Multibody Factorization with Uncertainty and Missing Data Using the EM Algorithm

Amit Gruber and Yair Weiss
School of Computer Science and Engineering
The Hebrew University of Jerusalem
Jerusalem, Israel 91904
{*amitg,yweiss*}*@cs.huji.ac.il*

Abstract

Multibody factorization algorithms [2, 1, 16] give an elegant and simple solution to the problem of structure from motion even for scenes containing multiple independent motions. Despite this elegance, it is still quite difficult to apply these algorithms to arbitrary scenes. First, their performance deteriorates rapidly with increasing noise. Second, they cannot be applied unless all the points can be tracked in all the frames (as will rarely happen in real scenes). Third, they cannot incorporate prior knowledge on the structure or the motion of the objects.

In this paper we present a multibody factorization algorithm that can handle arbitrary noise covariance for each feature as well as missing data. We show how to formulate the problem as one of factor analysis and derive an expectation-maximization based maximum-likelihood algorithm. One of the advantages of our formulation is that we can easily incorporate prior knowledge, including the assumption of temporal coherence. We show that this assumption greatly enhances the robustness of our algorithm and present results on challenging sequences.

1. Introduction

Common motion or "common fate" provides a powerful cue for segmenting objects. This principle, simply stated is that points that move together should be grouped together but there are number of ways of formulating "moving together" (see [16] for a recent review). In this paper, we focus on the problem of grouping based on common rigid 3 dimensional motion. Figure 1 shows Ullman's classical demonstration of how powerful this cue can be [12]. Two concentric cylinders rotate with different angular velocities and are rendered using an orthographic projection. Note that points on different cylinders may move in the same direction while points on the same cylinder may move in opposite directions. Nevertheless, humans obtain a powerful percept of the two independently moving objects.

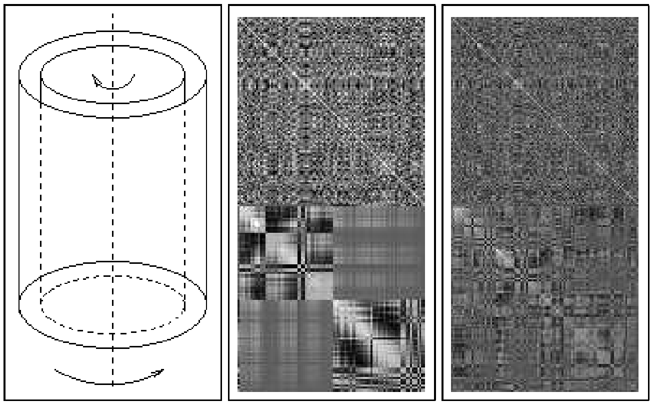

Figure 1: **a.** Ullman's [12] co-axial transparent cylinders demonstration. **b.** Costeira-Kanade [1] factorized matrix for noise free input. Top row is unsorted matrix, bottom is sorted. **c.** Costeira-Kanade factorized matrix for noisy input (unsorted on top row, sorted on bottom).

Multibody factorization algorithms [1, 2, 16] present an elegant and simple solution to this task. As we explain below, these methods factorize the measurement matrix so that in the absence of noise elements corresponding to different segments will be zero. Figure 1 shows the Costeira and Kanade factorized matrix in the noiseless case. Indeed the off-diagonal blocks are close to zero (and not zero due to input degeneracy) and the algorithm correctly segments the display. But once we add even small amounts of noise, (figure 1c) the off block elements are no longer separable from the others and the correct segmentation is far from obvious. Similar sensitivity to noise is shown by the algorithms in [2, 16].

An additional problem with existing factorization algorithms is that they require a full observation matrix: i.e. they work with points that are visible in all frames. In real scenes, where there are occlusions and failures of tracking, this is a severely limiting assumption. Finally, existing factorization methods have no way of incorporating prior knowledge on the motions or the structure. Due to the difficulty of the segmentation problem, one would like a way of incorporating these priors. For example, in almost any video sequence the camera location at time t is dependent on its location at time $t+1$ so that randomly permuting the order of frames would give a very different sequence. Yet existing factorization algorithms are invariant to such a factorization and thus neglect an important source of information.

In this paper we present a multibody factorization algorithm that can handle arbitrary noise covariance for each feature as well as missing data and can easily incorporate prior knowledge. The algorithm is an expectation-maximization based maximum-likelihood algorithm. We present results on challenging real and synthetic sequences.

1.1. Problem Formulation

A set of P feature points in F images are tracked along an image sequence. Let (u_{fp}, v_{fp}) denote image coordinates of feature point p in frame f.

