Incorporating Non-motion Cues into 3D Motion Segmentation

Amit Gruber and Yair Weiss

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

Abstract. We address the problem of segmenting an image sequence into rigidly moving 3D objects. An elegant solution to this problem is the multibody factorization approach in which the measurement matrix is factored into lower rank matrices. Despite progress in factorization algorithms, the performance is still far from satisfactory and in scenes with missing data and noise, most existing algorithms fail.

In this paper we propose a method for incorporating 2D non-motion cues (such as spatial coherence) into multibody factorization. We formulate the problem in terms of constrained factor analysis and use the EM algorithm to find the segmentation. We show that adding these cues improves performance in real and synthetic sequences.

1 Introduction

The task of segmenting an image or an image sequence into objects is a basic step towards the understanding of image contents. Despite vast research in this area, performance of automatic systems still falls far behind human perception.

Motion segmentation provides a powerful cue for separating scenes consisting of multiple independently moving objects. Multibody factorization algorithms [1, 2, 3, 4, 5] provide an elegant framework for segmentation based on the 3D motion of the object. These methods get as input a matrix that contains the location of a number of points in many frames, and use algebraic factorization techniques to calculate the segmentation of the points into objects, as well as the 3D structure and motion of each object. A major advantage of these approaches is that they explicitly use the full temporal trajectory of every point, and therefore they are capable of segmenting objects whose motions cannot be distinguished using only two frames [4].

Despite recent progress in multibody factorization algorithms, their performance is still far from satisfactory. In many sequences, for which the correct segmentation is easily apparent from a single frame, current algorithms often fail to reach it.

Given the power of single frame cues and the poor performance of 3D motion segmentation algorithms, it seems natural to search for a common framework that could incorporate *both cues*. In this paper we provide such a framework. We use a latent variable approach to 3D motion segmentation and show how to modify the M step in an EM algorithm to take advantage of 2D affinities. We show that these cues improve performance in real and synthetic image sequences.

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part III, LNCS 3953, pp. 84–97, 2006.

1.1 Previous Work

The factorization approach to 3D segmentation has been suggested by Costeira and Kanade [1] who suggested to search for a block structure in the 3D structure matrix by computing a $P \times P$ affinity matrix Q from the SVD of the measurements matrix. 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 the segments is an NP-complete problem. Instead, Costeira and Kanade, suggested a greedy suboptimal clustering heuristic which turns out to be very sensitive to noise. In addition, the rank of the noise free measurements matrix should be found from the noisy measurements matrix as an initial step. 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 as an affinity matrix. Again, in noisy situations the algorithm does not guarantee correct segmentation. Some assumptions regarding the rank of the motion matrix are needed. Zelnik et al. [3] incorporate directional uncertainty by applying Gear's method on a matrix defined by measurable image quantities (spatial and temporal derivatives). Kanatani [6] proposed an algorithm that takes advantage of the affine subspace constraint. Recent works by [5, 4] have addressed the problem of motion segmentation with missing data. Both methods do not make any assumptions nor require prior information regarding the rank of the motion matrix. In addition [4] handles correlated non-uniform noise in the measurements and utilizes probabilistic prior knowledge on camera motion and scene structure.

Several authors have addressed the related, but different problem of 3D rigid body segmentation based on two frames or instantaneous motion [7, 8, 9, 10]. While these methods show encouraging results, they lack the attractive property of factorization methods in which information from the full temporal sequence is used simultaneously.

A different approach for image segmentation is to use single image cues such as color, intensity, texture and spatial proximity. A common approach is to present segmentation problems as problems of partitioning a weighted graph where the nodes of the graph represent pixels and the weights represent similarity or dissimilarity between them. Then some cost function of the partition should be minimized to find the desired segmentation. In many cases this optimization problem is NP-complete. Shi and Malik [11] introduced the Normalized Cut criterion and suggested an approximation algorithm based on the spectral properties of a weighted graph describing the affinities between pixels. Shi and Malik extend their work to 2D motion segmentation [12].

Single image cues have been used to improve the performance of various segmentation algorithms. An EM framework for incorporating spatial coherence and 2D image motion was presented in [13]. Kolmogorov and Zabih [14] have discussed incorporating spatial coherence into various segmentation algorithms. In this work we show how to incorporate spatial coherence (as well as other 2D non-motion cues) into $3D$ motion segmentation, as an extension of [15, 4].

