A Factorization Approach to Structure from Motion with Shape Priors*

Alessio Del Bue
Institute for Systems and Robotics – Instituto Superior Técnico
Av. Rovisco Pais 1 – 1049-001 Lisboa – Portugal
http://www.isr.ist.utl.pt/~adb/

Abstract

This paper presents an approach for including 3D prior models into a factorization framework for structure from motion. The proposed method computes a closed-form affine fit which mixes the information from the data and the 3D prior on the shape structure. Moreover, it is general in regards to different classes of objects treated: rigid, articulated and deformable. The inclusion of the shape prior may aid the inference of camera motion and 3D structure components whenever the data is degenerate (i.e. nearly planar motion of the projected shape). A final non-linear optimization stage, which includes the shape priors as a quadratic cost, upgrades the affine fit to metric. Results on real and synthetic image sequences, which present predominant degenerate motion, make clear the improvements over the 3D reconstruction.

1. Introduction

Extracting 3D data from monocular image sequences is a problem extensively studied in Computer Vision. The idea embracing various scenarios is relatively similar: to infer both the 3D structure parametrization and the camera parameters from a set of 2D points extracted from a sequence which depicts a moving object. In this context, methods based on the bilinear factorization of the image data have been very successful in both proposing simple and closed-form approaches with the least assumptions as possible.

Back in the early '90s, Tomasi and Kanade [14] introduced the first factorization algorithm dealing with a rigid object viewed by a simple orthographic camera. Later on, studies were mostly focused on upgrading the approach to more complex viewing conditions such as paraperspective [11] and projective [13]. Only recently, the framework was extended to deal with objects which may also vary their shape. Deformable [5] and articulated [17, 20] factorizations are among these examples.

However, one of the main problems of Structure from Motion (SfM) consists in the higher complexity in which shapes may vary their 3D structure. This may add non-linearities and strong dependencies between both shape and motion components resulting in non-trivial solutions of the problem. Likewise, the higher number of degrees of freedom and the possible motion degeneracies in the measured data may lead to a local solution which correctly minimizes the 2D reprojection error but, however, results in a poor or even meaningless 3D reconstruction.

In order to counter this effect, prior information may be included to obtain reliable 3D reconstructions. Different priors have been shown to improve performances in rigid and non-rigid SfM. Forsyth et al. [7] first proposed priors over the specific camera constraints in a consistent Bayesian framework. Xiao et al. [19] use the prior information over a set of independent shapes to compute a closed-form solution. In a face modelling context, Solem and Kahl [12] used a learned shape model to aid the 3D inference over regions with no 2D information available. Del Bue et al. [6] enforce priors over the rigidity of some points to obtain reliable estimations of the object rigid components. Torresani et al. [15] propose the use of gaussian priors over the deformation parameters in order to avoid arbitrarily variations. Finally, Olsen and Bartoli [10] impose a prior over temporal variations of the camera parameters combined with constraints over the proximity of projected 2D points and reconstructed 3D points.

Differently from the mentioned solutions, priors are here introduced in the form of previously computed 3D shapes. In such way, it is possible to obtain a description of the object shape jointly given by the measured data and the prior information available. This approach especially supports the computation of reliable 3D shapes whenever the image sequence contains strong degeneracies. For instance, a common case is a talking head in front of a camera performing tiny pose changes. In this case depth would be lost if not irremediably mixed with the ongoing deformations. A known 3D metric description of the subject face may how-

[1] *To appear in CVPR08, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, Alaska, June 2008.*

ever recover the lost depth and disambiguate better motion and shape components. At this end, the proposed computational methods directly include the 3D metric priors in a factorization framework for SfM. First, the approach uses the data and the prior to compute an initial affine solution which is then upgraded to metric with an iterative non-linear optimization procedure.

The next section provides an introduction to SfM methods showing how different problems can be treated with a unique approach. Section 3 shows how to compute a metric solution for the motion and structure components The proposed algorithm is then explained in Section 4 showing computational tools to compute 3D models given prior information. The experiments in Section 5 show real and synthetic examples of 3D reconstruction in such cases.

2. Factorization for SfM

The key idea in SfM is to gather all the 2D image coordinates lying on a generic shape at each frame in a single measurement matrix $\mathtt{W}$. The location of a point j in a certain frame i can be defined by a vector $\mathbf{w}_{ij} = (u_{ij} \;\; v_{ij})^T$ where u_{ij} and v_{ij} are the horizontal and vertical image coordinates respectively. A compact matrix representation can be expressed as:

$$\mathtt{W} = \begin{bmatrix} \mathbf{w}_{11} & \dots & \mathbf{w}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{f1} & \dots & \mathbf{w}_{fp} \end{bmatrix} = \begin{bmatrix} \mathtt{W}_1 \\ \vdots \\ \mathtt{W}_f \end{bmatrix} \tag{1}$$

where f is the total number of image frames and p the number of image points. The image trajectories stored in $\mathtt{W}$ can be expressed as a bilinear product as $\mathtt{W} = \mathtt{M}_{2f\times r}\;\mathtt{S}_{r\times p}$. The matrices $\mathtt{M}$ and $\mathtt{S}$ refer to the motion and shape subspaces respectively with dimension r where $r \ll \min\{2f, p\}$. As a result, the rank of $\mathtt{W}$ is constrained to be $rank\{\mathtt{W}\} \leq r$. The matrices $\mathtt{M}$ and $\mathtt{S}$ can be further decomposed in:

$$\mathtt{M} = \begin{bmatrix} \mathtt{M}_1 \\ \vdots \\ \mathtt{M}_f \end{bmatrix} \qquad \mathtt{S} = \begin{bmatrix} \mathbf{S}_1 & \cdots & \mathbf{S}_p \end{bmatrix} \tag{2}$$

where $\mathtt{M}_i$ with $i = 1 \dots f$ is a $2 \times r$ matrix projecting the shape onto the image frame i. The size of $\mathtt{M}_i$ directly depends on the type of camera and motion that appears in the scene. The component $\mathbf{S}_j$ with $j = 1 \dots p$ is a r-vector that defines the 3D parametrization for each point j and its size depends on the shape properties (i.e. rigid or non-rigid).