Let $U = (u_{fp})$, $V = (v_{fp})$ and $W = (w_{ij})$ where $w_{2i-1,j} = u_{ij}$ and $w_{2i,j} = v_{ij}$ for $1 \leq i \leq F$, i.e. W is an interleaving of the rows of U and V.

Let K be the number of different motion components in the sequence. Let $\{G_k\}_{k=1}^{K}$ be a partition of the tracked feature points into K disjoint sets, each consists of all the points that conform to the kth motion, and let P_k be the number of feature points in G_k ($\sum P_k = P$).

Let M_i^j be a 2×4 matrix describing the jth camera parameters at time i, and let S_j be a $4 \times P_j$ matrix describing the $3D$ homogeneous coordinates of the P_j points in G_j moving according to the jth motion component.

Let

$$\left[M_i^j\right]_{2\times 4} = \begin{bmatrix} m_i^{jT} & d_i^j \\ n_i^{jT} & e_i^j \end{bmatrix} \quad \text{and} \quad S_j = \begin{bmatrix} X_{j1} & \cdots & X_{jP_j} \\ Y_{j1} & \cdots & Y_{jP_j} \\ Z_{j1} & \cdots & Z_{jP_j} \\ 1 & \cdots & 1 \end{bmatrix}_{4\times P_j} \tag{1}$$

m_i^j and n_i^j are 3×1 vectors that describe the rotation of the jth camera; d_i^j and e_i^j are scalars describing camera translation [1], and S_j describes points location in $3D$.

Let $\tilde{W}$ be a matrix of observations ordered according to the grouping $\{G_k\}_{k=1}^{K}$, i.e. the first P_1 columns of $\tilde{W}$ correspond to the points in G_1 and so on.

Under affine projection, and in the absence of noise, Costeira and Kanade [1] formulated this problem in the form:

$$\left[\tilde{W}\right]_{2F\times P} = [M]_{2F\times 4} \left[\tilde{S}\right]_{4\times P} \tag{2}$$

where

$$[M]_{2F\times 4K} = \begin{bmatrix} M_1^1 & M_2^1 & \cdots & M_K^1 \\ \vdots & & & \\ M_1^F & M_2^F & \cdots & M_K^F \end{bmatrix}_{2F\times 4K} \tag{3}$$

and

$$\left[\tilde{S}\right]_{4K\times P} = \begin{bmatrix} S_1 & 0 & \cdots & 0 \\ 0 & S_2 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & S_K \end{bmatrix}_{4K\times P} \tag{4}$$

If the segmentation $\{G_k\}_{k=1}^{K}$ was known, then we could have separated the point tracks (columns of the observations matrix) into K disjoint submatrices according to $\{G_k\}$, and run a single structure from motion algorithm (for example [11]) on each submatrix.

In real sequences, where segmentation is unknown, the observation matrix, W, is a column permutation of the ordered matrix $\tilde{W}$:

$$W = \tilde{W}\Pi = MS \tag{5}$$

where S is a $4K \times P$ matrix describing scene structure (with unordered columns). Substituting in equation 2 shows that the same $P \times P$ column permutation matrix Π that right multiplies the measurements matrix, multiplies also the block diagonal structure matrix, $\tilde{S}$. Hence, the structure matrix S is in general not block diagonal, but rather a column permutation of a block diagonal matrix.

$$S = \tilde{S}\Pi \tag{6}$$

The motion matrix, M, remains unchanged.

1.2. Previous Work

Costeira and Kanade [1] suggested to search for the block structure in S by factorizing W using the SVD:

$$W = U\Sigma V^T \tag{7}$$

Forming a matrix V' whose columns are the first $4K$ columns of V and calculating the $P\times P$ matrix $Q = V'V'^T$. It can be shown that in the absence of noise, $Q(i,j) = 0$ for points belonging to different segments.

In noisy situations the inter block elements $Q(i,j)$ are not zero, and in general they cannot be separated from the intra block elements by thresholding. Sorting the matrix Q to find block structure is equivalent to minimizing the energy of the off-diagonal blocks which is an NP-complete problem. Instead, Costeira and Kanade, suggested a greedy suboptimal clustering heuristic which turns out to be very

[1] We do not subtract the mean of each row from it, since in case of missing data the centroid of points visible in a certain frame does not coincide with the centroid of all points.

sensitive to noise. Results illustrating this sensitivity appear in section 3. In addition, as an initial step, the rank of the noise free measurements matrix should be found from the noisy measurements matrix. This is a difficult problem which is discussed extensively in [2].