1.2 Main Contribution of This Work

In this paper we present a unified framework for segmentation using information emerging from a diversity of cues. While previous segmentation methods can utilize either 3D motion information or 2D affinities but not both, we combine both sources of information.

We follow the constrained factorization approach for motion segmentation [4] based on the factorization formulation introduced by Costeira-Kanade [1]. We use 2D affinities to place priors on the desired semgnetation similar to [11] and show how the priors on the segmentation induce priors directly on the desired matrix factors. Then *constrained factorization with priors* is performed using the EM algorithm.

In contrast, previous approaches ([1, 2, 3, 5]) are based on algorithms (svd, reduced row echelon form, powerfactorization) which do not provide any apparent way to use priors on the segmentation.

Using the constrained factorization approach, we avoid the combinatorial search required by previous factorization approaches (e.g. [1, 2, 3]) in noisy scenarios. In our approach it is guaranteed to find a factorization where the interaction between points that belong to different motions is strictly 0 even in the presence of noise. In addition, with this formulation it is easy to deal with *missing data* and *directional uncertainty* (correlated non-uniform noise in the coordinates of tracked points such as the aperture problem). Another benefit of our formulation is that *no assumptions are made regarding the rank of the motion matrix* M (all affine motions are dealt with), and no prior knowledge about it is needed, unlike most previous methods for $3D$ motion segmentation that require some knowledge or assumptions regarding the rank of the motion matrix M (see [2] for discussion).

The EM algorithm is guaranteed to find a local maximum of the likelihood of S. Our experiments show that the additional information in the form of 2D affinities reduces the dependency in the initialization of the algorithm. Compared to the previous motion-only EM algorithm [4], the number of initializations required for success has diminished (details are given in the experiments section).

2 Model

2.1 3D Motion Segmentation – 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

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

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:

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

where

$$M = \begin{bmatrix} M_1^1 & \cdots & M_1^K \\ \vdots & & \\ M_F^1 & \cdots & M_F^K \end{bmatrix}_{2F\times 4K} \quad \text{and} \quad \tilde{S} = \begin{bmatrix} S_1 & 0 & \cdots & 0 \\ 0 & S_2 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & S_K \end{bmatrix}_{4K\times P} \tag{3}$$

If the segmentation $\{G_k\}_{k=1}^K$ were 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 [16]) on each submatrix. In real sequences, where segmentation is unknown, the observation matrix, W, is a column permutation of the ordered matrix $\tilde{W}$:

$$W = \tilde{W}\Pi = MS \Rightarrow S = \tilde{S}\Pi \tag{4}$$

where S is a $4K \times P$ matrix describing scene structure (with unordered columns) and $\Pi_{P\times P}$ is a column permutation matrix. Hence, the structure matrix S is in general not block diagonal, but rather a column permutation of a block diagonal matrix. The motion matrix, M, remains unchanged.

For noisy observations, the model is:

$$[W]_{2F\times P} = [M]_{2F\times 4K} [S]_{4K\times P} + [\eta]_{2F\times P} \tag{5}$$

where 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 $2D$ tracked feature points in frame t.

Let π be a labeling of all points, i.e. $\pi = (\pi_1, \ldots, \pi_P)$, where $\pi_p = k$ stands for point p is moving according to the kth motion ($p \in G_k$). Let s_p denote the $3D$

[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.

coordinates of point p, and let $\hat{S}$ denote $[s_1, \ldots, s_P]$ the $3D$ coordinates of all points (S contains both segmentation and geometry information, $\hat{S}$ contains only geometry information). Taking the negative log of the complete likelihood (which will be needed for the EM algorithm presented in section 3), the energy function due to 3D motion is:

$$E_{\text{3D-Motion}}(\hat{S}, \pi, M) = \sum_p E_{\text{3D-Motion}}(s_p, \pi_p, M) = \tag{6}$$
$$\sum_p \sum_t ((W_{t,p} - M_t^{\pi_p} s_p)^T \Psi_{t,p}^{-1} (W_{t,p} - M_t^{\pi_p} s_p))$$

Notice that if the motion M is given, $E_{\text{3D-Motion}}(\hat{S}, \pi, M)$ is a sum of functions of variables related to a single point independent of the others.