2.1. Rigid factorization

The projection of a 3D rigid shape by means of an orthographic camera is historically the first factorization problem studied [14]. In this case the camera motion $\mathtt{M}_i$ and the 3D point $\mathbf{S}_j$ can be expressed as:

$$\begin{aligned} \mathtt{M}_i &= \begin{bmatrix} r_{i1} & r_{i2} & r_{i3} & t_{ui} \\ r_{i4} & r_{i5} & r_{i6} & t_{vi} \end{bmatrix} = \begin{bmatrix} \mathtt{R}_i \;\middle|\; \mathbf{t}_i \end{bmatrix} \\ \mathbf{S}_j &= \begin{bmatrix} X_j & Y_j & Z_j & 1 \end{bmatrix}^T = \begin{bmatrix} \mathbf{X}_j \\ 1 \end{bmatrix} \end{aligned} \tag{3}$$

where $\mathtt{R}_i$ contains the first two rows of a rotation matrix (i.e. $\mathtt{R}_i\mathtt{R}_i^T = \mathtt{I}_{2\times 2}$), $\mathbf{S}_j$ is a 4-vector containing the homogeneous metric coordinates of the 3D point $\mathbf{X}_j$, and $\mathbf{t}_i$ is a 2-vector representing a translation into the image plane. Every point belonging to the rigid structure shares the same rotation and translation, thus we can compact 3D points in a single $4 \times p$ matrix $\mathtt{S}$ giving:

$$\mathtt{W}_i = \begin{bmatrix} \mathtt{R}_i \;\middle|\; \mathbf{t}_i \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 & \cdots & \mathbf{S}_p \end{bmatrix} = \mathtt{M}_i\mathtt{S}. \tag{4}$$

Stacking the rows of $\mathtt{W}_i$ for every frame, we obtain the full measurement matrix as:

$$\mathtt{W} = \mathtt{MS} = \begin{bmatrix} \mathtt{R} \;\middle|\; \mathbf{t} \end{bmatrix} \mathtt{S} \tag{5}$$

where $\mathtt{R}$ is the $2f \times 3$ collection of f rotation matrices, $\mathbf{t}$ is a $2f$-vector which contains the translation for every frame. The dimension of $\mathtt{M}$ and $\mathtt{S}$ is fixed to $r = 4$. If the 2D points in $\mathtt{W}$ are registered to the shape centroid (i.e. $\mathtt{W}\mathbf{1}^T = \mathbf{0}$), the maximum dimensionality of each subspace is $r = 3$ and equation (5) can be written as:

$$\mathtt{W} = \mathtt{MS} = \mathtt{R} \begin{bmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_p \end{bmatrix}. \tag{6}$$

2.2. Articulated factorization

If the measurements in $\mathtt{W}$ belong to two independent moving objects, the overall rank sums to eight since it is possible to write motion and shape components as:

$$\begin{aligned} \mathtt{M}_i &= \begin{bmatrix} \mathtt{R}_i^{(1)} \;\middle|\; \mathbf{t}_i^{(1)} \;\middle|\; \mathtt{R}_i^{(2)} \;\middle|\; \mathbf{t}_i^{(2)} \end{bmatrix} \\ \mathtt{S} &= \begin{bmatrix} \mathtt{S}^{(1)} & \mathtt{0} \\ \mathtt{0} & \mathtt{S}^{(2)} \end{bmatrix} \end{aligned} \tag{7}$$

such that:

$$\mathtt{W} = \begin{bmatrix} \mathtt{W}_i^{(1)} \;\middle|\; \mathtt{W}_i^{(2)} \end{bmatrix} = \mathtt{M}_i\mathtt{S} \tag{8}$$

where $\mathtt{W}_i^{(1)}$ and $\mathtt{W}_i^{(2)}$ are the measured data at frame i for the first and second shape respectively. The components of $\mathtt{M}_i$ and $\mathtt{S}$ for each shape are in the form as shown by equation (4) for a single rigid object. However, the rank r may decrease if the moving objects show a dependency such as a common rotational axis.

In articulated SfM [17, 20], this dependency is given by the joints which constrain the degrees of freedom of the moving objects. In the case of a *universal joint* [17] the distance between the center of the shapes is constant (for instance, the head and the torso of a human body) but they

show independent rotation components. At each frame the shapes connected by a joint satisfy:

$$\mathbf{t}^{(1)} + \mathtt{R}^{(1)}\mathbf{d}^{(1)} = \mathbf{t}^{(2)} + \mathtt{R}^{(2)}\mathbf{d}^{(2)} \tag{9}$$

where $\mathbf{t}^{(1)}$ and $\mathbf{t}^{(2)}$ are the 2D image centroid of the two objects, $\mathtt{R}^{(1)}$ and $\mathtt{R}^{(2)}$ the 2×3 orthographic camera matrices and $\mathbf{d}^{(1)}$ and $\mathbf{d}^{(2)}$ the 3D displacement vectors of each shape from the joint. The relation in equation (9) gives the reduced dimensionality in the motion and shape subspaces. Thus, the shape matrix $\mathtt{S}$ can be written as:

