A}^\dagger\mathbf{b}$. One must use a pseudo-inverse, since there is a r-dimensional ambiguity, making A rank deficient with a left nullspace of dimension r. This is a translational ambiguity between the basis shapes and the joint translation $\mathbf{t}$, that one can see by considering that $\forall\boldsymbol{\gamma} \in \mathbb{R}^r$, $\mathbf{x}_j = \mathcal{J}\mathbf{K}_j + \mathbf{t} = \mathcal{J}(\mathbf{K}_j - \boldsymbol{\gamma}) + \mathcal{J}\boldsymbol{\gamma} + \mathbf{t} = \mathcal{J}\mathbf{K}'_j + \mathbf{t}'$, with $\mathbf{K}'_j = \mathbf{K}_j - \boldsymbol{\gamma}$ and $\mathbf{t}' = \mathcal{J}\boldsymbol{\gamma} + \mathbf{t}$.

6.3. Reconstructing the Implicit Shape Points

We compute the basis shape vectors by non-rigid triangulation. This is done by minimizing the reprojection error. Assume that the j-th point is visible in the subsequence $[i, i']$, then this is formulated by:

$$\min_{\mathbf{K}_j} \|\bar{\mathbf{x}}_{[i,i']} - \mathcal{J}_{[i,i']}\mathbf{K}_j\|^2,$$

with $\bar{\mathbf{x}}_{[i,i']} = \mathbf{x}_{[i,i']} - \mathbf{t}_{[i,i']}$. The solution is $\mathbf{K}_j = \mathcal{J}^\dagger_{[i,i']}\bar{\mathbf{x}}_{[i,i']}$. We perform the minimization in a robust manner to eliminate erroneous image points. We use a RANSAC-like algorithm with adaptive number of trials. The number of image points sampled in the inner loop is $\lfloor \frac{r}{2} \rfloor + 1$.

6.4. Nonlinear Refinement

We complete the reconstruction algorithm by minimizing the reprojection error in order to finely tune the estimate:

$$\min_{\mathcal{J},\mathbf{t},\mathcal{K}} \|\mathcal{V}^+ \odot (\mathcal{X} - \mathcal{J}\mathcal{K} - \mathbf{t}\cdot\mathbf{1}^\mathsf{T})\|^2,$$

where $\mathcal{V}^+$ is obtained by duplicating[2] each row of the $(n \times m)$ visibility matrix $\mathcal{V}$. The minimization is done in a bundle adjustment manner. More precisely, we use a damped Gauss-Newton algorithm with a robust kernel. The damping is important to avoid singularities in the Hessian matrix, due to the $r(r+1)$ dimensional coordinate frame ambiguity. Contrarily to the explicit case, see [1, 13], no extra regularizing constraint is necessary.

7. Estimating the Non-Rigid Matching Tensors and Ranks

Our method estimates a non-rigid matching tensor over a (sub)sequence, *i.e.* for a complete measurement matrix, in a Maximum Likelihood framework. First, we tackle the case where the data do not contain outliers, and when the rank is given. Second, we examine the case where the data may contain outliers, and when the rank have to be estimated.

7.1. Outlier-Free Data, Known Rank

We describe a Maximum Likelihood Estimator, that handles minimal and redundent data. The translation $\mathbf{t}$ is obtained by averaging the point positions, and the measurement matrix is then centred as $\bar{\mathcal{X}} = \mathcal{X} - \mathbf{t}\cdot\mathbf{1}^\mathsf{T}$. The problem of finding the optimal $\mathcal{N}$ is formulated by $\min_{\hat{\mathcal{X}}} \|\bar{\mathcal{X}} - \hat{\mathcal{X}}\|^2$ s.t. $\mathcal{N}^\mathsf{T}\hat{\mathcal{X}} = \mathsf{0}$, where $\hat{\mathcal{X}}$ contains predicted point positions. This is a matrix approximation problem under rank deficiency constraint. It is solved by computing the SVD $\bar{\mathcal{X}} = \mathsf{U}\Sigma\mathsf{V}^\mathsf{T}$, from which $\hat{\mathcal{X}}$ is obtained by nullifying all but the r leading singular values in Σ and recomposing the SVD. Matrix $\mathcal{N}$ is given by the $2n - r$ last columns of U.