2.2 2D Affinities

We define affinity between pixels along a sequence similar to [11]. Shi et al. [11] define similarity weights in an image as the product of a feature similarity term and a spatial proximity term:

$$w_{i,j} = e^{\frac{-\|F(i)-F(j)\|_2^2}{\sigma_I^2}} \cdot \begin{cases} e^{\frac{-\|X(i)-X(j)\|_2^2}{\sigma_X^2}} & \text{if } \|X(i) - X(j)\|_2^2 < r \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $X(i)$ is the 2D coordinates of point i in the image and F is the vector of features used for segmentation. For example, if segmentation is performed according to spatial proximity, $F(i) = 1$ for all points.

The weights w_{ij} were defined in [11] for a single image. We adapt them to a sequence of images:

1. In order to use the information from all given frames rather than only one, we sum the energy terms $\frac{-\|X_t(i)-X_t(j)\|_2^2}{\sigma_X^2}$, $\frac{-\|F_t(i)-F_t(j)\|_2^2}{\sigma_I^2}$, over the entire sequence. In the summation, if one of the points is unobserved at a certain frame, this frame is omitted.
2. Since point locations along the sequence are the output of a tracking algorithm, they are given up to some uncertainty. We give weights to these locations according to $R_t^{-1}(i,j)$, the inverse covariance matrix of $(X_t(i) - X_t(j))$ in frame t (it can be shown that the posterior inverse covariance matrix of a 2D point location is $\begin{bmatrix} \sum I_x^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^2 \end{bmatrix}$. see [15,4,17]), thereby replacing $\|X_t(i) - X_t(j)\|_2^2$ with $(X_t(i) - X_t(j))^T R_t^{-1}(i,j)(X_t(i) - X_t(j))$. For frames where either point i or j is missing, $R_t^{-1}(i,j) = 0$. In other words, frame t is omitted from the summation.

The energy of an assignment due to spatial coherence is then:

$$E_{\text{2D-coherence}}(\pi) = \sum_{p,q} w_{p,q} \cdot (1 - \delta(\pi_p - \pi_q)) \tag{8}$$

Notice that $E_{\text{2D-coherence}}(\pi)$ is a sum of functions of two variables at a time.

3 An EM Algorithm for Multibody Factorization

Our goal is to find the best segmentation and 3D structure. We are looking for

$$\hat{S}, \pi = \arg\max_{\hat{S},\pi} \Pr(\hat{S}, \pi | W) = \arg\max_{\hat{S},\pi} < \Pr(\hat{S}, \pi | M, W) >_M \tag{9}$$

Maximizing the likelihood of W given $\hat{S}, \pi, M$ is equivalent to minimizing the energy $E(\hat{S}, \pi, M)$, i.e. the negative log of the likelihood function. This energy consists of two terms: a term of 3D motion information and a term of 2D coherence. These are the terms $E_{\text{3D-Motion}}(s_p, \pi_p, M)$ and $E_{\text{2D-coherence}}(\pi)$ introduced before.

$$E(\hat{S}, \pi, M) = E_{\text{3D-Motion}}(\hat{S}, \pi, M) + \lambda E_{\text{2D-coherence}}(\pi) \tag{10}$$

In order to find the optimal $\hat{S}$ and π, we minimize the energy with respect to $\hat{S}$ and π while averaging over M using the EM algorithm. The EM algorithm works with the expected complete log likelihood which is the expectation of the energy (taken with respect to the motion, M).

$$\begin{aligned} &E(\hat{S}, \pi) =< E(\hat{S}, \pi, M) >_M= \\ &< E_{\text{3D-Motion}}(\hat{S}, \pi, M) + \lambda E_{\text{2D-coherence}}(\pi) >_M= \\ &< E_{\text{3D-Motion}}(\hat{S}, \pi, M) >_M + \lambda E_{\text{2D-coherence}}(\pi) \end{aligned} \tag{11}$$

In the E step, sufficient statistics of the motion distribution are computed, such that $< E(\hat{S}, \pi, M) >_M$ can be computed for every $\hat{S}, \pi$. In the M-step, $< E(\hat{S}, \pi, M) >_M$ is minimized with respect to $\hat{S}$ and π.

3.1 Optimization with Respect to π

In this section, we focus on the optimization of $< E(\hat{S}, \pi, M) >_M$ with respect to π which is a part of the M-step. The missing details regarding the E step and optimization with respect to $\hat{S}$ in the M step are given in the next subsection.