$$\mathtt{S} = \begin{bmatrix} \mathtt{S}^{(1)} & \mathbf{d}^{(1)} \\ \mathtt{0} & \mathtt{S}^{(2)} - \mathbf{d}^{(2)} \\ 1 & 1 \end{bmatrix} \tag{10}$$

where $\mathtt{S}$ is a full rank-7 matrix. The motion for a frame i has to be accordingly arranged to satisfy equation (9) as:

$$\mathtt{M}_i = \begin{bmatrix} \mathtt{R}_i^{(1)} & \mathtt{R}_i^{(2)} & \mathbf{t}_i^{(1)} \end{bmatrix}. \tag{11}$$

Further details can be found in [17], alongside a description of additional joint models. Notice that is necessary to solve an assignment problem in order to form $\mathtt{W}$ as in equation (8). At this end, a recent approach able to detect articulated parts from images is presented in [20].

2.3. Deformable factorization

In the case of deformable objects, a single shape varies its 3D structure with respect to a set of deformation modes. The number of modes used to define the shape deformations results in a specific rank-constraint over the image trajectories in $\mathtt{W}$. The representation for the deformations is a simple model where any specific 3D configuration $\mathtt{X}$ is approximated by a linear combination of a set of k basis shapes $\mathtt{B}_d$ which represent the principal modes of deformation:

$$\mathtt{X} = \sum_{d=1}^{k} c_d \mathtt{B}_d \qquad \mathtt{X}, \mathtt{B}_d \in \Re^{3\times p} \quad c_d \in \Re \tag{12}$$

Bregler et al. [5] were the first to propose an extension of factorization algorithms able to deal with the case of deformable shapes assuming an orthographic camera model. In this case, the coordinates of the 2D image points observed at each frame i are related to the coordinates of the 3D points according to the following equation:

$$\mathtt{W}_i = \mathtt{R}_i \left(\sum_{d=1}^{k} c_{id} \mathtt{B}_d \right) + \mathtt{T}_i \tag{13}$$

where c_{id} is the configuration weight for basis d at frame i. When the image coordinates are registered to the object's centroid, equation (13) can be rewritten as:

$$\mathtt{W}_i = \begin{bmatrix} c_{i1}\mathtt{R}_i & \dots & c_{ik}\mathtt{R}_i \end{bmatrix} \begin{bmatrix} \mathtt{B}_1 \\ \vdots \\ \mathtt{B}_k \end{bmatrix} = \mathtt{M}_i\mathtt{S} \tag{14}$$

Again, by stacking the measurements at each frame we obtain the following compact matrix form:

$$\mathtt{W} = \begin{bmatrix} c_{11}\mathtt{R}_1 & \dots & c_{1k}\mathtt{R}_1 \\ \vdots & \ddots & \vdots \\ c_{f1}\mathtt{R}_f & \dots & c_{fk}\mathtt{R}_f \end{bmatrix} \begin{bmatrix} \mathtt{B}_1 \\ \vdots \\ \mathtt{B}_k \end{bmatrix} = \mathtt{M}\mathtt{S} \tag{15}$$

Since $\mathtt{M}$ is a $2f \times 3k$ matrix and $\mathtt{S}$ is a $3k \times p$ matrix, the rank of $\mathtt{W}$ when no noise is present must be $r \leq 3k$.

3. Affine and metric reconstruction

In order to extract the motion and shape components, the classical procedure solves two separate problems:

1. Find an affine fit $\tilde{\mathtt{M}}$ and $\tilde{\mathtt{S}}$ to the measured data $\mathtt{W}$ given the rank r.
2. Enforce the metric structure by computing a corrective transform $\mathtt{Q}_{r\times r}$ such that $\mathtt{W} = \tilde{\mathtt{M}}\mathtt{Q}\ \mathtt{Q}^{-1}\tilde{\mathtt{S}} = \mathtt{M}\,\mathtt{S}$.

The first step can be trivially solved using any rank revealing technique such as SVD. However this solution is merely one means of numerical computations, other approaches may be used as pointed out in [9]. Given the chosen rigid/non-rigid model, the numerical rank it is fixed to r and thus it is possible to approximate the decomposition as:

$$\mathtt{W} \xrightarrow{SVD} \sum_{i=1}^{n} \mathbf{u}_i \sigma_i \mathbf{v}_i^T = \sum_{i=1}^{r} \mathbf{u}_i \sigma_i \mathbf{v}_i^T = \mathtt{U}_r \Sigma_r \mathtt{V}_r^T \tag{16}$$

where $\mathtt{U}_r$ is a $2f \times r$ orthogonal matrix, Σ_r a $r \times r$ diagonal matrix and $\mathtt{V}_r^T$ a $p \times r$ orthogonal matrix. After simple operations the product can be arranged in the affine form $\mathtt{W} = \tilde{\mathtt{M}}\tilde{\mathtt{S}}$. Notice that this initial solution is independent from the problem considered, either the considered shape is rigid or non-rigid.