Gear [2] suggested a similar method that use the reduced row echelon form of the measurements matrix W. For reasons of numerical stability, QR decomposition is first applied to W and then the reduced row echelon form of R ($W = QR$), which is identical to the reduced row echelon form of W, is found. Let A be the reduced row-echelon form of R. If measurements are noise free, then if both $A_{ij} \neq 0$ and $A_{ik} \neq 0$ for some i, then points j and k are in the same set, and a simple connected component algorithm can find the segmentation. For noisy scenarios, [2] suggests an EM algorithm for clustering. Again, in noisy situations the algorithm does not guarantee correct segmentation. It is crucial for Gear's algorithm that the motion matrix (that contains all motion matrices for all times) is of full rank. Otherwise, there may be non-zero elements on the same row for two points (columns) belonging to different objects.

Zelnik et al [16] incorporates directional uncertainty by using Gear's method on the matrix $[\,G \quad H\,]$ where G and H are $F \times P$ matrices containing measurable image quantities.

Kanatani [7] proposed an algorithm that takes advantage of the *affine subspace constraint*. However, he shows that in spite of the model improvement, when comparing to a previous algorithm, each algorithm performs better on different cases.

Several authors have addressed the related, but different problem of 3D rigid body segmentation based on two frames and a perspective camera [14, 15, 13]. While these methods show encouraging results, they lack the attractive property of factorization methods in which information from all the frames is used to perform the segmentation. For example, for Ullman's two cylinder demonstration which is rendered using orthographic projection it is easy to show that a single fundamental matrix perfectly explains the data and hence such scenes cannot be segmented using two-frame algorithms.

Despite the elegance of factorization methods many problems remain:

- Once there is noise in the measurements, it is nontrivial to compute a segmentation from the factorization.
- All these methods assume that a full measurement matrix exists and there is no way to use data from frames where image locations of some of the points are unknown.
- these methods have no way of incorporating additional prior knowledge.

2. EM Algorithm for Multibody Factorization

2.1. Segmentation as Constrained Factorization

The permutation Π does not preserve the block diagonal structure of $\tilde{S}$, but it does preserve the property that in a column of the structure matrix corresponding to a point in G_k, only the four entries $4(k-1)+\,1,\ldots,4k$ can be non-zeros (entry $4k$ is always one).

Hence we look for a factorization of W to M and S where the non-zero entries of S for each column can appear only in one out of K specific locations. If we find such a factorization, we can group together columns (points) of S to form blocks, and hence segment the points. Unfortunately, a standard factorization algorithm such as SVD is not guaranteed to find a factorization that satisfies this extra constraint even if it exists. We now show how to modify the classical EM for factor analysis algorithm [10] to produce factorizations that satisfy this constraint.

2.2. Factorization as Factor Analysis

For noisy observations, the model is:

$$[W]_{2F\times P} = [M]_{2F\times 4K}\,[S]_{4K\times P} + [\eta]_{2F\times P} \qquad (8)$$

where M and S are as defined in the previous section and η is Gaussian noise.

We seek a factorization of W to M and S under the constraint that S is a permuted block diagonal matrix $\tilde{S}$, that minimizes the weighted squared error $\sum_t\ [(W_t - M_t S)^T \Psi_t^{-1}(W_t - M_t S)]$, where Ψ_t^{-1} is the inverse covariance matrix of the feature points in frame t.

In [4], the problem of structure from motion for a single motion was written as a factor analysis problem, and solved while placing a prior on the motion. We adapt this approach to the multi motion case and show how to group the feature points at the same time.

In standard factor analysis we have a set of observations $\{y(t)\}$ that are linear combinations of a latent variable $x(t)$:

$$y(t) = Ax(t) + \eta(t) \qquad (9)$$

with $x(t) \sim N(0, \sigma_x^2 I)$ and $\eta(t) \sim N(0, \Psi_t)$. If Ψ_t is a diagonal matrix with constant elements $\Psi_t = \sigma^2 I$ then in the limit $\sigma/\sigma_x \to 0$ the ML estimate for A will give the same answer as the SVD [9]. We now show how to rewrite the multibody factorization problem in this form.

In equation 2 the horizontal and vertical coordinates of the same point appear in different rows. To get an equation with all measurement taken from the same frame in the same line of the measurements matrix, It can be rewritten as:

$$[U\ V]_{F\times 2P} = [M_U\ M_V]_{F\times 8K} \begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix}_{8K\times 2P} + [\eta]_{F\times 2P} \quad (10)$$

where M_U is the submatrix of M consisting of rows corresponding to U (odd rows), and M_V is the submatrix of M consisting of rows corresponding to V (even rows).

Let $A = \begin{bmatrix} S^T & 0 \\ 0 & S^T \end{bmatrix}$. Identifying $y(t)$ with the tth row of the matrix $[U\ V]$ and $x(t)$ with the tth row of $[M_U\ M_V]$, then equation 10 is equivalent (transposed) to equation 9.

