Computer Vision and Image Understanding 113 (2009) 319–331

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Attention-from-motion: A factorization approach for detecting attention objects in motion

Yiqun Hu, Deepu Rajan *, Liang-Tien Chia

School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798, Singapore

ARTICLE INFO

Article history:
Received 31 October 2007
Accepted 18 August 2008
Available online 5 September 2008

Keywords:
Visual attention
Factorization
Motion segmentation

ABSTRACT

This paper introduces the notion of attention-from-motion in which the objective is to identify, from an image sequence, only those object in motions that capture visual attention (VA). Following the important concept in film production, viz, the tracking shot, we define the attention object in motion (AOM) as those that are tracked by the camera. Three components are proposed to form an attention-from-motion framework: (i) a new factorization form of the measurement matrix to describe dynamic geometry of moving object observed by moving camera; (ii) determination of single AOM based on the analysis of certain structure on the motion matrix; (iii) an iterative framework for detecting multiple AOMs. The proposed analysis of structure from factorization enables the detection of AOMs even when only partial data is available due to occlusion and over-segmentation. Without recovering the motion of either object or camera, the proposed method can detect AOM robustly from any combination of camera motion and object motion and even for degenerate motion.

1. Introduction

Research in cognition and neuroscience has shown that the primate visual system can selectively focus attention on some specific motion pattern in a scene, while ignoring other motion patterns in it [1,2]. Detection of motion that captures visual attention is useful in several applications like video surveillance [3], video summarization [4], and video editing [5].

Previous research on detecting motion patterns that capture visual attention use heuristic notions of salient motion. Williams and Draper [6] studied the effect of motion on visual attention by adding the motion channel to a saliency map that encodes the attention value of every pixel. They concluded that motion contrast cannot improve the performance of saliency-based selective attention. This shows that using information about contrast that appears as a result of motion is not a good cue to find the "interesting" object in the video sequence. Ma et al. [4] used motion vector fields to compute the motion energy to identify salient motion regions. Tian and Hampapur [7] discriminated salient motion from unimportant motion using a consistency measure. Wixson [8] also used a consistency measure to detect interesting motion by accumulating directionally consistent flow. In all the above, regions that attract visual attention due to motion are defined heuristically, e.g., based on features like motion energy and motion consistency. Such definitions are not applicable in many instances, for example, even if motions of all objects are consistent, only one of them might invite visual attention. Similarly, an object with low motion energy can also be the one on which visual attention is centered. Moreover, these features are also not robust to noise. In [9], Horaud et al. developed a framework in which a static camera is cooperated with an active camera to detect all moving objects. Here, there is no discrimination among detected objects in terms of visual attention. In [10], López et al. extracted motion features like motion presence, module and angle of velocity to segment the moving objects and the user is invited to specify the objects that capture visual attention. The extraction of interesting object in this case is not automatic. Other motion analysis algorithms like [11,12] ignore the concept of visual attention. In this paper, we describe an automatic method to extract the interesting moving object which is defined according to what the cameraman wishes the viewer to focus on. We call this problem as **Attention-from-Motion**.

1.1. Attention-from-Motion

Video production manuals provide insight into the procedures used during video creation and editing. One of the most important shots is the "tracking shot" in which the camera follows one or more moving objects [13]. Naturally, the intention of the cameraman is to focus the viewer's attention on the object(s) which he is following. Thus, we define an **Attention Object in Motion (AOM)** as an object that is followed by the camera. As illustrated in Fig. 1, although there are two cars in the field of view of the camera, it follows only one of them, so that the car on the top-left is an AOM while the one on the bottom-right is not. The notion of "camera following an object" embeds the subjective attention of the cameraman and will eventually

* Corresponding author. Fax: +65 67926559.
E-mail address: asdrajan@ntu.edu.sg (D. Rajan).

1077-3142/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.08.010

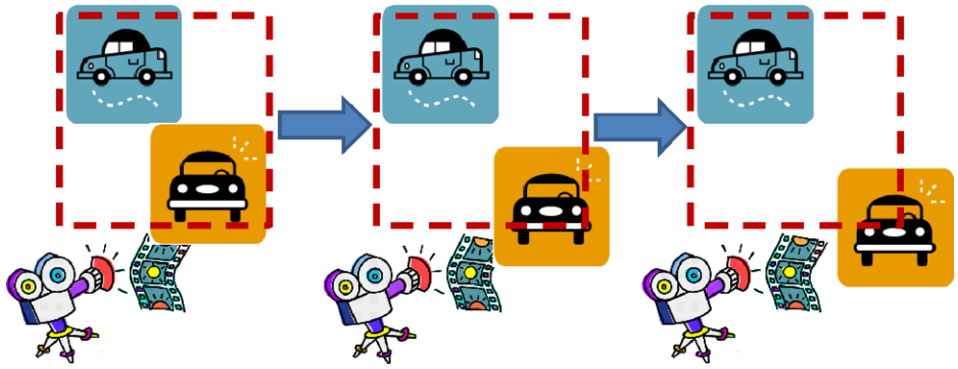

Fig. 1. Illustration of attention object in motion (the dashed square indicates the field of view of the camera). The camera follows the car on the top-left (AOM) while ignoring the car on the bottom-right (not AOM).

guide the attention of the viewers. This definition of AOM is applicable for videos produced professionally as well as for videos taken by general consumers of hand-held video cameras. We define the problem of **attention-from-motion** as detecting AOMs from the motion information in a video sequence. The motion information is provided as a measurement matrix of P feature points across F frames. The definition of AOM embeds the "following" relationship between camera motion and object motion.

In this paper, we develop an algorithm to detect AOMs based on factorization of a measurement matrix, under the assumption of orthographic projection and rigid body motion. When the camera tracks an AOM, the factorization relating to the motion matrix attains a special structure as described in Section 2. We describe the method to identify this structure *without carrying out the complete factorization* (Section 3). Finally, an iterative framework is proposed to robustly extract more than one AOM (Section 4). The proposed solution to the attention-from-motion problem is not constrained by the type of camera/object motion including degenerate as well as dependent motion. Furthermore, the algorithm can detect AOMs even when only partial data is available, e.g., due to occlusion.

2. Factorization method for AOM

Factorization methods have been used in [14,15] to solve the structure-from-motion problem. In [15], a moving object is observed by a static camera while in [14] a static object is observed by a moving camera. We generalize the above two methods to consider the case of a moving camera observing an object in motion under the assumptions of orthographic projection and rigid body motion. The factorization results in a motion matrix and a shape matrix such that the motion matrix corresponding to an AOM has a special structure.

2.1. Dynamic scene factorization

The *measurement matrix* W consists of a $F \times P$ submatrix X of horizontal feature coordinates x_{fp}, with F being the number of frames and P being the number of feature points and a similar $F \times P$ submatrix Y of vertical coordinates y_{fp}. Thus $W = \left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]$. According to rigid-body motion assumption, the 3-D coordinates of the pth feature point on an object at the fth frame is

$$s_{fp} = R_{o,f} s_{1p} + T_{o,f}, \quad p = 1, \ldots, P \quad \text{and} \quad f = 1, \ldots, F, \tag{1}$$

where $R_{o,f}$ is the rotation matrix and $T_{o,f}$ is the translation vector of the object at the fth frame with respect to the world coordinates. By placing the origin of the world coordinates at the centroid of object feature points at the 1st frame, we have

$$\frac{1}{P}\sum_{p=1}^{P} s_{1p} = 0. \tag{2}$$

Combining (2) with (1), we get

$$\frac{1}{P}\sum_{p=1}^{P} s_{fp} = \frac{1}{P}\sum_{p=1}^{P}(R_{o,f} s_{1p} + T_{o,f}) = T_{o,f}. \tag{3}$$

Under orthography, the image feature position (x_{fp}, y_{fp}) of point s_{fp} at frame f is given by the equations [14]

$$x_{fp} = i_f^T (s_{fp} - T_{c,f}), \quad y_{fp} = j_f^T (s_{fp} - T_{c,f}), \tag{4}$$

where, i_f and j_f are the unit vectors in frame f pointing along the rows and columns, respectively, of the image and defined w.r.t the world reference system, and $T_{c,f}$ is the translation vector of the camera w.r.t the world coordinates. If $R_{c,f}$ is the rotation matrix of the camera w.r.t. the world coordinates, then

$$\begin{aligned} i_f &= R_{c,f} i_0 + T_{c,f} - T_{c,f} = R_{c,f} i_0 \\ j_f &= R_{c,f} j_0 + T_{c,f} - T_{c,f} = R_{c,f} j_0, \end{aligned} \tag{5}$$

where $i_0 = [1,0,0]^{\mathrm{T}}$ and $j_0 = [0,1,0]^{\mathrm{T}}$ are the two axes of the canonical world coordinate system as shown in Fig. 2. Any other representation of i_0 and j_0 will not affect the subsequent analysis.