The motion energy term (averaged over M) can be written as a sum of functions, D_p, each of which is a function of variables (s_p, π_p) related to a single pixel:

$$\begin{aligned} &< E_{\text{3D-Motion}}(\hat{S}, \pi, M) >_M= \sum_p < E_{\text{3D-Motion}}(s_p, \pi_p, M) >_M= \\ &\sum_p D_p(s_p, \pi_p) \end{aligned} \tag{12}$$

The 2D coherence energy function is a sum of terms of pairwise energy $V_{p,q}(\pi_p, \pi_q)$ for each pair of pixels p, q:

$$\begin{aligned} &E_{\text{2D-coherence}}(\pi) = \sum_{p,q} E_{\text{2D-coherence}}(\pi_p, \pi_q) = \\ &\sum_{p,q} V_{p,q}(\pi_p, \pi_q) \end{aligned} \tag{13}$$

Therefore $< E(\hat{S}, \pi, M) >_M$ can be represented as a sum of terms D_p involving a single point and terms $V_{p,q}$ involving pairs of points.

$$< E(\hat{S}, \pi, M) >_M = \sum_p D_p(s_p, \pi_p) + \sum_{p,q} V_{p,q}(\pi_p, \pi_q) \tag{14}$$

With this representation, if s_p is known for all p, then π_p can be found for all p by solving a standard energy minimization problem in a Potts model. In the binary case (i.e. $K = 2$), the optimal minimum of the energy function can be found efficiently using graph cuts [18]. If there are more than two objects, an approximation can be found using either graph cuts [18] or loopy belief propagation [19].

Since s_p is not known, we define

$$D_p(\pi_p) = \min_{s_p} D_p(s_p, \pi_p) \tag{15}$$

and then

$$\begin{aligned} &\min_{\hat{S},\pi} < E(\hat{S}, \pi, M) >_M = \\ &\min_{\pi} \left[\sum_p \min_{s_p} D_p(s_p, \pi_p) + \sum_{p,q} V_{p,q}(\pi_p, \pi_q) \right] = \\ &\min_{\pi} \left[\sum_p D_p(\pi_p) + \sum_{p,q} V_{p,q}(\pi_p, \pi_q) \right] \end{aligned} \tag{16}$$

In the next section we show how $D_p(\pi_p)$ is computed for each possible value of π_p. The pairwise terms, $V_{p,q}(\pi_p, \pi_q)$ are computed directly from the images. Given $D_p(\pi_p)$, $V_{p,q}(\pi_p, \pi_q)$ for all possible values of π_p, π_q, then one of the standard minimization algorithms for Potts model can be applied to find the optimal π.

3.2 Complete Description of the EM Algorithm

In the E-step, sufficient statistics of $< E(\hat{S}, \pi) >_M$ are computed. Recall that only the 3D motion energy term of $< E(\hat{S}, \pi) >_M$ depends on M, therefore only calculation of the expectation $< E_{\text{3D-Motion}}(s_p, \pi_p, M_p) >_M$ is required ($E_{\text{2D-coherence}}(\pi)$ is constant with respect to M). We compute these sufficient statistics by representing the factorization problem of equation 5 as a problem of factor analysis [15, 4].

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

$$y(t) = Ax(t) + \eta(t) \tag{17}$$

with $x(t) \sim N(0, \sigma_x^2 I)$ and $\eta(t) \sim N(0, \Psi_t)$. We now show how to rewrite the multibody factorization problem in this form.

In equation 5 the horizontal and vertical coordinates of the same point appear in different rows. To get an equation with all the measurements taken from the same frame in the same line of the measurements matrix, It can be rewritten as:

$$[U \quad V]_{F\times 2P} = [M_U \quad M_V]_{F\times 8K} \begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix}_{8K\times 2P} + [\eta]_{F\times 2P} \tag{18}$$

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 18 is equivalent (transposed) to equation 17. For diagonal covariance matrices Ψ_t (the case where Ψ_t is not diagonal is discussed in [15, 4]) the standard algorithm [20] gives:

E step:

$$E(x(t)|y(t)) = \left(\sigma_x^{-2} I + A^T \Psi_t^{-1} A\right)^{-1} A^T \Psi_t^{-1} y(t) \tag{19}$$

$$V(x(t)|y(t)) = \left(\sigma_x^{-2} I + A^T \Psi_t^{-1} A\right)^{-1} \tag{20}$$

$$< x(t) > = E(x(t)|y(t)) \tag{21}$$

$$< x(t)x(t)^T > = V(x(t)|y(t)) + < x(t) >< x(t) >^T \tag{22}$$

Although in our setting the matrix A must satisfy certain constraints, the E-step (in which the matrix A is assumed to be given from the M-step) remains the same as in standard factor analysis. In [15], priors regarding the motion are incorporated into the E-step.