The second step, the computation of $\mathtt{Q}$, is where the metric properties are imposed. For instance, in the rigid orthographic case, this leads to the computation of the transformation $\mathtt{Q}_{3\times3}$ which renders each row of $\tilde{\mathtt{M}}_i$ orthonormal (i.e. $\tilde{\mathtt{M}}_i \rightarrow \mathtt{R}_i$). In the non-rigid case the computation is more complex since the relationships in $\mathtt{M}$ are strongly non-linear. For instance, in the deformable case, a closed form solution can be found only if a correct set of independent bases is chosen [19]. A wrong set of bases may lead to inaccurate solutions as shown in [4]. On the other hand, non-linear optimization or hybrid approaches have been proposed to solve the problem [4, 15]. In the case of articulated shapes, solutions are available [17] but based on the correct knowledge of the type of joint connecting the bodies.

4. Factorization with shape priors

It is clear from the previous section that the metric upgrade is dependent on the initial affine decomposition. An

inaccurate affine fit of $\tilde{\mathtt{M}}$ and $\tilde{\mathtt{S}}$ may irremediably compromise the following metric upgrade. In this sense, the introduction of a prior on the values of the shape subspace $\mathtt{S}$ can bring the estimation close to the desired solution. Moreover, shape priors can support the computation in case of deficient components in the motion subspace $\mathtt{M}$ – a likely case when the image motion is weak. In such cases, it is always possible to compute a factorization which well fit $\mathtt{W}$ numerically but nonetheless the factors $\tilde{\mathtt{M}}$ and $\tilde{\mathtt{S}}$ may contain meaningless components. This problem is hard to solve unless prior information is included in the computation.

4.1. Measurements and priors

Shape priors are introduced as a matrix $\mathtt{L} \in \Re^{l \times p}$ which holds a parametrization of each image trajectory stored in $\mathtt{W}$. A point trajectory $\mathbf{w}_j$ can be written as a $2f$-vector such that $\mathtt{W} = [\mathbf{w}_1 \cdots \mathbf{w}_p]$. Thus accordingly, we can write the shape prior [1] as $\mathtt{L} = [\mathbf{l}_1 \cdots \mathbf{l}_p]$. The size of each l-vector $\mathbf{l}_j$ depends on the type of prior chosen. Often $\mathtt{L}$ may represent a rigid 3D shape leading to $l = 3$. Such prior may be used in deformable SfM to obtain a more reliable estimation of an object with complex shape variations. Alternatively, it may support the estimation of the object's depth when the shape is moving planarly. In general, $\mathtt{L}$ may as well store more complex shape descriptions (i.e. $l > 3$) such as a set of deformable basis shapes computed previously from a similar shape.

The main idea is to join the information stored in $\mathtt{L}$ with the available measurements in $\mathtt{W}$ in order to extract an affine fit which is dependent on both components. This can be formulated as two bilinear models for the data:

$$\mathtt{W} = \tilde{\mathtt{M}}_{2f\times t}\tilde{\mathtt{S}}_{t\times p} = [\mathtt{M}_J|\,\mathtt{M}_I]\begin{bmatrix}\mathtt{S}_J\\ \mathtt{S}_I\end{bmatrix} \tag{17}$$

and for the shape prior:

$$\mathtt{L} = \mathtt{N}_{l\times l}\mathtt{S}_J \tag{18}$$

where the J subscript refers to the components obtained by the joint space between prior and image measurements while the I refers to the remaining ones. Notice that we always consider $\mathtt{L}$ being full rank thus the following properties hold:

$$rank\,(\mathtt{W}) = r,\ rank\,(\mathtt{L}) = l \text{ and } rank\left(\begin{bmatrix}\mathtt{W}\\ \mathtt{L}\end{bmatrix}\right) = t \tag{19}$$

where $t = \max\{r, l\}$ is the overall rank for both prior and measurements.

[1] Notice that both the prior and measurements are registered to the respective centroids i.e. $\mathtt{W}\mathbf{1}_{p\times 1} = \mathbf{0}_{2f\times 1}$ and $\mathtt{L}\mathbf{1}_{p\times 1} = \mathbf{0}_{l\times 1}$.

4.2. Generalized singular value decomposition

Once the shape prior and the data are defined, we seek a computational solution able to find the joint factorization for $\mathtt{W}$ and $\mathtt{L}$. This can be obtained using a *Generalized Singular Value Decomposition* (GSVD) which can decompose the matrix pair $\{\mathtt{W}, \mathtt{L}\}$ as:

$$\begin{aligned}\mathtt{W} &= \mathtt{U}\,\mathtt{D}_U\,\mathtt{X}^T\\ \mathtt{L} &= \mathtt{V}\,\mathtt{D}_V\,\mathtt{X}^T\end{aligned} \tag{20}$$

where $\mathtt{X}^T$ is a $p \times p$ matrix which span the common row space of $\{\mathtt{W}, \mathtt{L}\}$, $\mathtt{U}$ is a $2f \times 2f$ matrix with orthonormal columns ($\mathtt{U}^T\mathtt{U} = \mathtt{I}$) and $\mathtt{V}$ is a $l \times l$ matrix such that $\mathtt{V}^T\mathtt{V} = \mathtt{I}$. Notice that the matrix $\mathtt{X}$ is rank deficient with t non-zero singular values (for more details and proofs on GSVD see [2], Sec. 4.2.2). The diagonal value matrices $\mathtt{D}_U$ and $\mathtt{D}_V$ of size $2f \times p$ and $l \times p$ are arranged as:

$$\mathtt{D}_U = \begin{bmatrix}\Sigma_U & \mathtt{0}\\ \mathtt{0} & \mathtt{I}\end{bmatrix} \text{ and } \mathtt{D}_V = \begin{bmatrix}\Sigma_V & \mathtt{0}\\ \mathtt{0} & \mathtt{0}\end{bmatrix}. \tag{21}$$