Using (1) and (5), the entries of X can be derived as

$$\begin{aligned} x_{fp} &= i_f^{\mathrm{T}}(s_{fp} - T_{c,f}) \\ &= i_0^{\mathrm{T}} R_{c,f}^{\mathrm{T}} R_{o,f} s_{1p} + i_0^{\mathrm{T}} R_{c,f}^{\mathrm{T}} (T_{o,f} - T_{c,f}). \end{aligned} \tag{6}$$

Similarly, the entries of Y can also be derived as

$$y_{fp} = j_0^{\mathrm{T}} R_{c,f}^{\mathrm{T}} R_{o,f} s_{1p} + j_0^{\mathrm{T}} R_{c,f}^{\mathrm{T}} (T_{o,f} - T_{c,f}). \tag{7}$$

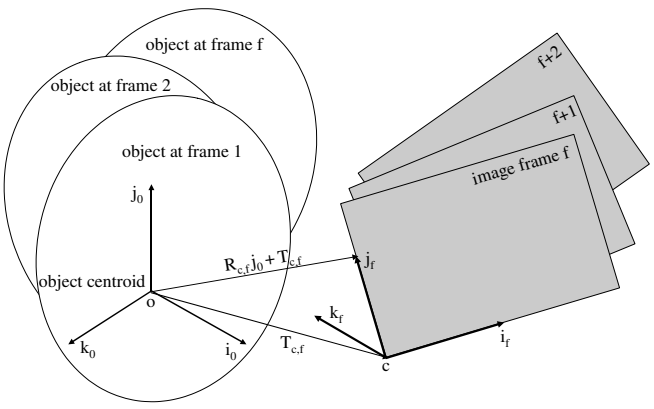

Fig. 2. The reference coordinate system used in the factorization of dynamic scene.

The above two sets of $F \times P$ Eqs. ((6) and (7)) form the unregistered measurement matrix W for the case of a moving camera observing a moving object. The matrix W can be factorized into $W = M \times S$ as

$$W = \begin{bmatrix} X \\ \hline Y \end{bmatrix} = \underbrace{\left[\begin{array}{cc} i_0^T R_{c,1}^T R_{o,1} & i_0^T R_{c,1}^T (T_{o,1} - T_{c,1}) \\ i_0^T R_{c,2}^T R_{o,2} & i_0^T R_{c,2}^T (T_{o,2} - T_{c,2}) \\ \vdots & \vdots \\ i_0^T R_{c,F}^T R_{o,F} & i_0^T R_{c,F}^T (T_{o,F} - T_{c,F}) \\ \hline j_0^T R_{c,1}^T R_{o,1} & j_0^T R_{c,1}^T (T_{o,1} - T_{c,1}) \\ j_0^T R_{c,2}^T R_{o,2} & j_0^T R_{c,2}^T (T_{o,2} - T_{c,2}) \\ \vdots & \vdots \\ j_0^T R_{c,F}^T R_{o,F} & j_0^T R_{c,F}^T (T_{o,F} - T_{c,F}) \end{array}\right]}_{M(2F\times 4)} \underbrace{\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1P} \\ 1 & 1 & \cdots & 1 \end{bmatrix}}_{S(4\times P)} \tag{8}$$

The matrix M in (8) contains the motion of both camera and object while in [15] and [14], M contains only object motion and only camera motion, respectively. Since M is $2F \times 4$ and S is $4 \times P$, (8) yields the **Rank Theorem** that states that *without noise, the measurement matrix W of a moving object observed by a moving camera is at most of rank four.*

2.2. Factorization of AOM

Recall that we define an AOM as an object in motion that is followed by a moving camera. The fact that the moving object is followed by the moving camera implies that, in the absence of noise,

$$\frac{1}{P}\sum_{p=1}^{P} x_{fp} = \frac{1}{P}\sum_{p=1}^{P} x_{1p} \quad \forall f \in \{1,\ldots,F\} \tag{9}$$

and

$$\frac{1}{P}\sum_{p=1}^{P} y_{fp} = \frac{1}{P}\sum_{p=1}^{P} y_{1p} \quad \forall f \in \{1,\ldots,F\}. \tag{10}$$

If this condition is true over a small number of frames, but not true in the long term, then according to our definition, the feature points belong to an AOM initially, but it is no longer an AOM when the conditions in (9) and (10) fail. Using Eqs. (1), (4) and (5), (9) can be rewritten as

$$\begin{aligned} &\frac{1}{P}\sum_{p=1}^{P} i_f^T (s_{fp} - T_{c,f}) = \frac{1}{P}\sum_{p=1}^{P} i_1^T (s_{1p} - T_{c,1}) \\ &\Rightarrow \frac{1}{P}\sum_{p=1}^{P} i_f^T (R_{o,f} s_{1p} + T_{o,f} - T_{c,f}) = \frac{1}{P}\sum_{p=1}^{P} i_1^T (s_{1p} - T_{c,1}) \\ &\Rightarrow \frac{1}{P}\sum_{p=1}^{P} i_0^T R_{c,f}^T (R_{o,f} s_{1p} + T_{o,f} - T_{c,f}) = \frac{1}{P}\sum_{p=1}^{P} i_0^T R_{c,1}^T (s_{1p} - T_{c,1}) \\ &\Rightarrow i_0^T R_{c,f}^T R_{o,f} \underbrace{\frac{1}{P}\sum_{p=1}^{P} s_{1p}}_{=0} + i_0^T R_{c,f}^T (T_{o,f} - T_{c,f}) = i_0^T R_{c,1}^T \underbrace{\frac{1}{P}\sum_{p=1}^{P} s_{1p}}_{=0} - i_0^T R_{c,1}^T T_{c,1} \\ &\Rightarrow i_0^T R_{c,f}^T (T_{o,f} - T_{c,f}) = -i_0^T R_{c,1}^T T_{c,1} \quad \forall f \in \{1,\ldots,F\} \end{aligned} \tag{11}$$

Similarly, it can be shown that

$$j_0^T R_{c,f}^T (T_{o,f} - T_{c,f}) = -j_0^T R_{c,1}^T T_{c,1} \quad \forall f \in \{1,\ldots,F\} \tag{12}$$

Interestingly, the left sides of (11) and (12) are the entries in the last column of M in (8). Hence, for an AOM, the motion matrix M can be represented as:

$$M = \left[\begin{array}{cc} i_0^T R_{c,1}^T R_{o,1} & -i_0^T R_{c,1}^T T_{c,1} \\ i_0^T R_{c,2}^T R_{o,2} & -i_0^T R_{c,1}^T T_{c,1} \\ \vdots & \vdots \\ i_0^T R_{c,F}^T R_{o,F} & -i_0^T R_{c,1}^T T_{c,1} \\ \hline j_0^T R_{c,1}^T R_{o,1} & -j_0^T R_{c,1}^T T_{c,1} \\ j_0^T R_{c,2}^T R_{o,2} & -j_0^T R_{c,1}^T T_{c,1} \\ \vdots & \vdots \\ j_0^T R_{c,F}^T R_{o,F} & -j_0^T R_{c,1}^T T_{c,1} \end{array}\right]. \tag{13}$$

In the above $2F \times 4$ matrix, the first 3 columns relate to the rotation component and the 4th column relates to the translation component of the dynamic scene. This structure allows us to impose the following motion constraints:

- *Rotation constraint:* Each pair of rows in the first 3 columns of M are the rotations of two reference orthonormal basis (e.g. $[1,0,0]^T$ and $[0,1,0]^T$) of M since $R_{c,f}^T R_{o,f}$ is a rotation matrix such that $i_0^T R_{c,f}^T R_{o,f}$ and $j_0^T R_{c,f}^T R_{o,f}$ are the rotations of i_0 and j_0, respectively.
- *Translation constraint:* The elements of the first half of the last column of M are constants and so also are the elements of the second half of the last column.

2.3. Why not simply check for stationary centroid?

Since the special structure of M is derived from the property of a stable centroid of *all* the feature points belonging to an AOM, it would seem that checking for the stationarity of the centroid would suffice to identify the AOM. However, this would imply that all the feature points are available to cover the whole object, otherwise the centroid would not be stable. This is a very strong constraint because feature points could be lost for a variety of reasons like occlusion and change in appearance of the object due to motion or change in illumination. Moreover, the limitations of the feature detection algorithm itself may cause only some of the feature points on an object to be detected resulting in the tracking of an erroneous centroid. Fig. 3 illustrates this problem in which four frames of a sequence show a rotating rectangle whose four corners are the feature points. Since the centroid of the feature points indicated by the black cross is at the same location, the rectangle is an AOM. However, if the feature detection algorithm returns only two feature points (red and yellow) then its centroid as indicated by the pink cross is not stationary.

Another case when checking for the stationarity of the centroid of feature points fails is that of an object revolving about a point. In this case, the cameraman will most likely wish to capture the entire circular motion by focusing on the centroid of the motion rather than on the centroid of the feature points on the object. Fig. 4 shows a few frames of this type of motion in which the ball is revolving about the point marked +; here the centroid of the feature points changes across all frames, but the camera will focus on the +. Furthermore, one would recognize the motion of the object as a circular motion about a point only if all the frames of at least one cycle are observable. If only some of the frames of the entire cycle are available, as shown in Fig. 4, it is not possible to recognize the circular motion because the centroid of the motion will not be the point about which the object rotates.