M step:

In the M-step, $< E(\hat{S}, \pi) >_M$ is minimized with respect to $\hat{S}$ and π. Section 3 describes how π is found provided that $D_p(k)$ is known. Here we describe how to compute $D_p(k)$ for all p and k before the algorithm from section 3 can be applied. We also describe how the optimal $\hat{S}$ is found.

Denote by s_p^k a vector of length 3 that contains the optimal $3D$ coordinates of point p assuming it belongs to motion model k. In other words,

$$s_p^k \triangleq \arg\min_{s_p} D_p(s_p, k) \tag{23}$$

For a diagonal noise covariance matrix Ψ_t (for a non-diagonal Ψ_t see [15, 4]), by taking the derivative of $< E(\hat{S}, \pi) >_M$, we get:

$$s_p^k = B_{pk} C_{pk}^{-1} \tag{24}$$

where

$$\begin{aligned} B_{pk} = \sum_t & \left[\Psi_t^{-1}(p,p)(u_{tp} - < d_t^k >) < m_k(t)^T > \right. \\ & \left. + \Psi_t^{-1}(p+P, p+P)(v_{tp} - < e_t^k >) < n_k(t) >^T \right] \\ C_{pk} = \sum_t & \left[\Psi_t^{-1}(p,p) < m_k(t) m_k(t)^T > \right. \\ & \left. + \Psi_t^{-1}(p+P, p+P) < n_k(t) n_k(t)^T > \right] \end{aligned} \tag{25}$$

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 $x(t)$). Notice that s_p^k depends only on the motion distribution, the observations of point p and k. It is independent on the other points and their assignments, given the motion distribution.

$D_p(k)$ is therefore

$$\begin{aligned}
&D_p(k) = \min_{s_p} D_p(s_p, k) = D_p(s_p^k, k) = \\
&< \sum_t (W_{t,p} - M_t^k s_p^k)^T \varPsi_{t,p}^{-1} (W_{t,p} - M_t^k s_p^k) >_M = \\
&\sum_t < (W_{t,p} - M_t^k s_p^k)^T \varPsi_{t,p}^{-1} (W_{t,p} - M_t^k s_p^k) >_M = \\
&\sum_t \left[W_{t,p}^T \varPsi_{t,p}^{-1} W_{t,p} - 2 < x_{t,k} >^T a_{p,k}^T \varPsi_{t,p}^{-1} W_{t,p} + \right. \\
&\left. \mathrm{trace}(a_{p,k}^T \varPsi_{t,p}^{-1} a_{p,k} < x_{t,k} x_{t,k}^T >) \right]
\end{aligned} \tag{26}$$

where $a_{p,k}$ is a 2×8 matrix $a_{p,k} = \begin{bmatrix} (s_p^k)^T & 0 \\ 0 & (s_p^k)^T \end{bmatrix}$, $x_{t,k}$ is the subvector of $x(t)$ corresponding to the k-th motion (entries $4(k-1)+1, \dots, 4k$ and $4K + 4(k-1) + 1, \dots, 4K + 4k$) and the required expectations $< x_{t,k} >_M$ and $< x_{t,k} x_{t,k}^T >_M$ were computed in the E-step.

Now that $D_p(k)$ is known for all p for every k, π is found as described in section 3. After finding π, then $s_p = s_p^{\pi_p}$. An outline of the algorithm is given in Algorithm 1.

Algorithm 1. An outline of the EM algorithm for segmentation

```
Iterate until convergence:

1. E-step:
   (a) for $t = 1, \dots, T$,
         – Compute $< x(t) >$, $< x(t) x(t)^T >$ using equations 19 - 22.
2. M-step:
   (a) for $p = 1, \dots, P$,
         – for $k = 1, \dots, K$,
              i. Compute $s_p^k$ using equation 24.
             ii. Compute $D_p(k)$ using equation 26.
   (b) find $\pi$ using a standard energy minimization algorithm (for example, graph cuts or BP).
   (c) for $p = 1, \dots, P$,
         – assign $s_p = s_p^{\pi_p}$
   (d) Update A
```

The proposed segmentation algorithm can handle correlated non-uniform noise and can be applied even when there is missing data (these are just points for which $\varPsi_t^{-1}(i,i) = 0$). See [15, 4] for further details. Even in the presence of noise and missing data, it is guaranteed to find a factorization where the structure matrix has at most 4 nonzero elements per column, resulting in increased robustness.