The EM algorithm for factor analysis is a standard algorithm for finding the ML estimate for the matrix A. It consists of two steps, (1) the expectation (or E) step in which expectations are calculated over the latent variables $x(t)$ and (2) a maximization (or M) step in which these expectations are used to maximize the likelihood of the matrix A.

For diagonal covariance matrices Ψ_t the standard algorithm gives:

E step:

$$E(x(t)|y(t)) = \left(\sigma_x^{-2}I + A^T\Psi_t^{-1}A\right)^{-1} A^T\Psi_t^{-1}y(t) \quad (11)$$

$$V(x(t)|y(t)) = \left(\sigma_x^{-2}I + A^T\Psi_t^{-1}A\right)^{-1} \quad (12)$$

$$< x(t) > = E(x(t)|y(t)) \quad (13)$$

$$< x(t)x(t)^T > = V(x(t)|y(t)) + <x(t)><x(t)>^T \quad (14)$$

M step:

$$A = (\sum_t y(t) < x(t)^T >)(\sum_t < x(t)x(t)^T >)^{-1} \quad (15)$$

In our setting, the matrix A must satisfy several constraints. First, it must be of the form $\begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix}$. Second, every column of S must have no more than 4 nonzero entries and finally, the last nonzero entry in each column should be equal to 1. We now show how to modify the M step so that it performs constrained factorization.

Denote by s_p^k a vector of length 3 that denotes the $3D$ coordinates of point p belonging to motion model k. then for a diagonal [2] noise covariance matrix Ψ_t the M step is:

$$s_p^k = B_{pk}C_{pk}^{-1} \quad (16)$$

where

$$\begin{aligned} B_{pk} &= \sum_t \big[\Psi_t^{-1}(p,p)(u_{tp} - < d_t^k >) < m_k(t)^T > \quad (17) \\ &\quad + \Psi_t^{-1}(p+P,p+P)(v_{tp} - < e_t^k >) < n_k(t) >^T\big] \\ C_{pk} &= \sum_t \big[\Psi_t^{-1}(p,p) < m_k(t)m_k(t)^T > \\ &\quad + \Psi_t^{-1}(p+P,p+P) < n_k(t)n_k(t)^T >\big] \end{aligned}$$

where the expectations required in the M step are the appropriate subvectors and submatrices of $< x(t) >$ and $< x(t)x(t)^T >$ (recall equation 1 and the definition of $x(t)$).

The task now is to find a grouping and $3D$ structure coordinates of the tracked points that maximize the complete log likelihood. In other words, we are looking for a factorization of the matrix $[U \quad V]$ to a $F \times 8K$ motion matrix, M, and a $8K \times 2P$ structure matrix, S, where S is subject to the constraint that in each of its columns there is only one fourth of non-zeros.

The M-step in the multi-motion case is therefore:

$$k' = \arg\max_k likelihood(k, s_p^k) \quad (18)$$

$$s_p = [0 \quad \ldots \quad 0 \quad s_p^{k'} \quad 0 \ldots \quad 0]^T$$

where s_p^k is the structure (computed from equation 16 or from one of the equations 19, 25 ahead) under the assumption that point p moves according to motion k.

By modifying the EM algorithm to deal with constrained factorization we now have an algorithm with the following attractive properties:

- It is guaranteed to find a factorization where the structure matrix has at most 4 nonzero elements per column, even in the presence of noise.
- It can be applied even when there is missing data (these are just points for which $\Psi_t^{-1}(i,i) = 0$).
- It is guaranteed to find a local maximum of the likelihood of S.

Regardless of uncertainty and missing data the complexity of the EM algorithm grows linearly with the number of feature points and the number of frames. At every iteration, the most computationally intensive step is an inversion of an 8×8 matrix.

In addition to these properties, the EM algorithm has the advantage of allowing us to incorporate additional priors.

2.3. Directional Uncertainty and Missing Data

A more realistic noise model for real images is that Ψ_t is *not diagonal* but rather that the noise in the horizontal and vertical coordinates of the same point are correlated with an arbitrary 2×2 inverse covariance matrix (it can be shown that the posterior inverse covariance matrix is $\begin{bmatrix} \sum I_x^2 & \sum I_xI_y \\ \sum I_xI_y & \sum I_y^2 \end{bmatrix}$). This problem is usually called *factorization with uncertainty* [5, 8]. It is easy to derive the M step in this case as well. It is similar to equation 16 except that cross terms A'_p, B'_p involving $\Psi_t^{-1}(p, p+P)$ (that equals $\Psi_t^{-1}(p+P, p)$) are also involved:

[2] We refer to the case where Ψ_t is not diagonal in the next subsection.

$$s_p^k = (B_{pk} + B'_{pk})(C_{pk} + C'_{pk})^{-1} \qquad (19)$$

where

$$\begin{aligned} B'_{pk} &= \sum_t \left[\Psi_t^{-1}(p, p+P)(v_{tp} - < e_t >) < m(t)^T > \right. \qquad (20) \\ &\quad + \Psi_t^{-1}(p+P, p)(u_{tp} - < d_t >) < n(t) >^T \left. \right] \\ C'_{pk} &= \sum_t \left[\Psi_t^{-1}(p, p+P) < n(t) m(t)^T > \right. \\ &\quad + \Psi_t^{-1}(p+P, p) < m(t) n(t)^T > \left. \right] \end{aligned}$$

With the addition of directional uncertainty, points residing on lines (with aperture effect), for example, can now provide reliable information in a certain direction.