The structure of M is valid in all of the above cases. For the case when only partial feature points are available, only a subset of feature points are required to calculate M and to analyze its structure to determine if they belong to an AOM. This is because it is enough to obtain the matrix M from the measurement matrix corresponding to a subset of feature points as long as its rank is equal to the

rank of the measurement matrix corresponding to the full set of feature points. Also, in our analysis, the requirement that the last column of M is constant is valid independent of the number of frames observed. This makes the proposed method robust to occlusion and avoids the propagation of tracking error over long intervals. The structure of M is valid, even when the centroid of the moving object is not stationary (e.g., in Fig. 4) because the AOM can be considered as a part of another object whose centroid is the centroid of motion. Thus, the proposed method works even if only partial data is available.

3. Attention-from-motion for a single AOM

In [14], it is shown that the factorization of the measurement matrix W using *singular value decomposition (SVD)* gives

$$W = U\Sigma V^T = (U\Sigma^{\frac{1}{2}}) \cdot (\Sigma^{\frac{1}{2}}V^T) = \hat{M}\hat{S}. \tag{14}$$

This factorization is not unique since for any 4×4 **invertible** matrix Q, the matrices $\hat{M}Q$ and $Q^{-1}\hat{S}$ are also a valid factorization of W, i.e.,

$$(\hat{M}Q)(Q^{-1}\hat{S}) = \hat{M}(QQ^{-1})\hat{S} = \hat{M}\hat{S} = W. \tag{15}$$

In the following, we describe a linear method to determine the existance and singularity of Q only from the affine version $\hat{M}$ of M. If there exists a Q that transforms $\hat{M}$ to M (of Eq. (13)), then the feature points in the measurement matrix W belong to an AOM. The result is then used to define an attention measure to identify an AOM.

3.1. Constraints on $\hat{M}$ and Q

By denoting the matrix Q as $Q = [Q_R|Q_T]$ where Q_R contains the first 3 columns of Q, and Q_T contains the last column of Q, M can be represented as

$$M = \hat{M}Q = [\hat{M}Q_R|\hat{M}Q_T]. \tag{16}$$

$\hat{M}Q_R$ corresponds to the first 3 columns of M which satisfies the rotation constraint and $\hat{M}Q_T$ corresponds to the last column of M which satisfies the translation constraint.

3.1.1. Rotation constraint on $\hat{M}$ and Q

Each of the $2F$ rows of matrix $\hat{M}Q_R$ is a unit norm vector and the first and second set of F rows are pairwise orthogonal. This orthonormality can be represented as

$$\begin{aligned}
\hat{M}_i Q_R Q_R^{\mathrm{T}} \hat{M}_i^{\mathrm{T}} &= 1, \\
\hat{M}_j Q_R Q_R^{\mathrm{T}} \hat{M}_j^{\mathrm{T}} &= 1, \\
\hat{M}_i Q_R Q_R^{\mathrm{T}} \hat{M}_j^{\mathrm{T}} &= 0,
\end{aligned} \tag{17}$$

for $i = 1 \ldots F$, $j = F + i$ and $\hat{M}_i$ denotes row i of $\hat{M}$. The set of Equations in (17) can be solved for Q_R by adding the additional constraints of the reference coordinate system (e.g., $i_0 = [1,0,0]^{\mathrm{T}}$ and $j_0 = [0,1,0]^{\mathrm{T}}$) to it. Instead, we solve for the entries of $Q_R Q_R^{\mathrm{T}}$, represented as

$$Q_R Q_R^{\mathrm{T}} = \begin{bmatrix} \mathcal{Q}_1 & \mathcal{Q}_5 & \mathcal{Q}_6 & \mathcal{Q}_7 \\ \mathcal{Q}_5 & \mathcal{Q}_2 & \mathcal{Q}_8 & \mathcal{Q}_9 \\ \mathcal{Q}_6 & \mathcal{Q}_8 & \mathcal{Q}_3 & \mathcal{Q}_{10} \\ \mathcal{Q}_7 & \mathcal{Q}_9 & \mathcal{Q}_{10} & \mathcal{Q}_4 \end{bmatrix}$$

and whose elements are

$$\begin{bmatrix} \mathcal{Q}_1 \\ \mathcal{Q}_2 \\ \mathcal{Q}_3 \\ \mathcal{Q}_4 \\ \mathcal{Q}_5 \\ \mathcal{Q}_6 \\ \mathcal{Q}_7 \\ \mathcal{Q}_8 \\ \mathcal{Q}_9 \\ \mathcal{Q}_{10} \end{bmatrix} = \begin{bmatrix} Q_{11}^2 + Q_{12}^2 + Q_{13}^2 \\ Q_{21}^2 + Q_{22}^2 + Q_{23}^2 \\ Q_{31}^2 + Q_{32}^2 + Q_{33}^2 \\ Q_{41}^2 + Q_{42}^2 + Q_{43}^2 \\ Q_{11}Q_{21} + Q_{12}Q_{22} + Q_{13}Q_{23} \\ Q_{11}Q_{31} + Q_{12}Q_{32} + Q_{13}Q_{33} \\ Q_{11}Q_{41} + Q_{12}Q_{42} + Q_{13}Q_{43} \\ Q_{21}Q_{31} + Q_{22}Q_{32} + Q_{23}Q_{33} \\ Q_{21}Q_{41} + Q_{22}Q_{42} + Q_{23}Q_{43} \\ Q_{31}Q_{41} + Q_{32}Q_{42} + Q_{33}Q_{43} \end{bmatrix} \tag{18}$$

by regarding (17) as a linear system about the 10 unique entries of $Q_R Q_R^{\mathrm{T}}$. If the linear system of (17) has a unique solution, we can directly obtain $\mathcal{Q}$ (the left hand side of (18)) using least squares technique. Note that unlike in [15,14] where Q is reconstructed after solving for $Q_R Q_R^{\mathrm{T}}$ in (17), *we do not need to reconstruct Q for carrying out attention analysis, as shown below.*

3.1.2. Translation constraint on $\hat{M}$ and Q

The fact that two halves of the last column of M are constants implies that $\hat{M}_1 Q_T = \hat{M}_2 Q_T = \cdots = \hat{M}_F Q_T$ and $\hat{M_{F+1}} Q_T = \hat{M_{F+2}} Q_T = \cdots = \hat{M_2 F} Q_T$. This constraint can be further transformed into a homogeneous linear system as

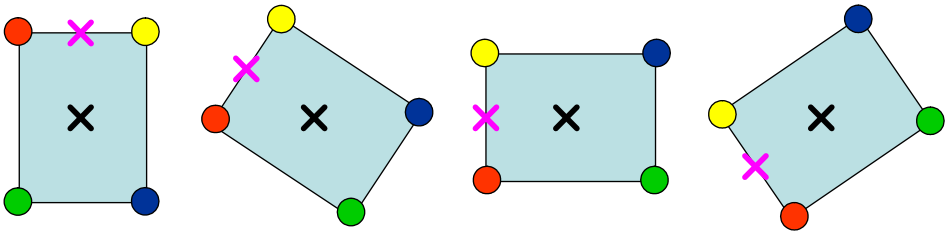

Fig. 3. An example of unstable centroid of AOM where only partial feature points on an AOM are observed.

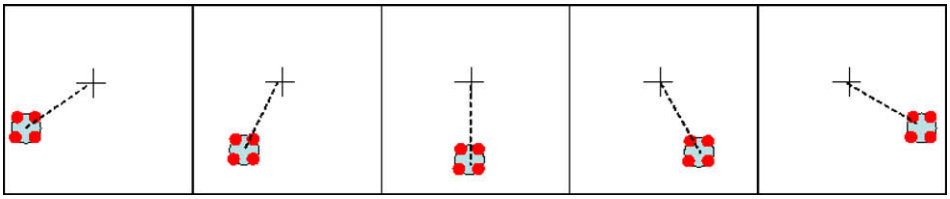

Fig. 4. Another example of unstable centroid of AOM where the centroid of motion is followed and only partial frames are observed.

$$\underbrace{\begin{bmatrix} \hat{M}_1 - \hat{M}_2 \\ \vdots \\ \hat{M}_{F-1} - \hat{M}_F \\ \hline \hat{M}_{F+1} - \hat{M}_{F+2} \\ \vdots \\ \hat{M}_{2F-1} - \hat{M}_{2F} \end{bmatrix}}_{P_T((2F-2)\times 4)} \underbrace{\begin{bmatrix} Q_{1,4} \\ Q_{2,4} \\ Q_{3,4} \\ Q_{4,4} \end{bmatrix}}_{Q_T} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{19}$$

For the existence of a non-zero Q_T, the rank of P_T should be smaller than 4 such that its null space is not empty. If $rank(P_T) = 3$, we can uniquely obtain a normalized version $\bar{Q}_T$ of Q_T by setting $\bar{Q}_{1,4} = 1$. If $rank(P_T) < 3$, then there is no unique solution. As we will see in the next section, the singularity of Q will have to be measured to determine a final attention measure for an AOM. This singularity is based on the linear dependence of rows/columns of Q. Since linear dependence is invariant to scale, the normalized version $\bar{Q}_T$ is enough to check for the singularity of Q. The trivial solution $[0,0,0,0]^T$ is invalid because it corresponds to constant 0 of the last column of Q which is not allowable because that would make Q non-invertible and therefore, singular.

