# Detection of Independently Moving Objects in Non-planar Scenes via Multi-Frame Monocular Epipolar Constraint

Soumyabrata Dey, Vladimir Reilly, Imran Saleemi, Mubarak Shah

Computer Vision Lab, University of Central Florida, Orlando, USA.
{sdey, vsreilly, imran, shah}@eecs.ucf.edu

**Abstract.** In this paper we present a novel approach for detection of independently moving foreground objects in non-planar scenes captured by a moving camera. We avoid the traditional assumptions that the stationary background of the scene is planar, or that it can be approximated by dominant single or multiple planes, or that the camera used to capture the video is orthographic. Instead we utilize a multiframe monocular epipolar constraint of camera motion derived for monocular moving cameras defined by an evolving epipolar plane between the moving camera center and 3D scene points. This constraint is parameterized as a polynomial function of time, and unlike repeated computations of inter-frame fundamental matrix, requires the estimation of fewer unknowns, and provides a more consistent separation between moving and static objects for different levels of noise. This constraint allows us to segment out moving objects in a general 3D scene where other approaches fail because their initial assumptions do not hold, and provides a natural way of fusing temporal information across multiple frames. We use a combination of optical flow and particle advection to capture all motion in the video across a number of frames, in the form of particle trajectories. We then apply the derived multi-frame epipolar constraint to these trajectories to determine which trajectories violate it, thus segmenting out the independently moving objects. We show superior results on a number of moving camera sequences observing non-planar scenes, where other methods fail.

## 1 Introduction

Moving object detection is a crucial stage in any automated surveillance system. If the camera is stationary, a popular framework for tackling the problem is to generate a model of the background of the static scene, and to treat moving objects as outliers to that model, as in [1–8]. The models reflect a diverse range of sophistication depending on problem domain and complexity of the scene. These methods can be extended in a straight forward manner to the moving camera case by performing global motion compensation prior to the generation of the background model, which is then utilized in the same manner as in the static case on the registered frames, [9, 10].

The above approach suffers from the severe limitation that the algorithms employed to compensate the motion of the camera, utilize an affine or homographic framework to model the transformations between the images. It is well known, that the homography is a limited image transformation model, which in the case of general camera motion is

**Fig. 1.** Illustration of drawbacks of homography based detection approach, where of out of plane objects (water tower) are incorrectly detected as moving objects. Left image shows original frames, and right image overlays red detection masks over the images. Detection is performed by image registration followed by background subtraction.

only valid when the observed scene is planar, or the camera is orthographic. If the scene is non-planar, and the camera is perspective, then the homography is valid if the camera motion is limited to rotation only. The homography is not valid when the camera undergoes translational motion while observing a non-planar scene. Under these circumstances, homography will be valid only for some areas of the images, the areas for which the homography is not valid will *appear* to "move" in the motion compensated imagery, and will be falsely detected as moving objects (figure 1).

The purpose of our paper is to present a novel method for segmenting moving objects, in a video captured by a generally moving camera, observing a non-planar 3D scene. We avoid performing homography based motion compensation, or background subtraction, and instead capture all of the motion in the video by computing optical flow, and then performing particle advection on that flow for a number of frames. The resulting set of particle trajectories contains two subsets. In the first subset the motion of the particles was induced purely by the motion of the camera (this is a set of particles belonging to stationary background in the world). In the second subset, the trajectories combine the motion of the camera, as well as the motion of independently moving objects. We separate the two sets using the constraint, defined by the proposed Multiframe Monocular Fundamental Matrix (MMFM).

We define this constraint as the evolution of the camera model with time, relative to some initial camera position. The evolving camera model defines an evolving epipolar plane between the initial center of the camera, its subsequent centers, and the static scene point. This multi-frame epipolar constraint can then be expressed as a dynamically changing fundamental matrix. When we assume that the evolution of camera parameters can be represented by polynomial functions of time, this monocular multiframe fundamental matrix can then also be represented as a polynomial function under assumption that the inter-frame camera rotation is small. Thus, we can obtain the coefficients of this matrix for frame segments of length $N$ and use them to determine whether particle advection trajectories belong to moving or static objects.