The diagonal matrices $\Sigma_U = diag(\sigma_1, \ldots, \sigma_l)$ and $\Sigma_V = diag(\mu_1, \ldots, \mu_l)$ of size $l \times l$ are constrained such that $\Sigma_U^2 + \Sigma_V^2 = \mathtt{I}$ and they have the diagonal elements ordered as:

$$0 \le \sigma_1 \le \ldots \le \sigma_l \le 1 \text{ and } 1 \ge \mu_1 \ge \ldots \ge \mu_l > 0$$

The ratio between the diagonal values are called *Generalized Singular Values* (GSV) and they are defined as $\gamma_i = \sigma_i/\mu_i$. As a further note, the data and prior matrices are usually pre-scaled such that $\|\mathtt{W}\|^2 = \|\mathtt{L}\|^2$. This condition [8] guarantees a well-conditioning of the matrix $\mathtt{X}$ and it is the only data scaling performed in the decomposition.

4.3. Generalized factorization for SfM

The GSVD decomposes the image measurements in $\mathtt{W}$ with a common row space which is dependent on both the measured data stored in $\mathtt{W}$ and the 3D prior information stored in $\mathtt{L}$. However, to obtain a solution in the form of equation (17), we must reduce the decomposition of $\mathtt{W}$ given by GSVD into the two standard affine components $\tilde{\mathtt{M}}_{2f\times t}$ and $\tilde{\mathtt{S}}_{t\times p}$. Thus, it is convenient to split $\mathtt{X}$ in the components which are dependent on the prior (the first l) and the one dependent on the data (the remaining $p - l$) such that:

$$\mathtt{X} = \left[\,\mathtt{X}_J \;\middle|\; \mathtt{X}_I\,\right]. \tag{22}$$

In such way, we aim to preserve the common row space $\mathtt{X}_J^T$ component which was computed by GSVD.

Thus, equation (20) can be accordingly separated in two components $\mathtt{W}_J$ and $\mathtt{W}_I$ giving:

$$\mathtt{W} = \mathtt{W}_J + \mathtt{W}_I = \mathtt{U}_J\Sigma_U\mathtt{X}_J^T + \mathtt{U}_I\mathtt{X}_I^T. \tag{23}$$

The matrix $\mathtt{X}_J^T$ of size $l \times p$ alone contains the row space components which mixes measurements and priors. The

row space in $\mathtt{X}_I$ of size $(p-l)\times l$ still entails the rank deficiency and it requires a further decomposition to extract the remaining $t-l$ components.

In order to obtain this further affine fit, the proposed method performs two projections of $\mathtt{W}_I$: first over the subspace defined by $\mathtt{X}_J^T$ and then along its orthogonal complement. This is obtained by defining the orthogonal projector $\mathtt{P}$ such that:

$$\mathtt{P} = \mathtt{X}_J \left(\mathtt{X}_J^T \mathtt{X}_J\right)^{-1} \mathtt{X}_J^T \tag{24}$$

giving:

$$\mathtt{W}_I = \mathtt{W}_I\mathtt{P} + \mathtt{W}_I\mathtt{P}_\perp = \mathtt{W}' + \mathtt{W}'' \tag{25}$$

where $\mathtt{P}_\perp = \mathtt{I} - \mathtt{P}$ which has the result to further split the remaining components in $\mathtt{W}'$ which still belongs to the subspace given the prior and $\mathtt{W}''$ which is its orthogonal complement. The components in $\mathtt{W}''$ can be reduced via SVD to obtain the remaining $t-l$ components giving:

$$\mathtt{W}'' \overset{SVD}{\longrightarrow} \mathtt{U}_c\mathtt{D}_c\mathtt{V}_c^T \tag{26}$$

which can be re-arranged as:

$$\mathtt{M}_I = \mathtt{U}_c\mathtt{D}_c \text{ and } \mathtt{S}_I = \mathtt{V}_c^T \tag{27}$$

where $\mathtt{M}_I$ is a $2f\times(t-l)$ matrix and $\mathtt{S}_I$ a $(t-l)\times p$ matrix. The remaining data $\mathtt{W}'$ projected along $\mathtt{X}_J^T$ is merged to the joint space obtaining the measurements $\mathtt{W}_g$ such that:

$$\begin{aligned}\mathtt{W}_g &= \mathtt{W}_J + \mathtt{W}' = \mathtt{U}_J\Sigma_U\mathtt{X}_J^T + \mathtt{U}_I\mathtt{X}_I^T\mathtt{X}_J\left(\mathtt{X}_J^T\mathtt{X}_J\right)^{-1}\mathtt{X}_J^T \\ &= \left(\mathtt{U}_J\Sigma_U + \mathtt{U}_I\mathtt{X}_I^T\mathtt{X}_J\left(\mathtt{X}_J^T\mathtt{X}_J\right)^{-1}\right)\mathtt{X}_J^T = \mathtt{M}_J\mathtt{S}_J\end{aligned}$$

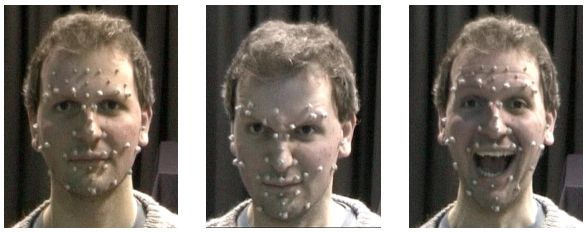
where $\mathtt{S}_J = \mathtt{X}_J^T$ and $\mathtt{M}_J$ is given by the remaining factors. Given the factors $\mathtt{M}_J$, $\mathtt{M}_I$, $\mathtt{S}_J$ and $\mathtt{S}_I$, it is possible now to form the bilinear decomposition as in equation (17).