3.2. Attention measure from $\bar{Q}_T$ and $\mathcal{Q}$

Whether a group of feature points belongs to an AOM depends on whether there *exists* an *invertible* Q that satisfies the two constraints described in the previous section. We show that both the existence and singularity of Q can be analyzed from $\bar{Q}_T$ and $\mathcal{Q}$, *without calculating* Q. In fact, for degenerate cases such as pure translation and pure rotation, it is not possible to calculate Q.

3.2.1. Existence of Q from $\bar{Q}_T$

In the absence of noise, if the feature points belong to an AOM, then there exists a solution for the system of equations in (19) implying the existence of a $\bar{Q}_T$ that transforms $\hat{M}$ to a column consisting of two constant parts. However, it is not assured that the column will indeed be the last column of (16). If the column is one of the first 3 columns of Q, $\bar{Q}_T$ will be linearly dependent with one column of Q_R and, consequently, Q will be singular, in which case we need to check for the singularity of Q. In the case of noisy data, a solution $\bar{Q}_T$ of (19) always exists in the least square sense. Due to its sensitivity to noise, use of the smallest *SVD* value of P_T in (19) to measure the existence of $\bar{Q}_T$ is not robust.

To overcome this problem, we propose a measure to check for the existence of $\bar{Q}_T$ based on the variance of $\tilde{Q}_T = \hat{M} \times \bar{Q}_T$. This measure consists of two factors. The first factor measures the variances of the first F rows and the last F rows of $\tilde{Q}_T$, i.e.,

$$M_c(i) = var(\tilde{Q}_T^i), \quad i = 1, 2, \tag{20}$$

where $\tilde{Q}_T^1$ and $\tilde{Q}_T^2$ are the first and second halves of $\tilde{Q}_T$, respectively. The second factor is related to the randomness in the change of $\tilde{Q}_T$. The change can be caused due to unstable following of the object by the camera or due to the feature points moving in a particular direction not being followed. The former causes more randomness in the values of $\tilde{Q}_T$ than the latter; this trait is captured in the following measure

$$R_c(i) = \frac{\max(\tilde{Q}_T^i) - \min(\tilde{Q}_T^i)}{1 + \sum_{k=2}^{F} |\tilde{Q}_T^i(k) - \tilde{Q}_T^i(k-1)|}. \tag{21}$$

R_c is closer to 1 when the randomness of the variation in $\tilde{Q}_T$ is small. These two factors are combined to obtain the final measure S_q computed as

$$S_q = \max(R_c(1) \cdot M_c(1), R_c(2) \cdot M_c(2)). \tag{22}$$

If S_q is small, then $\tilde{Q}_T$ is more close to constant. Otherwise, it is more possible that $\tilde{Q}_T$ is not constant column.

3.2.2. Singularity of Q from $\bar{Q}_T$ and $\mathcal{Q}$

The singularity of Q can be indicated by the linear dependence in either its rows or its columns. In this section, we show that the linear dependency can be established by analyzing only $\bar{Q}_T$ and $\mathcal{Q}$.

In the case of linear dependence in rows or in columns of Q, we have

$$rank([Q_R|\bar{Q}_T]) = rank(Q) < 4. \tag{23}$$

Since $[Q_R|\bar{Q}_T]^T$ has the same rank as $[Q_R|\bar{Q}_T]$, it is also true that

$$rank([Q_R|\bar{Q}_T][Q_R|\bar{Q}_T]^T) = rank([Q_R|\bar{Q}_T]) < 4 \tag{24}$$

The matrix $[Q_R|\bar{Q}_T][Q_R|\bar{Q}_T]^T$ can be calculated directly from $\mathcal{Q}$ and $\bar{Q}_T$ without knowing Q, since

$$[Q_R|\bar{Q}_T][Q_R|\bar{Q}_T]^T = Q_R Q_R^T + \bar{Q}_T \bar{Q}_T^T = \begin{bmatrix} \mathcal{Q}_1 & \mathcal{Q}_5 & \mathcal{Q}_6 & \mathcal{Q}_7 \\ \mathcal{Q}_5 & \mathcal{Q}_2 & \mathcal{Q}_8 & \mathcal{Q}_9 \\ \mathcal{Q}_6 & \mathcal{Q}_8 & \mathcal{Q}_3 & \mathcal{Q}_{10} \\ \mathcal{Q}_7 & \mathcal{Q}_9 & \mathcal{Q}_{10} & \mathcal{Q}_4 \end{bmatrix} + \bar{Q}_T \bar{Q}_T^T \tag{25}$$

A measure of linear dependence of Q, denoted by L_q, can then be determined as $L_q = \sigma_{min}$ where σ_{min} is the smallest singular value of $[Q_R|\bar{Q}_T][Q_R|\bar{Q}_T]^T$. For noisy case, we improve the robustness of this measure by considering the ratio of the two smallest singular values.

3.2.3. Degenerate cases

Here, we analyze the existence and singularity of Q for degenerate cases, i.e., the case in which the rank of the measurement matrix is less than 4. Examples of degenerate motions are pure rotation and pure translation. Eq. (18) can be divided into 6 homogeneous linear systems in the following form:

$$\begin{bmatrix} Q_{i1}^2 + Q_{i2}^2 + Q_{i3}^2 \\ Q_{j1}^2 + Q_{j2}^2 + Q_{j3}^2 \\ Q_{i1}Q_{j1} + Q_{i2}Q_{j2} + Q_{i3}Q_{j3} \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix} \tag{26}$$

where $i \neq j \in \{1,2,3,4\}$ and C_1,C_2 and C_3 are corresponding entities in $\mathcal{Q}$, e.g., $\mathcal{Q}_1$, $\mathcal{Q}_2$ and $\mathcal{Q}_5$ form one such linear system. Each linear system relates 2 of the 4 rows of Q_R. For degenerate cases, only some of the rows of Q_R can be determined. The singularity of Q can still be indicated by the linear dependence among the available rows. We detect such linear dependence in degenerate cases by analyzing only $\bar{Q}_T$ and $\mathcal{Q}$. For example, if only 3 rows of Q_R can be determined for degenerate motion, then the rotation constraint on $\hat{M}$ gives

$$\begin{bmatrix} \mathcal{Q}_1 \\ \mathcal{Q}_2 \\ \mathcal{Q}_3 \\ \mathcal{Q}_4 \\ \mathcal{Q}_5 \\ \mathcal{Q}_6 \end{bmatrix} = \begin{bmatrix} Q_{11}^2 + Q_{12}^2 + Q_{13}^2 \\ Q_{21}^2 + Q_{22}^2 + Q_{23}^2 \\ Q_{31}^2 + Q_{32}^2 + Q_{33}^2 \\ Q_{11}Q_{21} + Q_{12}Q_{22} + Q_{13}Q_{23} \\ Q_{11}Q_{31} + Q_{12}Q_{32} + Q_{13}Q_{33} \\ Q_{21}Q_{31} + Q_{22}Q_{32} + Q_{23}Q_{33} \end{bmatrix} \tag{27}$$

If we denote the matrix containing 3 rows of Q_R as Q_A and the corresponding three entries of $\bar{Q}_T$ as $\bar{Q}_{TA}$, we have the following property for singular Q:

$$rank([Q_A|\bar{Q}_{TA}][Q_A|\bar{Q}_{TA}]^T) = rank([Q_A|\bar{Q}_{TA}]) < 3 \tag{28}$$

As before, the matrix $[Q_A|\bar{Q}_{TA}][Q_A|\bar{Q}_{TA}]^T$ can be calculated from available $\mathcal{Q}_i$ in (27) and the corresponding entries in $\bar{Q}_T$. So the linear dependence of Q can still be measured from the singular values of $[Q_A|\bar{Q}_{TA}][Q_A|\bar{Q}_{TA}]^T$.

3.3. Integrated attention measure

Finally, we combine both measures S_q and L_q to obtain an integrated measure to determine if a collection of feature points belongs to an AOM. We estimate the attention of a moving object as

$$Attention = L_q/S_q. \tag{29}$$

Since S_q measures the variance as well as the randomness of the last column of M, a small value of S_q indicates how constant the last column of M is. L_q is the smallest singular value of $[Q_R|\bar{Q}_T][Q_R|\bar{Q}_T]^T$ that indirectly indicates the singularity of Q. If L_q is close to 0, the matrix Q is singular (non-invertible). When L_q is large and S_q is small, the feature points are more likely to belong to AOM(s) and their attention values should be large. Hence, we use the ratio L_q/S_q as the measure of attention. If *Attention* is greater than a threshold, then the feature points whose coordinates form the measurement matrix W belong to an AOM. In this paper, the threshold value is decided by Otsu thresholding method [16] for all experiments. Other advanced measures can also be designed based on S_q and L_q. If there are multiple solutions for Q_T in (19) or for $Q_R Q_R^T$ in (17), then we apply the above analysis individually on X and Y. If there are still multiple solutions, then we can conclude that the feature points belong to an AOM.