7.2. Contaminated Data, Unknown Rank

In most previous work, the rank of the sequence is assumed to be given. One exception is Aanæs *et al.* [1] who use the BIC model selection criterion to select the rank, but do not deal with blunders. When one uses subsequences, the subsequence rank may be lower than the sequence rank, and must be estimated along with the matching tensor. In addition, one has to deal with erroneous image points. We propose to use the robust estimator MAPSAC in conjunction with the GRIC model selection criterion proposed in [8]. GRIC is a modified BIC for robust least-squares problems. Our algorithm maximizes the GRIC score, as follows. In the inner

[2] This is simply to make it the same size as $\mathcal{X}$.

loop of the robust estimator, we sample point tracks and not only compute a single matching tensor, but multiple ones by varying the rank. Obviously, an upper bound r_{max} on the rank is necessary to fix the number of point tracks that one samples at each trial. One must take into account that the computational cost rises with r_{max}. One possible solution is to divide the sequence of trials into groups using gradually narrower intervals of possible rank values. The GRIC score is given by:

$$\text{GRIC} = \sum_{j=1}^{m} \rho\left(\frac{e_j^2}{\sigma^2}\right) + \lambda d + rm\log(m),$$

where e_j is the prediction error for the j-th point track, $\lambda = 4d\log(z) - \log(2\pi\sigma^2)$ and z is chosen as the image side length. Function ρ is $\rho(x) = x$ for $x < t$ and $\rho(x) = t$ otherwise, where the threshold $t = 2\log(\theta) + d\lambda/(2n)$ with θ the ratio of the percentage of inliers to the percentage of outliers. The noise level is robustly estimated using the weakest model, *i.e.* for a tensor dimension $d = 1$, as $\sigma^2 = \text{med}(e_j^2)/0.6745^2$. We refer the reader to [8] for more details.

8. Experimental Results

Most other methods do not handle missing data, and hence can not be compared to our. The method from Torresani *et al.* [10] handles missing data but uses the explicit model.

8.1. Simulated Data

We simulated $n = 180$ cameras observing a set of $m = 1000$ points generated from $l = 5$ basis shapes, hence with rank $\underline{r} = 3l = 15$. The configuration weights are chosen in order to give a decaying energy to successive deformation modes. The simulation setup produces a complete measurement matrix $\tilde{\mathcal{X}}$, from which we extract a sparse, band-diagonal measurement matrix $\mathcal{X}$, similar to what a real intensity-based point tracker would produce. A Gaussian centred noise with variance $\sigma^2 = 1$ is added to the image points.

In the experiments, we measured the *reprojection error* and the *generalization error*, which are dubbed in a machine learning context *training* and *test* error respectively. The reprojection error is $\mathcal{E} = \sqrt{\frac{1}{e}\|\mathcal{V}^+ \odot (\mathcal{X} - \mathcal{J}\mathcal{K} - \mathbf{t}\cdot\mathbf{1}^\mathsf{T})\|^2}$, where e is the total number of visible image points. In other words, the reprojection error reflects the difference between the measures and the predictions. The generalization error is given by $\mathcal{G}_\gamma = \sqrt{\frac{1}{e_\gamma}\|\tilde{\mathcal{V}}_\gamma^+ \odot (\tilde{\mathcal{X}} - \mathcal{J}\mathcal{K} - \mathbf{t}\cdot\mathbf{1}^\mathsf{T})\|^2}$, where γ indicates the percentage of hidden image points in $\tilde{\mathcal{X}}$ involved in the estimation and e_γ is the total number of image points used in the calculation. The $(n \times m)$ matrix $\tilde{\mathcal{V}}_\gamma$ indicates which image points are used in the calculation: it is constructed by including points further away from the visible points area while γ grows, *i.e.* $\tilde{\mathcal{V}}_0 = \mathcal{V}$ and $\tilde{\mathcal{V}}_{100} = 1_{(n\times m)}$. For example, $\mathcal{G}_0 = \mathcal{E}$ and $\mathcal{G}_{100} = \sqrt{\frac{1}{nm}\|\tilde{\mathcal{X}} - \mathcal{J}\mathcal{K} - \mathbf{t}\cdot\mathbf{1}^\mathsf{T}\|^2}$, *i.e.* all the visible and hidden image points are used to compute the error. Obviously, we expect the generalization error to be greater than the reprojection error, and to grow with γ.