4 Experiments

In this section we test our algorithm and compare it to previous algorithms on synthetic and real sequences.

EM guarantees convergence to a local maximum which is dependent on the initialization. In these experiments, 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 (synthetic) noise free scenes, the global maximum is usually found with a single initialization. As the amount of noise increases, the number of initializations needed for success also increases. For the experiments reported here, the maximal number of initializations is 10 (for both EM versions).

The global minimum of equation 16 was found using graph cuts in the experiments with two objects. Loopy belief propagation was used for minimization in experiments with more than two objects.

4.1 Experiments with Synthetic Data

We begin with a series of synthetic examples that demonstrate our approach vs. previous approaches of: [2] [1, 11, 4] and normalized cut with affinities that are a linear combination of 2D affinities and the Costeira-Kanade motion interaction matrix (referred as NCut Motion+2D in table 1). The following scenarios are tested (see figure 1):

1. A scene containing two objects with different 3D motions that are located far away from each other,
2. A scene with two coaxial objects (and thus cannot be separated spatially) rotating with different angular velocities,
3. And a scene containing two objects that are close to each other and have similar (yet different) motions.

We test each of these scenes in the presence and absence of mild amount of noise ($\sigma = 0.5$) and significant amount of noise ($\sigma = 10$, that was selected to show the difference in the performance of the two versions of EM).

In the first scenario, both 3D motion and spatial proximity provide a good separation between the objects. All algorithms have shown perfect results when there was no noise, as expected. Once mild amount of noise was added, the performance of CK (Costeira-Kanade, [1]) deteriorated while the segmentation results of the 3 other algorithms remained unchanged.

In the second scenario, the objects cannot be separated spatially in each individual image, but in the overall sequence some spatial information exists as the objects have different angular velocities. Despite of the existence of some spatial information, both versions of Normalized Cut failed to segment the objects in both the clean and noisy scenes. The other 3 algorithms have separated the objects perfectly in the noise free scenario due to their different motions (which were chosen to create a full rank motion matrix, M). Once mild amount of noise was added, again CK failed while the results of

[2] For the experiments with Normalized Cut we used the code available at http://www.seas.upenn.edu/~timothee/software_ncut/software.html

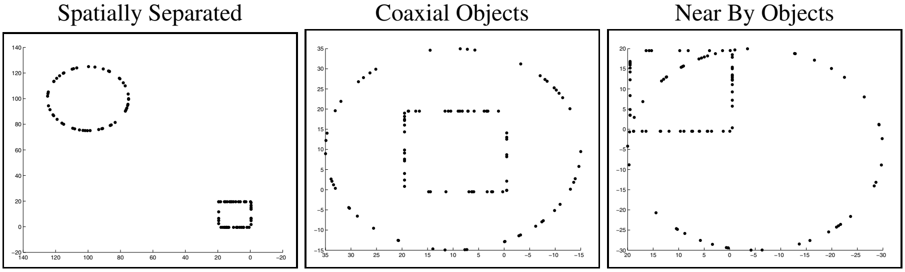

Fig. 1. Three scenarios we use to demonstrate the power of combining 3D motion information and 2D affinities. We compare the EM algorithm proposed in this paper with [1],[11] and [4] on clean and noisy ($\sigma = 0.5$ and $\sigma = 10$) input for scenes where (a) objects are spatially separable (and have different motions), (b) motion is the only cue for separation and (c) information from both sources is available, and in the noisy case required for separation. results are reported in table 1.

Table 1. Numbers of points that were misclassified by each of the algorithms for scenes from figure 1. These examples demonstrate the additional power of combining motion and non-motion information and the robustness of EM.