4. Attention-from-motion for multiple AOMs

In the previous section, we have presented the theory for detecting a single AOM in the video sequence. Clearly, in some situations, there could be multiple moving objects and the measurement matrix would contain feature points from all of them. We develop an iterative framework to detect multiple AOMs, in which motion segmentation is performed. Since the proposed method to detect single AOM can handle partial data, we allow over-segmentation of the object, i.e., feature points can be belong to two or more groups. The proposed framework can robustly detect multiple AOM(s) for partial dependent motion [17] as well as for degenerate motion.

4.1. Shape space initialization

As noted in (14), the SVD of W is carried out as $W = U\Sigma V^T$. We denote the space formed by the basis of V as *shape space*. It encodes the information in the row space of the measurement matrix W. In [15], multibody segmentation is performed through the shape interaction matrix $A = VV^T$.

In our algorithm, initially, each feature point constitutes a group and the interaction of each point with all the other points are computed as follows. At each iteration, the feature point with the maximum interaction, computed as the sum of interaction with the points already in the group, is appended to the group. This procedure is continued till all the feature points are included in a group. The above process of expanding a group is performed on all the P groups. This means that every group will have all the feature points in decreasing order of interaction. Based on the ordering, we hope that points having strong interaction belong to the same object. But, such sorting could be erroneous since two points i and j might belong to the same object even if $A_{i,j} = 0$. However, it is very likely that the first 4 points on a majority of the sorted groups belong to the same object. The group is expanded further based on motion information (next section) to determine if there are other feature points belonging to this group. Instead of relying on the block-diagonal structure of canonical A for segmentation [15], which could degrade when there is even small amount of noise or when there is partial dependent motion [17], we only extract the first 4 points from every sorted list according to the rank theorem of dynamic scene.

Although we cannot determine whether two points with small interaction belong to same object or not, if there is large interaction between them, it is very likely that they belong to the same object. That is the reason why *majority* of the 4-tuple are correct. However, some of the 4-tuple may be still erroneous due to accidental interaction between points on different objects. A voting scheme will select the most reliable grouping with maximum support in the final stage, making the framework robust to such error in grouping.

4.2. Motion space expansion

For grouping feature points having similar motion, it is imperative to use motion information in addition to using shape information. Indeed, if two points have similar motion, they must be grouped together irrespective of their 3D coordinates. In this section, we expand the list of feature points previously obtained using shape information by adding points that have similar motion.

The similarity of motion among feature points is computed as follows. The matrix U obtained from the SVD decomposition of the measurement matrix W forms a basis for the column space of W, which is called the *motion space*. From the factorization analysis in Section 2.1, we know that the motion space of a moving object observed by a moving camera is of at most rank 4. Hence, we carry out the SVD on the measurement matrix W of the 4-tuple (obtained from the previous section) to determine the motion space of the corresponding object. We compute the projection of every column of W (there are P columns) onto the basis of this motion space (projection denoted as L) as well as orthogonal to it (projection denoted as H). The ratio $\frac{H}{L}$ indicates the similarity between the feature point for that column and the 4-tuple. If the two motions are similar, $\frac{H}{L}$ will be small; in the absence of noise, it will be 0 if the motions are exactly the same. The initial 4-tuple is expanded by appending to it, in increasing $\frac{H}{L}$ ratios, those feature points that satisfy the rank theorem of Section 2.1. The expanded list is then split into two, based on a thresholding mechanism such that the feature points on the top of the list are retained as candidates for voting. The threshold is chosen as that feature point which has the maximum difference of the ratio $\frac{H}{L}$ from its neighbor.

All the 4-tuples from the shape space initialization are expanded according to the above procedure. In order to ensure that, for each group, the measurement matrix corresponding to the expanded list of feature points is not rank deficient, we develop an iterative technique by which the list is expanded. Specifically, we repeat the same process as above starting from the SVD of the measurement matrix up to finding a new list of feature points after thresholding. If the new group is the same as the old one, iteration is stopped, otherwise a new measurement matrix for the new group is formed and the process is repeated.

In order to accommodate for possible errors in the groups, we assign a confidence measure to each of them. The confidence measure for a group G is defined as

$$C_G = \frac{\min_{i \notin G}(H(i)/L(i))}{\sum_{j \in G}(H(j)/L(j))}, \tag{30}$$

where $L(i)$ and $H(i)$ are the projections of the column in the measurement matrix W corresponding to feature point i, onto the motion space and orthogonal to it, respectively. When there is no clear difference among all $\frac{H}{L}$ ratios, it is not possible to obtain a good threshold. This could possibly lead to an erroneous group, whose confidence measure C_G will be small. Large confidence measure C_G indicates that the threshold can be easily found and the group is more likely to be correct.

By combining both shape space as well as motion space, the proposed method for motion segmentation is more robust to previous methods, which either consider shape space or motion space. Compared to the methods using shape space only [15,18], and the cluster-and-test approaches using motion space only [11,12], forming the 4-tuple from shape space and relying on motion space expansion, provides a systematic way to define the motion space, overcoming major disadvantages of previous methods. First, the 4-tuple with largest interaction values has more confidence to belong to the same object than random selection of a 4-tuple. The incorporation of confidence measure increases the accuracy of the motion space and allows errors in the shape interaction matrix. Second, the iterative motion space expansion and voting scheme (introduced in next section) can improve the robustness of motion segmentation when some groups of feature points are erroneous. Third, we iteratively extract the most reliable grouping one by one and avoid estimating the number of objects. Note that the degenerate motion space (e.g., low-rank subspace of motion space) does not affect the proposed method since we allow for over-segmentation of the object. Interestingly, the case of partially dependent motion as enunciated in [17] is also handled efficiently using the proposed method. For two objects o_1 and o_2 having partially dependent motion, the trajectory of o_1 has only *some* components lying on the motion space of o_2 so that the orthogonal component H for o_1 will be larger than that for o_2. It follows that the above process will separate the two objects.

4.3. Iterative AOM extraction by confidence voting

At this point, we have P candidate groups and their associated confidence measures from which only the most reliable group is selected based on a voting scheme on the confidence measures. We first refine these candidate groups by combining those groups that contain any common feature points. Beginning with the first candidate group as the only existing group, we iteratively add a new candidate group G_{new}. Specifically, if G_{new} contains any common feature point with some existing group G_{exist}, we combine it with G_{exist}, i.e.,

$$G_{\text{exist}} = G_{\text{exist}} \cup G_{\text{new}} \tag{31}$$

and update the corresponding confidence measure as

$$C_{G_{\text{exist}}} = C_{G_{\text{exist}}} + \frac{|G_{\text{exist}} \cap G_{\text{new}}|}{|G_{\text{new}}|} \cdot C_{G_{\text{new}}} \tag{32}$$

where $\cup$ denotes the union operator, $\cap$ denotes the intersection operator and $|\cdot|$ indicates the size of the set. If G_{new} does not overlap with any existing group, it is formed as a new group with the associated confidence measure. After combination, the group with the maximum confidence measure is chosen as the most reliable motion group. This group is analyzed using the method described in Section 3 for a single AOM and a value of *Attention* is assigned to it. The measurement matrix W is updated by removing the columns corresponding to the points in the most reliable group and the entire process is repeated. The complete algorithm is summarized in Algorithm 1.

Algorithm 1. Recursive Multiple AOM(s) Detection **Require:** measurement matrix W

While W is not empty **do**
  Calculate A from SVD of W (Eq. (14));
  for each feature point i
    Obtain its 4-tuple from shape space (Section 4.1);
    Expand the 4-tuple from motion space to obtain the candidate group G_i (Section 4.2);
    for each existing candidate group G_{exist} **do**
      if G_i overlaps with G_{exist} **then**
        update G_{exist} as well as $C_{G_{\text{exist}}}$ according to Eq. (32);
      else
        add G_i as a new existing group and associate its confidence measure as C_{G_i};
      end if
    end for
  end for
  Calculate *Attention* for the group with maximum confidence, say G_{CMax}, which is one of the existing groups with the maximum confidence measure *CMax* (Eq. (29)); Update W by removing the corresponding columns of those feature points in G_{CMax};
end while
Threshold *Attention* to detect AOM(s)

The reliability of a group depends not only on the number of instances of that group but also on their confidence measures. The confidence measure of every instance of a group may be different because of the difference in the initial motion space. The proposed scoring mechanism takes both these factors into account. Moreover, once the most reliable motion group is obtained, it is removed from further analysis so that the remaining groups can be determined more confidently one by one.