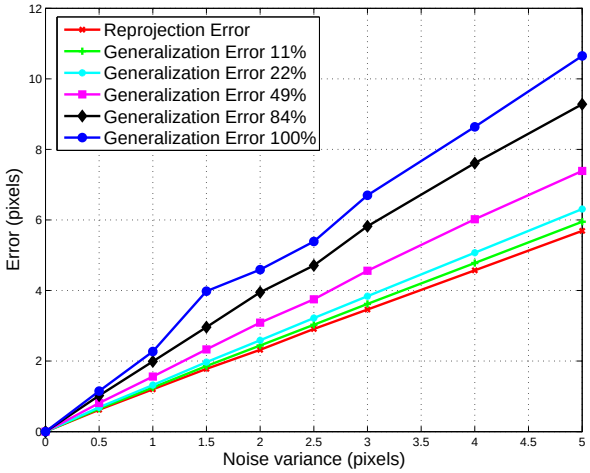

Figure 1: Reprojection and generalization error versus the variance of added noise σ for different percentages γ of hidden points to compute the generalization error.

The first experiment we performed consists in varying the level of added noise σ for different percentages γ of hidden points to compute the generalization error. The results are shown on figure 1. We observed that the reprojection error is slightly higher than the level of noise. The ability to generalize is accurate for a 1 pixel noise level, and smoothly degrades for larger noise levels, but is still reasonable: in the tested rang $\sigma = 0, \ldots, 5$ pixels, the $\gamma = 100\%$ generalization error is slightly higher than twice the noise level.

The second experiment we performed consists in varying the rank used in the computation, namely we tested $r = 11, \ldots, 27$, for different percentages γ of hidden points to compute the generalization error. The results are shown on figure 2. We observed that it is preferable to overestimate rather than to underestimate the rank, up to some upper limit. A similar experiment with roughly equal magnitude configuration weights to generate the data shows that r can be slightly underestimated and largely overestimated. The conclusion is that in practice, overestimating the rank is safe.

The third experiment is devised to assess the quality of the rank estimation based on GRIC in the presence of outliers. We tested for true ranks in the range $r = 3, \ldots, 18$ which covers what one expects to meet in practice. The results we obtained are shown in table 2, which shows av-

	3	6	9	12	15	18
0%	3.82	6.06	8.48	11.28	13.82	16.22
10%	3.86	6.02	8.60	11.02	13.66	16.24
20%	3.72	5.98	8.48	11.20	13.84	16.44
30%	3.64	5.94	8.52	11.00	13.52	16.58
40%	3.60	5.98	8.44	11.00	13.58	16.28
50%	3.40	5.88	8.30	10.86	13.68	16.16

	3	6	9	12	15	18
0%	0.38	0.42	0.57	0.66	0.65	1.12
10%	0.35	0.37	0.49	0.65	0.55	1.14
20%	0.45	0.37	0.50	0.60	0.58	0.50
30%	0.48	0.37	0.57	0.53	0.61	0.67
40%	0.49	0.32	0.57	0.53	0.64	1.08
50%	0.49	0.62	0.70	0.63	0.71	1.17

Table 2: (left) Average estimated rank r and (right) its standard deviation σ_r versus the true rank $\underline{r}$ and percentage of outliers.