The presented algorithm allows the use of any arbitrary inverse covariance matrix, Ψ_t^{-1}, for any point p at any time t. There is no requirement for any relation between these matrices, as opposed to [5, 16], a property which is important for handling missing data.

2.4. Adding Temporal Coherence

We follow [4] and use temporal coherence prior on each of the motion components:

The factor analysis algorithm for factorization assumes that the latent variables $x(t)$ are independent. In SFM this assumption means that the camera location in different frames is indepenent and hence permuting the order of the frames makes no difference for the factorization. As mentioned in the introduction, in almost any video sequence this assumption is wrong. Typically camera location varies smoothly as a function of time.

Figure 2a shows the graphical model corresponding to most factorization algorithms: the independence of the camera location is represented by the fact that every time step is isolated from the other time steps in the graph. But it is easy to fix this assumption by adding edges between the latent variables as shown in figure 2b.

Specifically, we use a second order approximation to the motion of the camera:

$$\begin{aligned} x(t) &= x(t-1) + v(t-1) + \frac{1}{2}a(t-1) + \epsilon_1 && (21) \\ v(t) &= v(t-1) + a(t-1) + \epsilon_2 && (22) \\ a(t) &= a(t-1) + \epsilon_3 && (23) \\ y(t) &= Ax(t) + \eta(t) && (24) \end{aligned}$$

Note that we are not assuming that the $2D$ trajectory of each point is smooth. Rather we assuming the $3D$ trajectory of the camera is smooth.

It is straightforward to derive the EM iterations for a ML estimate of S using the model in equation 24. The M step is unchanged from the classical factor analysis and is given by equation 16. The only change in the E step is that $E(x(t)|y)$ and $V(x(t)|y)$ need to be calculated using a Kalman smoother. We use a standard RTS smoother [3]. Note that the computation of the E step is still linear in the number of frames and data points.

2.5. Adding Prior on Structure

Up to this point, we have assumed nothing regarding the $3D$ coordinates of the feature points we are trying to reconstruct. $3D$ Reconstructions with the true coordinates are considered (a priori) as likely as any other reconstruction, even one that suggest the object is located at an infinite position, or behind the camera, for example. But usually when sequences are acquired for structure reconstruction, the object is located just in front of the camera in the center of the scene, and not at infinity [3]. Therefore, we would like to prefer suggestions for reconstructions that place the feature points around certain coordinates in the world, denoted by S_0 (typically X and Y are scattered around zero and Z is finite). We model this preference with the following prior: $p(S) \propto e^{-\lambda \|S - S_0\|_F^2}$.

Derivation of the modified M-step with the addition of prior on structure yields (the following modification of equation 19):

$$s_p^k = (B_{pk} + B'_{pk})(C_{pk} + C'_{pk} + \lambda(I - S_0))^{-1} \qquad (25)$$

Experimental results show an improvement in reconstruction results in noisy scenes after the addition of this naive prior.

3. Experiments

EM guarantees convergence to a local maximum which is dependent in the initialization. To find the global maximum, we start with several (random) initializations for each input, and choose the output that achieves maximal likelihood to be the final result. Empirical results show that for noise free scenes, the global maximum is usually found with a single initialization. As the amount of noise increases, number of initialization needed for success also increases. We work with maximal number of 10 initializations for an input sequence in the experiments reported in figure 4 and maximal number of 20 initializations in the experiments reported in figure 5.

Figure 3 shows a comparison of EM and Costeira and Kanade's algorithm on Ullman's two cylinder demonstration as depicted in figure 1. 50 points were sampled uniformly from the surface of both objects, the trajectories of these points were computed and then projected onto each

[3] although for objects to comply with affine model they have to located relatively far from the camera, they are not placed at infinity.