The failure to detect feature points occurs due to two reasons: occlusion (or out-of-view) and errors in tracking. Fig. 5 shows an example in which both these phenomena occur. They introduce two types of errors in the measurement matrix W either an entry is missing due to occlusion or the entry is wrong due to tracking error. In the former case, since the proposed algorithm can handle partial data, we simply remove all columns in W corresponding to the occluded points even if they were available in other frames. In the case of missing data, if feature points are missing in a majority of the frames, only the available data is analyzed without degrading performance. On the other hand, if feature points are missing in only few frames, SVD in the above procedure can be replaced by the PowerFactorization method [19] to utilize the additional information that is available. However, the PowerFactorization method requires the rank of the matrix as the prior knowledge, which is not available when there are multiple objects. Hence, for the entire sequence, we remove all the feature points that have been occluded. For example, the points within the green ellipse of Fig. 5 are removed from W. Since the proposed algorithm can handle partial data, it can still detect an AOM if, despite occlusion, sufficient number of points on it is available. The proposed algorithm can detect AOM in a very short interval because of its large discriminating power as well as the ability to handle partial period of motion. Hence, for occlusion changing over time, we can detect AOM without occlusion in some interval, then detect AOM when the object is partially occluded and only miss it when it is fully occluded. However, when the object re-appears and there are sufficient feature points on it, the analysis resumes. In the case of tracking error, the proposed algorithm will be inevitably affected by the failure of feature point detection/tracking (e.g., the tracking error in red ellipse of Fig. 5). For moderate error in W, the proposed method

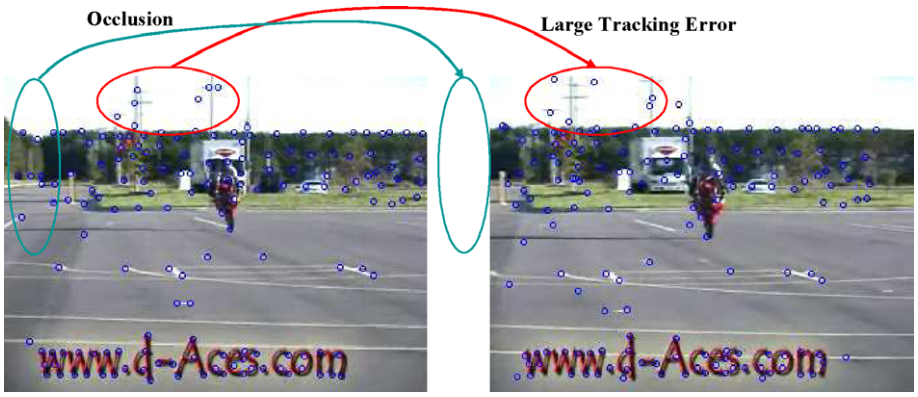

Fig. 5. An example of the failure to detect feature points (indicated by small blue circles). There are two types of errors introduced: 1, out-of-view/occlusion error (area indicated by green ellipse); 2, tracking error (area indicated by red ellipse). (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

can still correctly detect AOM due to its large discriminant power. But the performance will degrade when the output of tracking algorithm is too erroneous to correctly reflect motion information. The large discriminant power and the robustness to error in W of the proposed algorithm will be evaluated in the experiment section.

5. Experimental results

The performance of the proposed method for attention-from-motion problem was tested on synthetic and real image sequences. The algorithm was tested to evaluate its ability to handle arbitrary number of objects, different types of motion and the case of partial data.

5.1. Synthetic data

We consider image sequences containing synthetic objects like cube, sphere and pyramid and a static background. We choose 7 feature points from the pyramid, 12 from the cube, 30 from the sphere and 20 from the background. The object motions are synthesized and the camera motion is synthesized to follow one or more objects. The 2D image coordinates of the feature points in each frame are calculated with round-off errors to form the measurement matrix W. The synthetic sequences incorporate a combination of the following characteristics—(i) different number of AOMs, (ii) different object motion and camera motion including degenerate and dependent cases, (iii) AOM with all feature points and, (iv) AOM with partial feature points.

The top row in Fig. 6(a) shows a sequence containing a cube and a pyramid as well as some background points in which the cube undergoes a 3D motion including both rotation and translation. The camera motion is also synthesized to undergo a different 3D motion but to follow the cube. Hence, from our definition of AOM, the cube has been detected as AOM and marked in red in the bottom row. Since the pyramid undergoes different motion so that the camera does not follow it, it is not an AOM. In Fig. 6 (b), the two objects undergo different 3D motions–rotation and translation for the cube and translation only for the truncated pyramid. Both objects are followed by a moving camera. We can see the proposed algorithm detected both of them as AOMs and they are marked in red in the bottom row. The sequence in Fig. 6(c) consists of a sphere in addition to the cube and pyramid. However, only half of the sphere is observable and only feature points from this half are included in W. In the sequence, the camera undergoes *only* translation, i.e., this is a case of degenerate camera motion. The pyramid and the sphere undergo different rotations and the same translation as the camera, while the cube has a different translation. The bottom row shows that the two AOMs—sphere and pyramid—are correctly identified in red.

We also test the robustness of the proposed algorithm to errors in the measurement matrix W. Gaussian noise with standard deviation varying from 0.5 to 10 was added to W of the synthetic data to introduce the errors in the trajectories of feature points. Such noise perturbs the image coordinates of feature points. Table 1 shows the average offset as well as the maximum offset in the image coordinates of the feature points. To measure the performance of AOM detection, we calculate the precision and recall of the detection based on 10 trials at each noise level. Here, the precision is defined as the ratio between the number of correctly detected AOM points and the number of total detected points. Similarly, the recall is defined as the ratio between the number of correctly detected AOM points and the number of all AOM points. Each trial generates a new synthetic sequence where motions are different and the number of AOMs is different. As shown in Fig. 7, high precision and recall rates are obtained for small noise level indicating that the proposed algorithm is robust to moderate errors in W. When the noise level is very high, the performance of the proposed algorithm degrades because of the incorrect motion information of W. Even so, we note that at a noise level of 10 when the average offset is about 2 pixels and the maximum error is about 11 pixels, the precision is still very high and the recall is reasonably good at about 77%. We pushed the noise level further to standard deviations of 25 and 30 to obtain a recall of 38% and 12%, respectively, while maintaining a precision of close to 1. This shows that the proposed algorithm never identifies an incorrect AOM, but will miss a true AOM when noise in the measurement matrix is very high, resulting in erroneous motion information.

5.2. Real image sequences

Next, we apply the proposed AOM detection method on real video sequences shown in Figs. 8 and 9. The top row shows the feature points used to form W, while the second row shows feature points belonging only to the AOMs (different AOMs are indicated by different colors). The feature points are detected as follows: The corner points in the first frame are detected and tracked in the subsequent frames using [20]. Only the feature points which are correctly tracked are used to form W. For some sequence, the AOM(s) do not have enough corner points, we manually label minimum points to cover them. Fig. 8 shows the results of AOM detection on three sequences. In Fig. 8(a), the top row contains 70

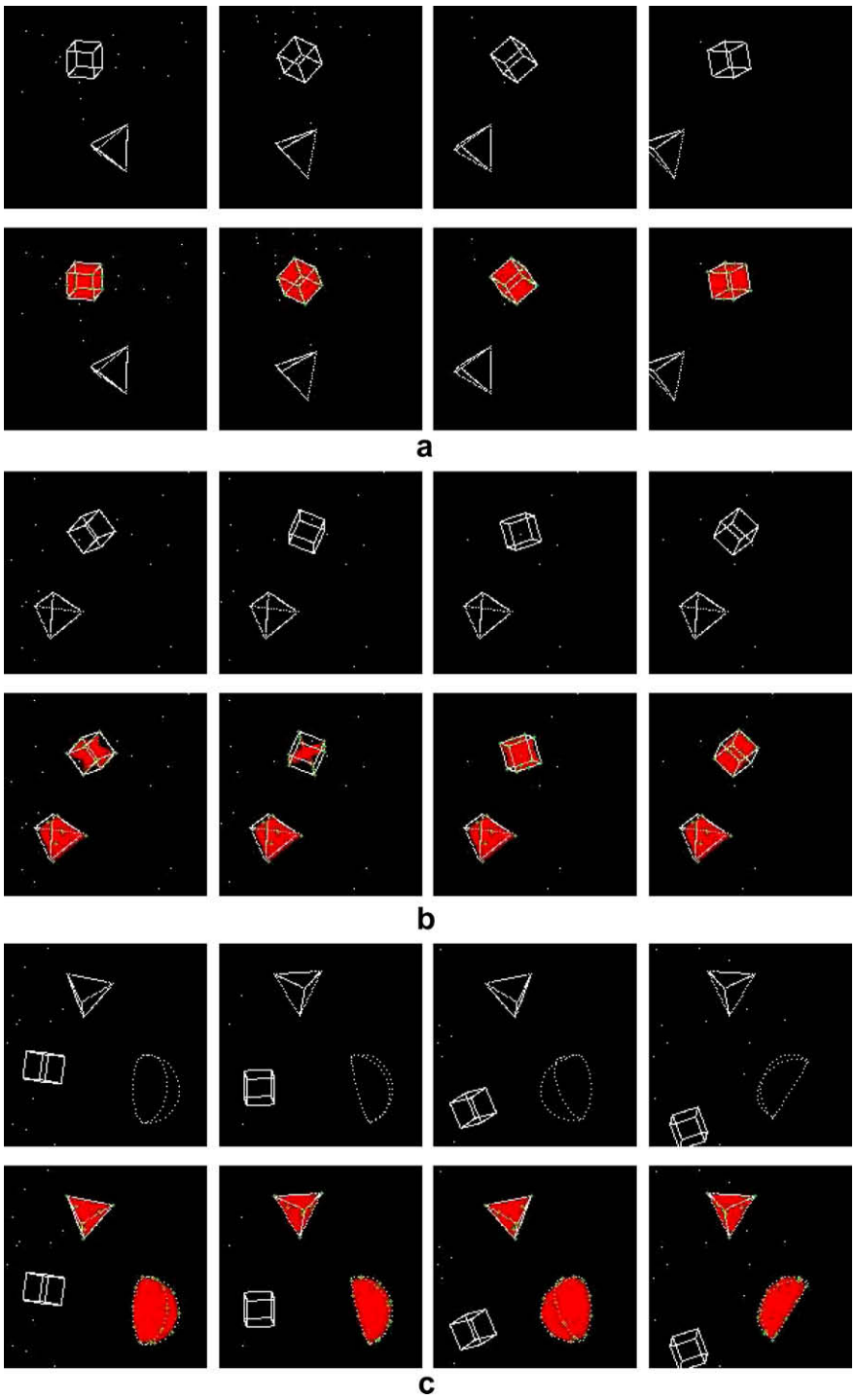

Fig. 6. (a–c) *Top row:* Objects in synthetic image sequences. *Bottom row:* Detected AOMs in red. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