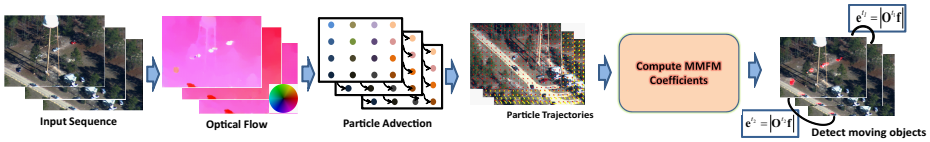


**Fig. 2.** A flow chart illustrating the different steps of the proposed framework.

## 2 Related Work

A number of techniques have previously addressed the problem of moving object detection under conditions of non-planar scene, and general camera motion. One approach is to assume that even though the scene itself is non-planar, it is nevertheless dominated by a ground plane, which has several out of plane objects upon it. This is known as the plane plus parallax model, and was utilized in [11–13]. When these assumptions hold, and a robust registration method is utilized, a valid homography can be estimated for pixels belonging to the ground plane. Since the homography does not hold for the out of plane objects, they too are detected as 'movers' during background subtraction. Finally detections are filtered as either moving or stationary out of plane by applying two or three view geometric constraints. The issue with the above framework is that while it is more general than the initial planar scene assumption, it is not sufficiently general to deal with scenes that do not have a dominant plane, or scenes that are so cluttered with out of plane objects as to prevent the correct estimation of homography. Additionally the systems themselves are rather complex and involve a large number of steps.

In the case of video taken from an unmanned aerial vehicle (UAV), even more elaborate systems have been proposed. In [14], out of plane objects are assumed to be buildings and trees, and are explicitly detected and segmented in every frame using wavelet features and Bayesian segmentation. In [15], the 3D nature of the scene is dealt with, by exploiting metadata associated with the imagery, to aid in stereo-like 3D reconstruction of buildings present in the scene, as well as color, texture, and depth based classification to detect vegetation. The 3D knowledge is exploited to suppress false detections that appear on out of plane objects. The limitation here of course, is that this method is not really applicable outside of aerial surveillance, and the metadata may not always be available or accurate, and has to be refined via a complex geo-registration process. Pollard and Mundy [16], essentially extend probabilistic modeling of background, to the $3^{rd}$ dimension. Rather than constructing probability distributions of background appearance on a per pixel or per block level, they use metadata to reconstruct the 3D structure of the scene as a series of probabilistic models of color within 3D *voxels*. Moving objects are then detected by comparing their appearance against the probabilistic models of the voxels of the static scene. While the work is original, the results are unfortunately not very good, which is explained by issues with metadata errors, and the selection of proper voxel granularity. Also in general we believe that full 3D reconstruction for object detection is overkill and should be avoided since it is a rather complex problem in and of itself.

Another body of work exists on segmenting video into layers of motion, as seen in [17–21]. This framework essentially attempts to segment the scene into regions that exhibit similar planar motion within themselves, which occurs due to different planes in the scene, objects with high parallax, and moving objects. This approach, however is non-trivial, since it requires some form of clustering of local planar transformations in the video into similar groups. Which brings up a myriad of issues typically associated with clustering, such as the space in which the data lies, the manual selection of the number clusters (which implies knowing the number of layers that will be extracted from the video a-priori), or the tuning of parameters, if the clustering method determines the number of clusters automatically. On top of this, once the layers are generated one has to determine which layers actually belong to moving objects, and not static scene parts.

Recently [22] used a completely different approach to segment foreground moving objects from background. The main idea of the work is that the motion of trajectories of the static 3d points in the image frames is caused by camera motion only. Under an orthographic camera assumption, the observation matrix composed of the point trajectories has rank 3, and therefore, any static trajectory can be expressed by a linear combination of 3 other trajectories. The authors exploit this fact to construct a projective subspace of background trajectories in a RANSAC framework. Trajectories corresponding to moving objects do not belong to the subspace, and can be detected by computing the error by projecting them upon the subspace. The main limitation here is the orthographic camera assumption, where performance of the method degrades as the amount of depth variation in the scene increases, and the velocity of the objects decreases.

In contrast to these techniques, our framework avoids assumptions of orthographic camera or dominating planes, and solves the problem of object detection under moving camera in general 3D scenes. We show that translating camera centers form a dynamically evolving epipolar plane with a static 3D point, and by projecting particle trajectories onto this plane we can detect moving objects. Using the rigorous experiments with the synthetic data, we justify the use of the dynamically evolving epipolar plane, because it provides a more consistent separation between moving and static objects than the standard fundamental matrix. This is detailed in the following sections.

## 3 Derivation of a Multiframe Monocular Epipolar Constraint

We begin by deriving our multiframe epipolar constraint for a *single* moving camera. Assume that a point $\mathbf{X}$, is a 3D point in the world, and point $\mathbf{x}$, is a point in the camera coordinate system, which in the case of a stationary camera is obtained as,

$$\mathbf{x} = \mathbf{R}\mathbf{X} + \mathbf{T}, \tag{1}$$

where $\mathbf{R}$ is the rotation of the camera, and $\mathbf{T}$ is the translation. If the camera is moving in a static scene, then the rotation $\mathbf{R}$ and translation $\mathbf{T}$ will be different at different time instances. Hence the camera coordinates of the world point $\mathbf{X}$, at a time $t$ is given by,