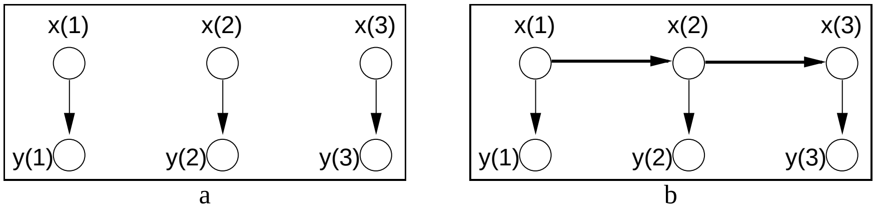

Figure 2: **a.** The graphical model assumed by most factorization algorithms for SFM. The camera location $x(t)$ is assumed to be independent of the camera location at any other time step. **b.** The graphical model assumed by our approach. We model temporal coherence by assuming a Markovian structure on the camera location.

frame to create the ordered noise free measurements matrix $\tilde{W}$. Then the columns of $\tilde{W}$ were shuffled by a random permutation $\pi \in S_{100}$ to create a noise free W. Each input matrix was added normally distributed zero-mean noise with the specified standard deviation. As can be seen, both algorithms work in the noise-free case (top two rows) but with small amounts of noise ($\sigma = 0.1$) Costeira and Kanade's algorithm segments the data incorrectly (and hence obtains the wrong structure) while EM continues to work. In order to help Costeira and Kanade's algorithm we gave it the correct rank of the measurement matrix (as discussed in [2] this step is often a failure source for Costeira and Kanade's algorithm). The algorithms of [2, 16] cannot be applied because they require W to be of full rank.

Figure 4 shows quantitative comparisons of EM and Costeira and Kanade for for three different synthetic sequences as a function of noise level. It is apparent that all algorithms give perfect segmentation when there is no noise at all. As the amount of noise increases, the performance of [1] deteriorates rapidly, while EM-based segmentation continues to succeed for low amounts of noise and shows moderate increase in number of error for larger amounts of noise. It is also clear that EM with temporal coherence performs significantly better than EM without temporal coherence for noisy inputs. The algorithms of [2, 16] perform similar to [1] provided the actual rank of observation matrix in non-degenerate cases.

Figure 5 shows the performance of EM with time coherence as a function of the percentage of missing data. While all existing factorization algorithms cannot work with missing data, EM continues to perform well even when 50% of the data is missing. For comparison, we also show the algorithm of [1] when the observation matrix is first filled in using Jacobs' algorithm [6] and the correct rank is given to all algorithms.

Finally, we tested the different algorithms on a real sequence of two cans rotating horizontally around parallel different axes in different angular velocities. 149 feature points were tracked along 20 frames, from which 93 are from one can, and 56 are from the other. Some of the feature points were occluded in part of the sequence, due to the rotation. Figure 6a shows the first frame of the sequence and the tracks superimposed. Note that because all the motions are horizontal, a single fundamental matrix can explain the data and hence this sequence cannot be segmented using two-frame methods. Using EM with temporal coherence. 8 points were misclassified and the structure correctly shows the curved surface of the two cylinders. To test Costeira-Kanade, we took the maximal full submatrix of the measurements matrix. The result was 30 misclassified points and a failure in $3D$ structure reconstruction.

4. Discussion

Motion segmentation is a "chicken and egg" problem: In order to divide the input into different sets, the motion related to each set should be known, whereas on the same time, in order to compute the motion, the segmentation should be given.

In this paper we introduced an EM framework for finding best segmentation and $3D$ structure, while *averaging* over all possible motions.

Moreover, the EM framework allows us to place priors on both structure and motion and to deal with directional uncertainty and missing data. The EM iterations described in this paper are simple and computationally efficient. Using this framework, we achieve good results on challenging inputs and outperform other existing methods.

An interesting future work would be to combine informative priors on the grouping of points. An example of such a prior is proximity of feature points in image domain: usually points residing on the same object, appear close to each other in their pictures.

References

[1] J. Costeira and T. Kanade. A multi-body factorization method for motion analysis. In *ICCV*, pages 1071–, 1995.

[2] C.W. Gear. Multibody grouping from motion images. *IJCV*, pages 133–150, 1998.

[3] A. Gelb, editor. *Applied Optimal Estimation*. MIT Press, 1974.