Table 1
Average and maximum error of image coordinates in W under different noise level

Std dev. of noise	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
Ave offset (pixel)	0.80	1.12	1.36	1.59	1.78	1.95	2.13	2.24	2.36	2.53
Max offset (pixel)	3.62	4.89	6.00	6.78	8.29	8.77	9.79	10.07	10.34	11.03

feature points. The bottom row shows that the moving car as well as the time stamp have been correctly identified as AOM, while the points in the background have been removed. The yellow car undergoes a full 3D motion and is detected as an AOM. The time stamp is also identified as another AOM since it is fixed on the screen. The sequence in Fig. 8(b) contains two moving cars of

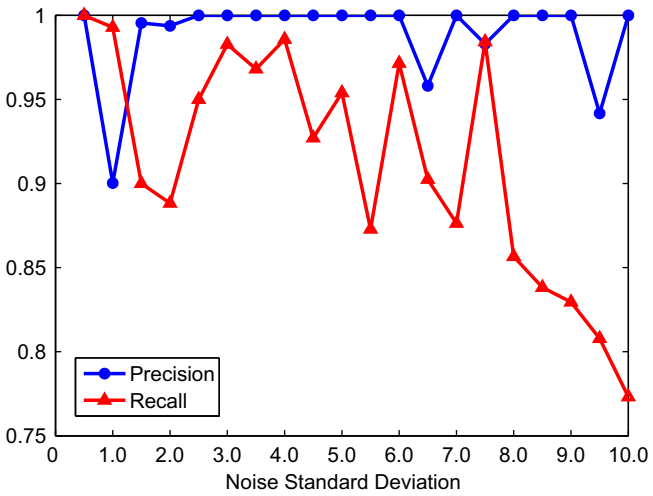

Fig. 7. Precision and recall of feature points of AOM(s) w.r.t noise level.

which only one is followed by the camera. This is also a challenging sequence since the cluttered background results in numerous tracking errors resulting in a very noisy measurement matrix W. As the bottom row shows, the tracked moving car has been identified as the AOM. In Fig. 8(c), the crane in the foreground and the duck in the background are being followed by the camera and the algorithm correctly identifies them as AOMs. Note that the duck gets partially occluded by the crane so that there is missing data in the measurement matrix. However, the algorithm succeeds in identifying the AOMs, without explicitly interpolating for the missing data. Fig. 9 shows the results of AOM detection on three more sequences. The sequence in Fig. 9(a) is the "mobile-and-calendar" in which there are 3 objects—ball, train and calendar—and a stationary background. The ball and train are followed by the camera and hence they are the two AOMs. Although the ball and train are close to each other and their motions are degenerate and partially dependent, the proposed method succeeds in separating them and identifying both of them as AOMs from a small number of points, e.g., only 6 on ball. The sequence shown in Fig. 9(b) is a cluttered one with many persons. Fifty-eight feature points par-

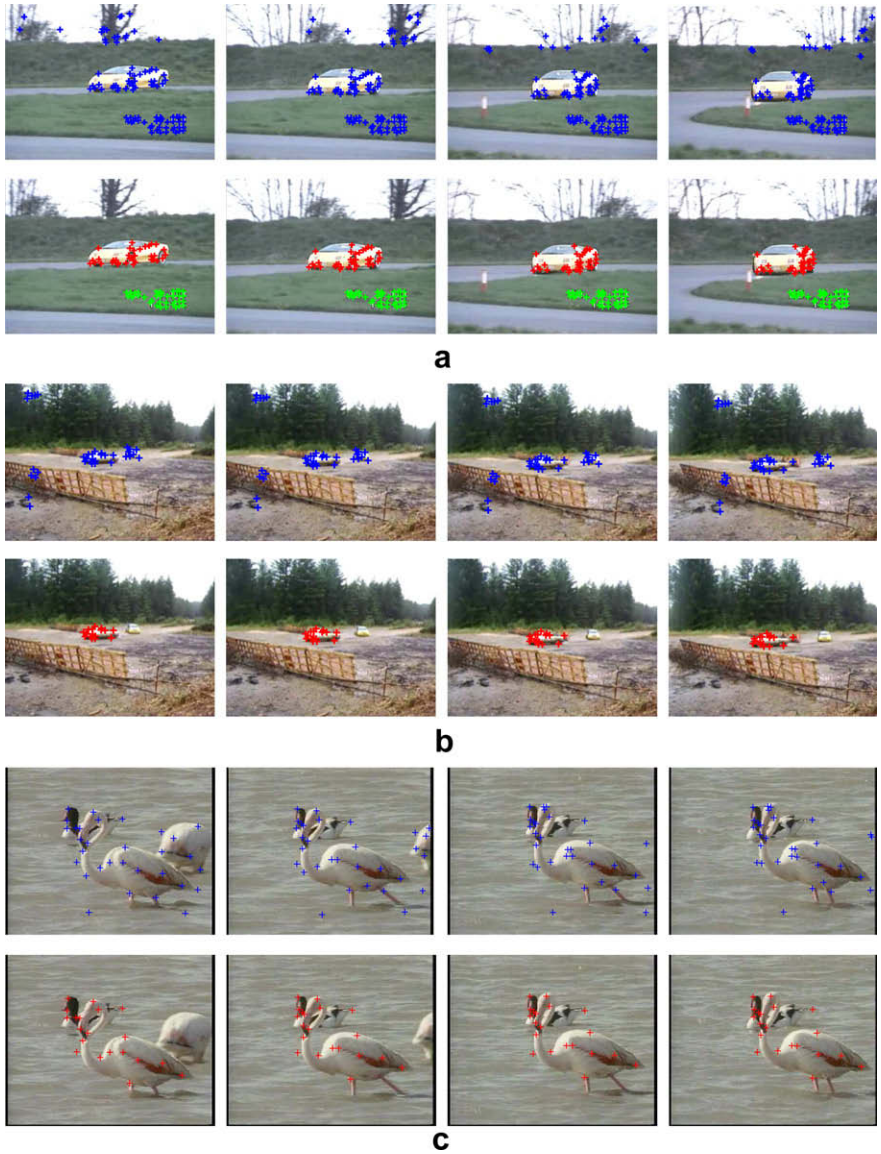

Fig. 8. *Top row:* All feature points. *Bottom row:* Detected AOMs for (a) Seq 1, (b) Seq 2, (c) Seq 3. In the top row, the number of feature points *on* an AOM need not be larger than that outside the AOM (figure best viewed in color). (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