Scene Properties	Costeira-Kanade [1]	NCut [11]	NCut Motion+2D	Motion-Only EM [4]	Spatially-Coherent EM
Clean, Spatially separated	0	0	0	0	0
Noisy ($\sigma = 0.5$), Spatially separated	26	0	0	0	0
Noisy ($\sigma = 10$), Spatially separated	34	0	0	9	0
Clean, Coaxial	0	22	23	0	0
Noisy ($\sigma = 0.5$), Coaxial	14	22	25	1	0
Noisy ($\sigma = 10$), Coaxial	46	29	28	33	31
Clean, Near by	0	25	24	0	0
Noisy ($\sigma = 0.5$), Near by	8	25	24	3	1
Noisy ($\sigma = 10$), Near by	48	35	27	27	1

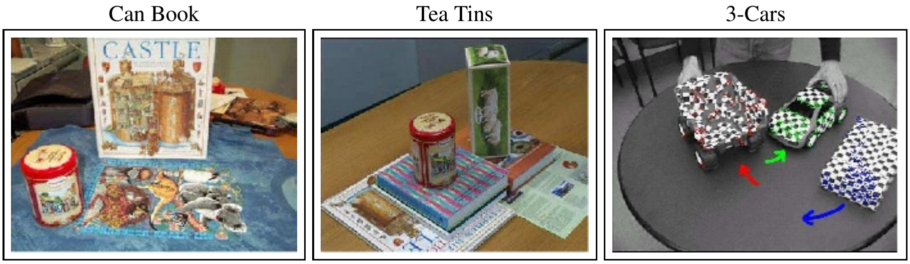

Fig. 2. The first image from the sequences: Can Book, Tea Tins and 3-Cars used for comparing the proposed EM algorithm and [5]. Results of this comparison are summarized in table 2.

both versions of the EM algorithm did not change. When a significant amount of noise was added, all algorithms failed because there was not enough information neither in the 3D motion nor in the spatial proximity.

Table 2. Misclassification error of segmentation using PowerFactorization and GPCA ([5]) and the algorithm proposed in this paper (EM) for the inputs from [5]

Sequence	Points	Frames	Motions	GPCA [5]	EM
Can Book	170	3	2	1.18%	0.00%
Tea Tins	84	3	2	1.19%	0.00%
3-Cars	173	15	3	4.62%	0.00%
Puma	64	16	2	0.00%	0.00%
Castle	56	11	2	0.00%	0.00%

In the last scenario we tested what do we gain from the combination of 2D spatial coherence information and 3D motion information. Although objects were not coaxial as in the previous scenario, they were not separated spatially well enough and both versions of Normalized Cut failed in both the noise free and noisy scenarios. As in previous cases CK found perfect segmentation when there was no noise, but failed in the noisy case. In the presence of significant amount of noise, we see that spatially coherent EM utilizes spatial information if it exists, and outperforms motion-only based EM.

4.2 Experiments with Real Data

We compared our algorithm to GPCA [5] by using 5 sequences that appeared in [5]. These sequences contain degenerate and non-degenerate motions, some contain only 3 frames. In this experiment, the maximal number of initializations of EM was 5, and the results of GPCA were taken from [5]. The results are presented in table 2: EM shows perfect results on all these input sequences, even when the number of frames is small (3) or when the motions matrix, M, is rank deficient.

Next, we checked the performance of spatially coherent EM on the sequence used in [4] and compared it to the motion based only EM algorithm from [4]. The input sequence consists 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. Using motion-only based

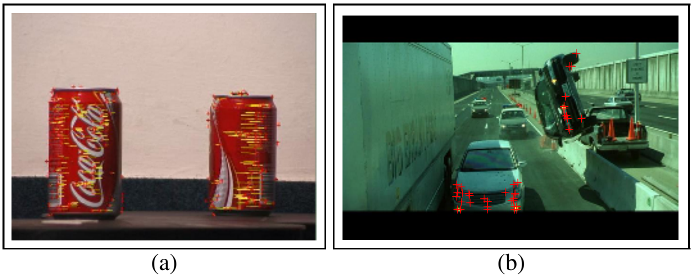
(a) (b)

Fig. 3. (a) A sequence of two cans rotating around different parallel axes. Spatial coherent EM succeeds to find correct segmentation and 3D structure up to 2 segmentation errors, comparing to 8 of motion-only EM and a failure of other methods. (b) First out of 13 frames taken from "Matrix Reloaded". 6 points were misclassified comparing to 14 by motion-only EM.