4.4. Finding a metric solution

For the case of rigid shapes, the affine fit given by the priors can be finally upgraded to metric by forcing metric constraints in closed form for different type of cameras [9]. Differently for non-rigid shapes, we opt for a non-linear optimization stage based on bundle adjustment [18] where a prior on the rigid basis shape is included as an additional quadratic cost (i.e. $\mathtt{L}$ is a $3\times p$ matrix). In principle, the prior may correspond to a full parametrization of a deformable shape, i.e. $l > 3$, however here we focus on priors which are describing the rigid component of a deformable object. Thus, the cost function minimized reflects the deformable model presented in Section 2.3 giving:

$$\min_{\mathtt{R}_i\mathbf{B}_{dj}c_{id}} \sum_{i,j} \| \mathbf{w}_{ij} - (\mathtt{R}_i \sum_d c_{id}\mathbf{B}_{dj}) \|^2 + \sum_j \| \mathbf{B}_{1j} - \mathtt{Cl}_j \|^2$$

where $\mathbf{B}_{dj}$ is the 3×1 basis component for the point j such that $\mathtt{B}_d = [\mathbf{B}_{d1}\cdots\mathbf{B}_{dp}]$. The matrix $\mathtt{C}$ performs a metric alignment of $\mathtt{L}$ to the first basis shape $\mathtt{B}_1$ and it can be computed using standard Procrustes analysis. The minimization of the first sum of quadratic costs is equivalent to a Maximum Likelihood (ML) estimate of the model parameters if i.i.d gaussian noise is affecting the measurements. However, by including the second sum, we obtain a Maximum A Posteriori (MAP) estimate given the shape prior.



Neutral Anger Surprise

Figure 1. Three images sampled from a 160 frames sequence showing different facial expressions.

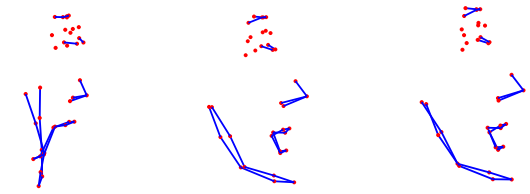
In order to initialize the non-linear optimization, the affine fit given by equation (17) is used to compute an initial metric solution for the prior constrained components $\mathtt{M}_J$ and $\mathtt{S}_J$. This procedure has analogies with the approach first proposed in [16] where the rigid component was used to initialize a non-linear optimization procedure. The main difference here is that the rigid shape is given by a mixture of prior and measured data. To compute each rotation matrix $\mathtt{R}_i$ and the first configuration weight c_{i1} each frame-wise component of $\mathtt{M}_J$ can be decomposed using a orthonormal decomposition [3]. The remaining values c_{id} with $d = 2\ldots k$ are initialized close to zero. Finally notice that the non-linear optimization with $k = 1$ can be used to infer the 3D structure of a rigid object. For the articulated case, the algorithm presented in [17] can be extended to compute an affine fit using shape priors which can be then corrected by forcing specific metric constraints for the given joint. Non-linear optimization can be then applied by adding additional priors for each articulated part.

5. Experiments

The experiments are mainly focused on deformable and articulated shapes. First, synthetic tests are performed on a deformable face in order to verify the validity of the reconstruction using shape priors. The results from non-linear optimization are compared against ground truth obtained from a VICON motion capture system. Then, two further real tests show the method performances with real imaging condition for the deformable and articulated case.

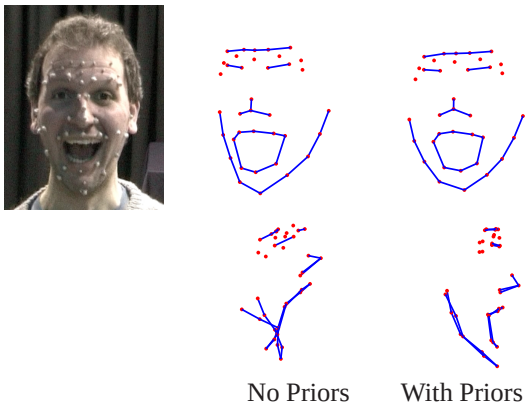
5.1. Deformable face with ground truth

In this experiment 37 point tracks from a 160 frames long sequence were obtained with a VICON system which cap-



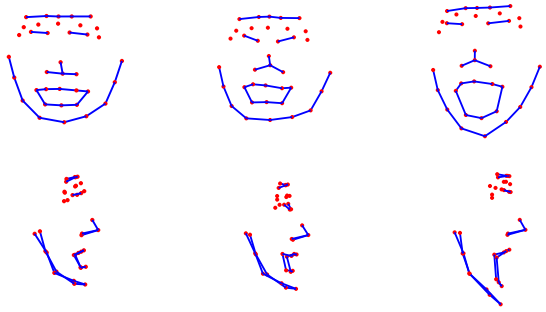
(a) (b) (c)

Figure 2. Shape prior and its effect over the estimated basis shapes. (a) The 1^{st} deformable basis computed without priors. (b) The 3D prior used in the test. (c) The 1^{st} basis computed using priors.



No Priors With Priors

Figure 3. Comparison between the solution with and without priors. Even if the frontal view appears correct in both cases, only the solution with priors can properly estimate the shape depth.