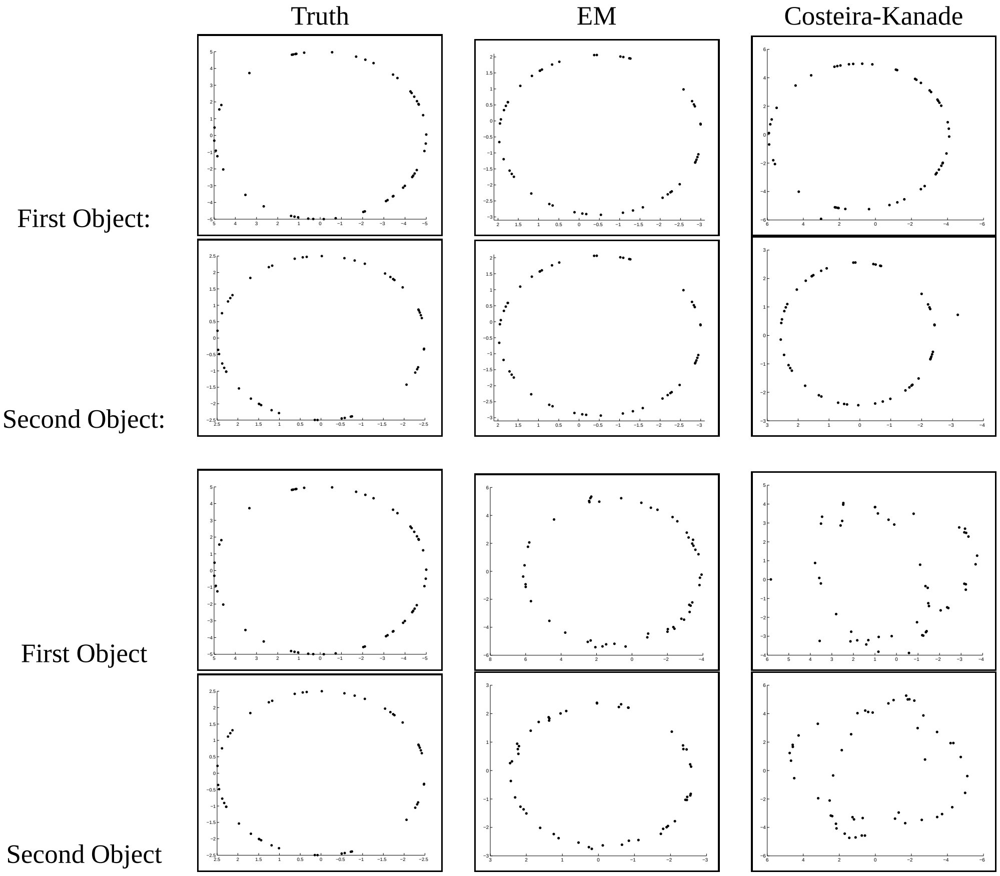

Figure 3: Comparison of EM and Costeira's algorithm on synthetic sequences, consisting of two coaxial cylinders rotating in different angular velocities around their elongated axis (as shown in figure 1). On top, results are shown for clean data. On bottom, results are shown for noisy data with standard deviation $\sigma = 0.1$. Results are displayed in top view. The reconstruction result of [1] in bottom right is similar to the reconstruction obtained when applying [11] on the observation matrix of the two objects (without segmentation).

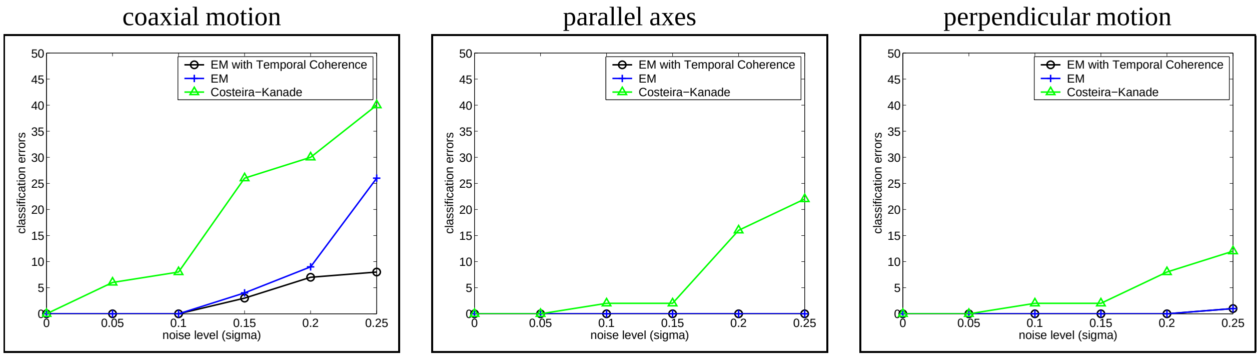

Figure 4: Comparison of different factorization algorithms for motion segmentation on synthetic inputs. Graphs display total number of misclassified points as a function of the noise standard deviation for $\sigma = 0, \ldots, 0.25$. In some of the experiments, the graphs of the two factor analysis versions overlap. **a.** sequence of concentric cylinders rotating in different speeds. Due to the input degeneracy only EM and [1] are compared. **b.** a cylinder and a cube rotating in the same speed around different parallel axes. **c.** A cube and a cylinder rotating around perpendicular axes.