EM 8 points were misclassified. With the addition of spatial coherence, only 2 points were misclassified and the 3D was correctly reconstructed. Figure 3(a) shows the first frame of the sequence and the tracks superimposed. For comparison, Costeira-Kanade (using the maximal full submatrix of the measurements matrix) resulted in 30 misclassified points and a failure in $3D$ structure reconstruction.

Our last input sequence is taken from the film “Matrix Reloaded”. In this experiment 69 points were tracked along 13 frames: 28 on the car rotating in the air and 41 points were tracked on the front car approaching on the left (see figure 3(b)). On each object, points were selected from to be roughly in the same depth to avoid projective effects. Spatially coherent EM misclassified 6 points comparing to 14 points that were misclassified by motion-only EM and 19 points that were misclassified by Ncut.

5 Discussion

In this paper we presented an algorithm for incorporating 2D non-motion affinities into 3D motion segmentation using the EM algorithm. We showed that using a coherence prior on the segmentation is easily implemented and gives rise to better segmentation results. In the E step, the mean and covariance of the 3D motions are calculated using matrix operations, and in the M step the structure and the segmentation are calculated by performing energy minimization.

With the EM framework, missing data and directional uncertainty are easily handled. Placing meaningful priors and imposing constraints on the desired factorization greatly increase the robustness of the algorithm.

Future work includes incorporation of other sources of information, for examples other principles of perceptual organization suggested by the Gestalt psychologists. Another direction for future work is to place the spatial coherence prior directly on the 3D of the points: solving together for structure and segmentation, where the prior on the segmentation depends directly on the reconstructed 3D structure.

Acknowledgements. Supported by EC Contract 027787 DIRAC. The authors would like to thank V. Kolmogorov for the graph cut code and T. Meltzer for the BP code.

References

1. Costeira, J., Kanade, T.: A multi-body factorization method for motion analysis. In: ICCV. (1995)
2. Gear, C.: Multibody grouping from motion images. IJCV (1998) 133–150
3. Zelnik-Manor, L., Machline, M., Irani, M.: Multi-body segmentation: Revisiting motion consistency (2002)
4. Gruber, A., Weiss, Y.: Multibody factorization with uncertainty and missing data using the EM algorithm. In: Computer Vision and Pattern Recognition (CVPR). (2004)
5. Vidal, R., Hartely, R.: Motion segmentation with missing data using powerfactorization and gpca. In: Computer Vision and Pattern Recognition (CVPR). (2004)
6. Kanatani, K.: Evaluation and selection of models for motion segmentation. In: ECCV. (2002) (3) 335–349

7. Vidal, R., Soatto, S., Ma, Y., Sastry, S.: Segmentation of dynamic scenes from the multibody fundamental matrix (2002)
8. Wolf, L., Shashua, A.: Two-body segmentation from two perspective views. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2001) 263–270
9. Feng, X., Perona, P.: Scene segmentation from 3D motion. CVPR (1998) 225–231
10. MacLean, W.J., Jepson, A.D., Frecker, R.C.: Recovery of egomotion and segmentation of independent object motion using the em algorithm. In: BMVC. (1994)
11. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 888–905
12. Shi, J., Malik, J.: Motion segmentation and tracking using normalized cuts. In: ICCV. (1998) 1154–1160
13. Weiss, Y., Adelson, E.: A unified mixture framework for motion segmentation: incorporating spatial coherence and estimating the number of models. In: Proceedings of IEEE conference on Computer Vision and Pattern Recognition. (1996) 321–326
14. Zabih, R., Kolmogorov, V.: Spatially coherent clustering with graph cuts. In: Computer Vision and Pattern Recognition (CVPR). (2004)
15. Gruber, A., Weiss, Y.: Factorization with uncertainty and missing data: Exploiting temporal coherence. In: Neural Information Processing Systems (NIPS). (2003)
16. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: A factorization method. Int. J. of Computer Vision **9** (1992) 137–154
17. Irani, M., Anandan, P.: Factorization with uncertainty. In: ECCV (1). (2000) 539–553
18. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts ? Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2004)
19. Weiss, Y., Freeman, W.T.: On the optimality of solutions of the max-product belief propagation algorithm in arbitrary graphs. IEEE Transactions on Information Theory **47** (2001) 723–735
20. Rubin, D., Thayer, D.: EM algorithms for ML factor analysis. Psychometrika 47(1) (1982) 69–76