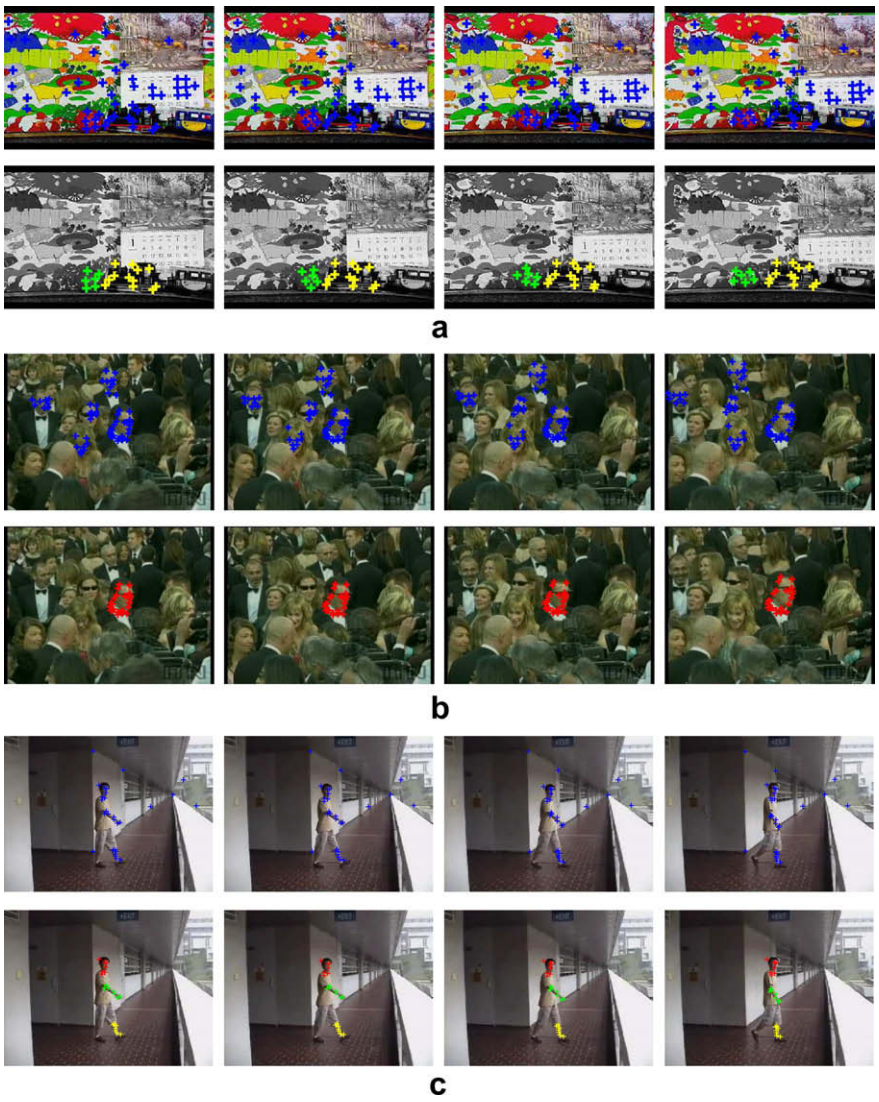

Fig. 9. *Top row:* All feature points. *Bottom row:* Detected AOMs for (a) Seq 4, (b) Seq 5 and (c) Seq 6. In the top row, the number of feature points *on* an AOM need not be larger than that outside the AOM (figure best viewed in color). (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

Table 2
Comparison with simply checking for stationary centroid

Seq 4 (Fig. 9(a))			Seq 5 (Fig. 9(b))			Seq 6 (Fig. 9(c))		
object	S_c	L_q/S_q	object	S_c	L_q/S_q	object	S_c	L_q/S_q
Ball	**6.49**	**383.64**	Person 1	28.38	0.24	**Head**	**5.73**	**5.26**
Train	**1.59**	**103.57**	Person 2	25.35	0.004	**Arm**	**13.65**	**5.40**
Calendar	10.80	0.26	Person 3	22.20	0.04	**Leg**	**13.58**	**2.93**
Bg	11.71	4.28	**Person 4**	**8.06**	**3.14**	Bg	18.48	0.04
/	/	/	Person 5	9.44	0.37	/	/	/

We calculate the proposed attention measure L_q/S_q and S_c of (33) on the objects for the 3 sequences in Fig. 9. The ground truth AOM(s) are highlighted in **bold**.

tially cover five persons. The proposed method again succeeds to separate them and identify the person followed by the camera. Although the assumption in this paper was of rigid body motion, we also experimented the proposed method on articulated motion. For the walking sideview sequence in Fig. 9(c), the centroids of the arm and the leg and also the centroids of their motion are not stationary. Here the joints of the shoulder and the knee are followed by camera and the rotation motions around them satisfy the special structure of the motion matrix M. The algorithm returned three AOMs corresponding to the three "rigid body motions" of the head (red), an arm (green) and a leg (yellow). All three AOMs cover the whole walking person followed by the camera.

To compare the proposed AOM detection method with the simple check for stationary centroid, we compare the proposed measure L_q/S_q and the following measure for the stationary of object centroid.

$$S_c(X,Y) = std(X) + std(Y) \tag{33}$$

where *std* denotes the standard deviation. From Table 2, we can see the proposed method is much more discriminant than the simple

Fig. 10. Detection of AOM with partial occlusion over a very short interval (6 frames). First row shows 3 frames with all feature points and second row shows the detected feature points belonging to AOM.

check for stationary of centroid. Although the S_c of AOM(s) are smaller than that of other objects which are not AOM, the fact that there is no optimal threshold to separate AOM(s) from others will degrade the AOM detection. For example, only the train will be detected as an AOM while the ball will be missed for Seq 4. Similarly, both arm and leg part will be missed in Seq 6. For Seq 5, another person (person 5—the lady in left bottom part) which is not an AOM will be wrongly detected as an AOM. However, it is clear that the proposed method can well separate AOM from others according to the attention values of L_q/S_q for all these 3 sequences.

We also conduct an experiment to evaluate the ability of the proposed method to detect AOM in a very short time interval and to handle partial occlusion that changes over time. In Fig. 10, we show 3 frames from a short sequence of *only* 6 frames where the body of the monkey (which is the AOM) is partially occluded. The occlusion changes over time. Due to the discriminating power of the algorithm, the AOM can be detected over such very small intervals. From the results, we can see that the proposed algorithm can detect the feature points on the monkey's head as part of the AOM, while ignoring those in the occluded parts and in the background.

6. Conclusion

In this paper, we have introduced the concept of attention object in motion, which is the object that the cameraman wishes to focus on. Then we define the problem of attention-from-motion that extracts AOMs from an image sequence. We propose the factorization of measurement matrix to describe a dynamic scene where a moving camera observes a moving object. Based on this factorization, we analyze the special structure in the motion matrix of single AOM and propose a method to estimate the existence of such structure without complete factorization. Finally, we describe an iterative framework to detect multiple AOMs by integrating shape space and motion space as well as voting scheme.

The proposed algorithm provides a robust framework for identifying multiple AOMs. The analysis of structure in the motion matrix of an AOM from factorization enables the detection of AOM(s) when only partial data is available due to occlusion and over-segmentation, a partial observation of complete object motion, or a special following pattern where the centroid of object is not stationary. Without recovering the motion of either object or camera, the proposed method can detect AOM robustly from any combination of camera motion and object motion, even for degenerate motion. For multiple AOMs, partial groups of individual objects can be robustly extracted. This framework can be extended to the entire family of affine camera models.

References

[1] R.S. Zemel, T.J. Sejnowski, A model for encoding multiple object motions and self-motion in area MST of primate visual cortex, Journal of Neuroscience 18 (1) (1998) 531–547.
[2] C. Pack, S. Grossberg, E. Mingolla, A neural model of smooth pursuit control and motion perception by cortical area MST, Journal of Cognitive Neuroscience 13 (1) (2001) 102–120.
[3] J.K. Tsotsos, Y. Liu, J.C. Martinez-Trujillo, M. Pomplun, E. Simine, K. Zhou, Attention to visual motion, Computer Vision and Image Understanding 100 (1–2) (2005) 3–40.
[4] Y.-F. Ma, X.-S. Hua, L. Lu, H.-J. Zhang, A generic framework of user attention model and its application in video summarization, IEEE Transactions on Multimedia 7 (5) (2005) 907–919.
[5] X.-S. Hua, L. Lu, H.-J. Zhang, AVE-automated home video editing, in: Proceedings of the Eleventh ACM International Conference on Multimedia, ACM Press, New York, NY, USA, 2003, pp. 490–497.
[6] T.J. Williams, B.A. Draper, An evaluation of motion in artificial selective attention, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPRW'05).
[7] Y.-L. Tian, A. Hampapur, Robust salient motion detection with complex background for real-time video surveillance, in: Proceedings of IEEE Computer Society Workshop on Motion and Video Computing, vol. 2, Breckenridge, Colorado, USA, 2005, pp. 30–35.
[8] L. Wixson, Detecting salient motion by accumulating directionally-consistent flow, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 774–780.
[9] R.P. Horaud, D. Knossow, M. Michaelis, Camera cooperation for achieving visual attention, Machine Vision and Applications 16 (6) (2006) 331–342.
[10] M.T. López, M.A. Fernández, A. Fernández-Caballero, J. Mira, A.E. Delgado, Dynamic visual attention model in image sequence, Image and Vision Computing 25 (5) (2007) 597–613.
[11] T.E. Boult, L.G. Brown, Factorization-based segmentation of motions, in: Proceedings of the IEEE Workshop on Visual Motion, Princeton, NJ, 1991, pp. 179–186.
[12] C.W. Gear, Multibody grouping from motion images, International Journal of Computer Vision 29 (2) (1998) 133–150.
[13] D. Bordwell, K. Thompson, Film Art: An Introduction, seventh ed., McGraw-Hill, 2003.
[14] C. Tomasi, T. Kanade, Shape and motion from image streams under orthography: a factorization method, International Journal of Computer Vision 9 (2) (1992) 137–154.
[15] J. a'o PauloCosteira, T. Kanade, A multibody factorization method for independently moving objects, International Journal of Computer Vision 29 (3) (1998) 159–179.
[16] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1) (1979) 62–66.

[17] L. Zelnik-Manor, M. Machline, M. Irani, Multi-body factorization with uncertainty: revisiting motion consistency, International Journal of Computer Vision 68 (1) (2006) 27–41.
[18] N. Ichimura, A robust and efficient motion segmentation based on orthogonal projection matrix of shape space, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, Hilton Head Island, SC, 2000, pp. 446–452.
[19] R. Hartley, F. Schaffalitzky, Powerfactorization: 3D reconstruction with missing or uncertain data, in: Proceedings of Australia–Japan Advanced Workshop on Computer Vision, Adelaide, Australia, 2003.
[20] B. Georgescu, P. Meer, Point matching under large image deformations and illumination changes, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (6) (2004) 674–688.