Neutral Anger Surprise

Figure 4. Front and side views of the 3D Reconstructions after non-linear optimization with shape priors of the neutral, anger and surprise facial expressions.

tured 2D and 3D locations of a set of markers overlaid on a deforming face (see Figure 1). The subject was performing tiny head pose changes which in turn affected the extraction of satisfactory 3D reconstructions from 2D trajectories. Figure 2(a) shows the first basis shape $\mathtt{B}_1$ obtained using deformable non-linear optimization without shape priors. The basis shape was generally quite flat and the mouth bend inward into the head. The shape prior $\mathtt{L}_{3\times 37}$ is taken from a 3D reconstruction of the VICON system itself when the subject was performing a neutral pose as shown in Figure 2(b). After performing non-linear optimization with priors, the basis shape $\mathtt{B}_1$ much resemble the prior with some variations located over the temple area (Figure 2(c)).

Figure 3 shows a comparison between 3D reconstructions for the surprise expression. Notice that both the frontal views of the 3D reconstructions apparently estimated correctly the face shape. However, the inclusion of the prior is critical as shown in the side views. Given a tiny rigid motion, the ML solution alone is very ambiguous since strong variations in depth results in small displacements onto the image plane. The inclusion of a prior over the rigid shape component constrains the object depth and deformation estimates. Figure 4 presents front and side views of the final 3D reconstruction after 21 iterations of non-linear optimization. Facial symmetry is well preserved and generally the depth of the shape is correctly estimated.

In order to compute quantitatively the algorithm performances, we performed 100 trials for each test with different conditions. Gaussian noise of different levels was added to the image points while the prior accurateness was as well altered by adding gaussian noise to L. These tests were performed in order to show how much the algorithm is resilient to inaccurate priors and increasing image noise. Results are presented in Figure 5 showing that the algorithm can deliver satisfactory performances even with some degrees of inaccuracy on the shape prior.

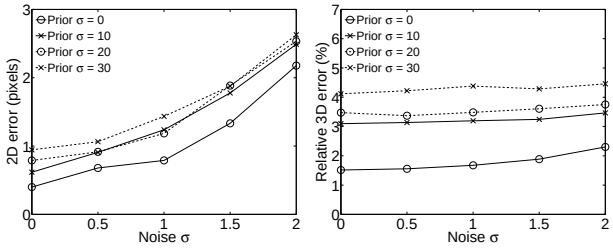


Figure 5. Synthetic experiments results for both 2D reprojection error in pixel (left) and 3D error in units (right). The image shape is approximately of size 114×84 pixels and zero-mean gaussian noise of variance $\sigma = \{0, 0.5, 1, 1.5, 2\}$ pixels is added to W. The 3D prior L is contained in a box of size $169 \times 193 \times 102$ and zero-mean gaussian noise of variance $\sigma = \{0, 10, 20, 30\}$ units is added to simulate inaccurate shape priors.

5.2. Image data with motion degeneracy

The aim of this experiment is to enforce a shape prior belonging to a subject with measurements obtained from a different subject. A rigid 3D shape of a face is first extracted from the image sequence shown in Figure 6 using a rigid factorization approach. Then, the prior is used to infer the 3D structure of a 45 frames sequence with deformations mainly localized in the mouth region. In both sequences, the 65 image points were extracted using an AAM tracker.

Figure 6. The first two images show snapshots from a brief sequence of 75 frames showing dominant rigid motion. The right image shows the 3D rigid shape prior computed from the sequence.

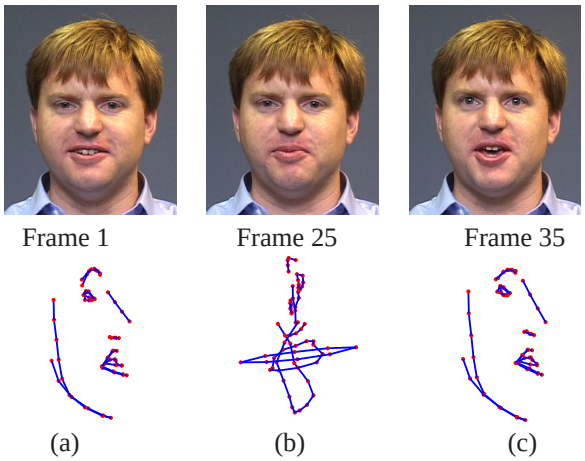


Figure 7. The first row shows three frames of a sequence with nearly planar motion and deformations mainly located in the mouth. The second row presents the 3D shape comparisons between the prior (a), the 1^{st} basis extracted without prior (b) and the same basis computed using shape priors (c).

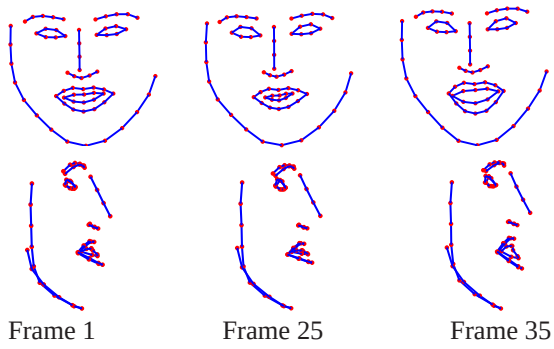


Figure 8. Front and side views after non-linear optimization. Deformations are mainly localized in the mouth region.

The second row of Figure 7 shows the priors and the difference between the computed rigid basis shape $\mathtt{B}_1$. From Figure 7(b) it is evident that the solution without priors presents depth estimates which were rather compromised. Differently, after the shape prior inclusion, depth is computed correctly and the face characteristics are adapted to the new subject. It is possible to notice this in Figure 7(c) where the nose is more elongated and eyebrows are less bent. Finally, Figure 8 shows the 3D reconstruction for three frames after non-linear minimization.