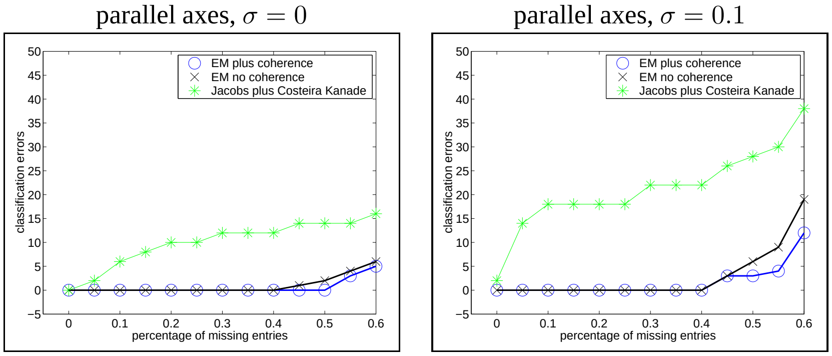

Figure 5: Performance of EM for segmentation with and without time coherence. Graphs show number of misclassifications as a function of the percentage of missing data. **a.** Cube and cylinder rotating around different parallel axes without noise. **b.** Cube and cylinder rotating around different parallel axes with noise with standard deviation $\sigma = 0.1$.

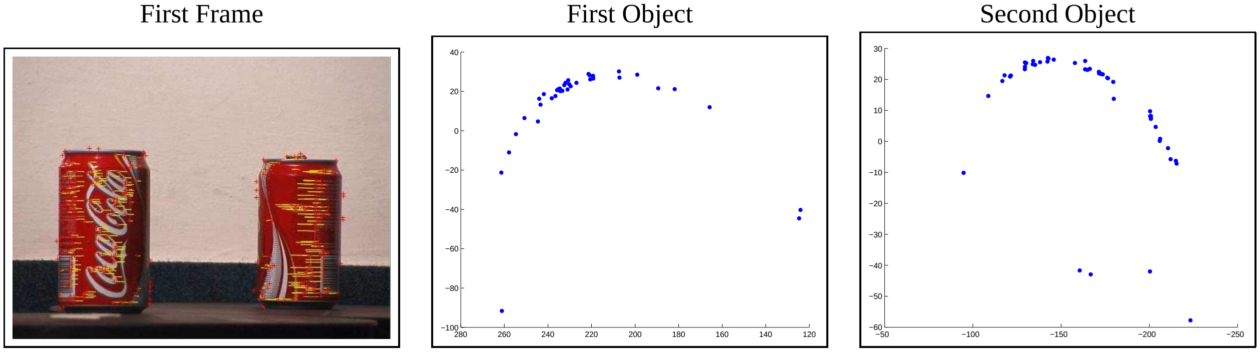

Figure 6: A real sequence of two cans rotating around different parallel axes. EM with temporal coherence succeeds to find correct segmentation and $3D$ structure reconstruction while other existing algorithms fail. See text for further details. **a.** First image from the input sequence with tracks found by tracking software superimposed. **b.** First segment, top view. **c.** Second segment, top view.

[4] Amit Gruber and Yair Weiss. Factorization with uncertainty and missing data: exploiting temporal coherence. In *NIPS 2003*, 2003.

[5] M. Irani and P. Anandan. Factorization with uncertainty. In *ECCV (1)*, pages 539–553, 2000.

[6] D. Jacobs. Linear fitting with missing data: Applications to structure-from-motion and to characterizing intensity images. In *CVPR*, pages 206–212, 1997.

[7] K. Kanatani. Evaluation and selection of models for motion segmentation. In *ECCV*, pages (3) 335–349, 2002.

[8] D. D. Morris and T. Kanade. A unified factorization algorithm for points, line segments and planes with uncertain models. In *ICCV*, pages 696–702, January 1999.

[9] S. Roweis. EM algorithms for PCA and SPCA. In *NIPS*, pages 431–437, 1997.

[10] D. Rubin and D. Thayer. EM algorithms for ML factor analysis. *Psychometrika 47(1)*, pages 69–76, 1982.

[11] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *Int. J. of Computer Vision*, 9(2):137–154, November 1992.

[12] S. Ullman. *The interpertation of visual motion*. MIT Press, 1979.

[13] R. Vidal, S. Soatto, Y. Ma, and S. Sastry. Segmentation of dynamic scenes from the multibody fundamental matrix.

[14] L. Wolf and A. Shashua. Two-body segmentation from two perspective views. pages 263–270, Dec. 2001.

[15] X.Feng and P.Perona. Scene segmentation from 3D motion. *CVPR*, pages 225–231, 1998.

[16] L. Zelnik-Manor, M. Machline, and M. Irani. Multi-body segmentation: Revisiting motion consistency, 2002.