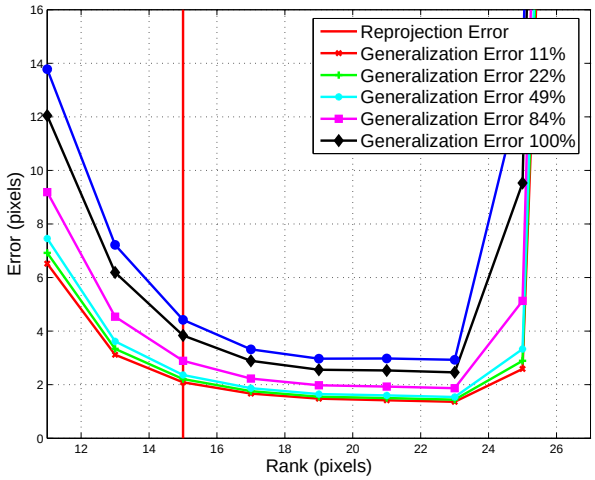

Figure 2: Reprojection and generalization error versus the rank r for different percentages γ of hidden points to compute the generalization error. The true rank $\underline{r} = 15$ is indicated with a vertical bar.

erages over 50 trials. We observed that these results are acceptable, even if the GRIC criterion we used is slightly biased since low ranks, *i.e.* less than 6, are slightly overestimated, while larger ranks, *i.e.* greater than 9 are slightly underestimated. It is however possible to correct for this bias in accordance with our conclusions on the previous experiment.

8.2. Real Data

We tested our algorithm on several image sequences. For one of them, extracted from the movie 'Groundhog Day', we show results. The sequence shows a man driving a car with a groundhog seated on his knees. The head of the man is rotating and deforming since he is speaking, and the animal is looking around, deforming its fur, opening and closing its mouth. Finally, the interior of the car is almost static, while the exterior is rigid, but moving with respect to the car.

The sequence contains 154 images, see figure 3 (top). We ran a KLT-like point tracker. We obtained a total of 1502 point tracks after having removed the small point tracks, namely which last less than 20 views. The visibility matrix, shown on figure 3 (bottom) is filled to 29.58%.

Figure 4: One frame with points and motion vectors reprojected from the reconstructed model.

For some parts of the sequence, where the motion of the different moving and deforming parts in the images is slow, computing the matching tensors is quite easy. Indeed, blunders can clearly be detected and classified as outliers. However, other parts in the sequence contain significant motion between single frames and motion blur occurs, making the point tracks slightly diverging from their 'true' position, and making the detection of outliers difficult. Large illumination changes sometimes make the tracker fails for entire areas of the image.

The reprojection errors we obtained at the non-rigid matching tensors estimation stage were distributed between 0.5 and 0.9 pixels, and 0.65 pixels on average. We used a user-defined rank $r = 15$. The initialization step yielded 58021 inliers over 68413 image points, *i.e.* the inlier rate was 84.8%, with a reprojection error of 1.19 pixels. The robust bundle adjustment yielded 61151 inliers, *i.e.* the inlier rate was 89.4%, with a reprojection error of 0.99 pixels. We

Figure 3: (top) 5 out of the 154 frames and (bottom) the visibility matrix $\mathcal{V}$ for the 'Groundhog Day' sequence.

believe it is a successful result on this challenging image sequence.

Figure 5: Closeup on the actor, the groundhog and the background overlaid with points and motion vectors reprojected from the reconstructed model (white dots), original points (light grey squares) and outliers (dark grey diamonds).

9. Conclusions

We proposed an implicit imaging model for non-rigid scenes, from which we derived non-rigid matching tensors and closure constraints. Based on these theoretical concepts, we proposed a robust batch implicit Structure-From-Motion algorithm for monocular image sequences of non-rigid scenes, dealing with missing data and blunders. Future work will be devoted to comparing various model selection criteria, and segmenting the scene based on the configuration weights, to recover objects that move or deform independently.

References

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

[2] B. Bascle and A. Blake. Separability of pose and expression in facial tracing and animation. In *Proceedings of the International Conference on Computer Vision*, 1998.

[3] A. Björck. *Numerical Methods For Least-Squares Problems*. Society For Industrial and Applied Mathematics, 1996.

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

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

[6] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[7] M. Irani. Multi-frame optical flow estimation using subspace constraints. In *Proceedings of the International Conference on Computer Vision*, 1999.

[8] 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.

[9] L. Torresani and C. Bregler. Space-time tracking. In *Proceedings of the European Conf. on Computer Vision*, 2002.

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

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

[12] L. Wolf and A. Shashua. On projection matrices $P^k \rightarrow P^2$, $k = 3, \ldots, 6$, and their applications in computer vision. *International Journal of Computer Vision*, 48(1), June 2002.

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

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