Figure 9. The left image shows a frame from a sequence with two articulated bodies with a universal joint. The right image shows the overall 3D reconstruction which has wrong depth estimates.

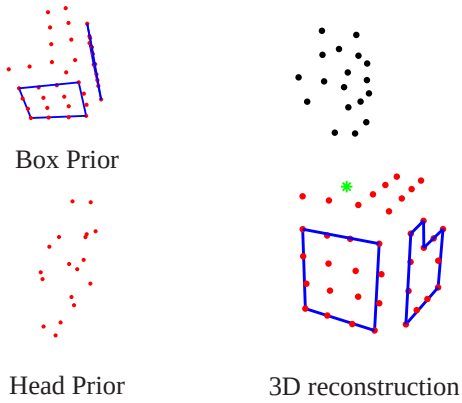


Figure 10. The two images on the left show the priors used to infer the 3D articulated structure. On the right it is shown the full reconstruction along with the estimation of the joint position (green).

5.3. Articulated shape with priors

This test was aimed to show the relevance of the inference with priors also in the case of articulated SfM. The experiment dealt with a universal joint between two shapes as shown by Figure 9. The 61 frames image sequence however contained motion degeneracies only in the box shape while the head was rotating enough to assure reliable 3D reconstructions. The shape priors were obtained from separated rigid factorizations of both objects from different sequences. Generalized factorization is then applied to obtain a better initial fit for the box shape. The overall affine structure is finally upgraded to metric with the closed-form

solution proposed in [17]. Figure 10 shows the box sides now preserving orthogonality and the depth was accurately estimated along with the position of the universal joint.

6. Conclusion

This paper presented a method capable of including shape priors in a factorization framework for SfM. These priors were in the form of previously computed 3D shapes which represented a close representation of the measured data. In such way it was possible to obtain reliable 3D reconstructions especially when the motion appearing in the image sequence was degenerate. At this end, a closed-form solution is used to compute an affine fit of motion and shape components which are then upgraded to metric using non-linear optimization with shape priors. Notice the strong relation of this approach to other methods based on factorization. For instance, in the case of photometric stereo [1], priors on the normals of the object may be directly applied using slight variations of our method. As ongoing work, in order to widen the approach applicability, it is necessary to develop an efficient algorithm for automatically matching 2D point trajectories and 3D shape priors.

Acknowledgments

This work was supported by Fundação para a Ciência e a Tecnologia (ISR/IST pluriannual funding) through the POS_Conhecimento Program (include FEDER funds) and grant PTDC/EEA-ACR/72201/2006, “MODI - 3D Models from 2D Images”. E. Muñoz, J. Xiao and P. Tresadern provided the sequences used in the experimental sections for the synthetic, deformable and articulated test respectively. Thanks to L. Agapito and X. Lladó for suggestions and for carefully reading this paper.

References

[1] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, 72(3):239–257, 2007. 8

[2] Å. Björck. *Numerical methods for least squares problems*. SIAM Philadelphia, 1996. 4

[3] M. Brand. Morphable models from video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, volume 2, pages 456–463, December 2001. 5

[4] M. Brand. A direct method for 3d factorization of nonrigid motion observed in 2d. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California*, pages 122–128, 2005. 3

[5] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, pages 690–696, June 2000. 1, 3

[6] A. Del Bue, X. Lladó, and L. Agapito. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, New York, NY*, pages 1191–1198, New York, June 2006. 1

[7] D. Forsyth, S. Ioffe, and J. Haddon. Bayesian structure from motion. *Proc. 7th International Conference on Computer Vision, Kerkyra, Greece*, 01:660, 1999. 1

[8] P. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. Society for Industrial Mathematics, 1998. 4

[9] K. Kanatani and Y. Sugaya. Factorization without factorization: complete recipe. *Memories of the Faculty of Engineering, Okayama University*, 38(1–2):61–72, 2004. 3, 5

[10] S. Olsen and A. Bartoli. Using priors for improving generalization in non-rigid structure-from-motion. *Proc. British Machine Vision Conference*, 2007. 1

[11] C. J. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. In *Proc. 3rd European Conference on Computer Vision, Stockholm*, volume 2, pages 97–108, 1994. 1

[12] J. Solem and F. Kahl. Surface reconstruction using learned shape models. *Advances in Neural Information Processing Systems*, 17, 2005. 1

[13] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *Proc. 4th European Conference on Computer Vision, Cambridge*, pages 709–720, April 1996. 1

[14] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *International Journal of Computer Vision*, 9(2):137–154, 1992. 1, 2

[15] L. Torresani, A. Hertzmann, and C. Bregler. Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008. 1, 3

[16] L. Torresani, D. Yang, E. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, 2001. 5

[17] P. Tresadern and I. Reid. Articulated structure from motion by factorization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California*, volume 2, pages 1110–1115, June 2005. 1, 2, 3, 5, 8

[18] B. Triggs, P. McLauchlan, R. I. Hartley, and A. Fitzgibbon. Bundle adjustment – A modern synthesis. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, LNCS, pages 298–375. Springer Verlag, 2000. 5

[19] J. Xiao, J. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. *International Journal of Computer Vision*, 67(2):233–246, April 2006. 1, 3

[20] J. Yan and M. Pollefeys. A factorization-based approach for articulated non-rigid shape, motion and kinematic chain recovery from video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5), May 2008. 1, 2, 3