$$\mathbf{x}(t) = \mathbf{R}(t)\mathbf{X} + \mathbf{T}(t). \tag{2}$$

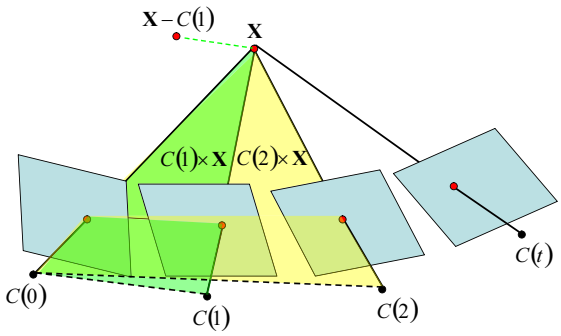


**Fig. 3.** This figure shows a single moving camera, and the epipolar planes that are formed between its initial center $C(0)$, centers at time $t$ $C(t)$ and a static point $\mathbf{X}$.

If the center of the camera at time $t$ is written as $\mathbf{C}(t)$, following the coplanarity constraint, we can write,

$$\Big(\mathbf{X} - \mathbf{C}(t)\Big)^\top \Big(\mathbf{C}(t) - \mathbf{C}(0)\Big) \times \Big(\mathbf{X} - \mathbf{C}(0)\Big) = 0, \tag{3}$$

and form a plane at each time instant, by taking the cross product between vectors $\mathbf{C}(t) - \mathbf{C}(0)$, and $\mathbf{X} - \mathbf{C}(0)$. If the camera center in the first frame is chosen to be the origin, i.e., $\mathbf{C}(0) = [0 \quad 0 \quad 0 \quad 1]^\top$, then equation 3 transforms to,

$$\Big(\mathbf{X} - \mathbf{C}(t)\Big)^\top \mathbf{C}(t) \times \mathbf{X} = 0. \tag{4}$$

Figure 3 illustrates the above equation for two planes in a monocular moving camera sequence, and the plane at $t = 1$ is shown in green, and the one at $t = 2$ in yellow. The point $\mathbf{X} - \mathbf{C}(t)$ is simply the world point $\mathbf{X}$ shifted along the plane. We can use equation 2, to express world point $\mathbf{X}$ in terms of it's camera coordinates $\mathbf{x}(t)$ as,

$$\mathbf{X} = \mathbf{R}(t)^\top \Big(\mathbf{x}(t) - \mathbf{T}(t)\Big) \tag{5}$$

Combining equations 4 and 5, and the fact that the first frame of the camera is the reference, i.e., $\mathbf{x}(0) = \mathbf{X}$, we obtain,

$$\Big(\mathbf{R}(t)^\top \left(\mathbf{x}(t) - \mathbf{T}(t)\right) - \mathbf{C}(t)\Big)^\top \mathbf{C}(t) \times \mathbf{x}(0) = 0, \tag{6}$$

and since $-\mathbf{R}(t)^\top \mathbf{T}(t) = \mathbf{C}(t)$, the equation can be simplified as follows:

$$\begin{aligned} \Big(\mathbf{x}(t)^\top \mathbf{R}(t)\Big)\mathbf{C}(t) \times \mathbf{x}(0) &= 0 \\ \Big(\mathbf{x}(t)^\top \mathbf{R}(t)\Big)\Big(-\mathbf{R}^\top(t)\mathbf{T}(t)\Big) \times \mathbf{x}(0) &= 0 \\ \Big(\mathbf{x}(t)^\top \mathbf{R}(t)\Big)\mathbf{S}(t)\mathbf{x}(0) &= 0, \end{aligned} \tag{7}$$

where $\mathbf{S}(t)$ is the skew-symmetric matrix of vector $-\mathbf{R}^\top(t)\mathbf{T}(t)$. The multiframe Essential matrix for the single moving camera is therefore obtained as,

$$\mathbf{x}(t)^\top \mathbf{E}(t)\mathbf{x}(0) = 0. \tag{8}$$

Finally assuming that the focal length of the camera remains constant during the capture, we derive the Multiframe Monocular Fundamental Matrix (MMFM) constraint for a moving camera as,

$$\begin{aligned} \mathbf{x}(t)^\top \mathbf{K}^{-\top}\mathbf{E}(t)\mathbf{K}^{-1}\mathbf{x}(0) = 0 \\ \mathbf{x}'(t)^\top \mathbf{F}(t)\mathbf{x}'(0) = 0, \end{aligned} \tag{9}$$

where $\mathbf{x}'(t) = [x(t) \quad y(t)]^\top$ are image coordinates of the world point, at time $t$.

We define the rotational and translational velocities of the camera at each time instance as $\mathbf{\Omega}(t)$, and $\mathbf{\Theta}(t)$, respectively.

$$\mathbf{\Omega}(t) = \begin{bmatrix} 1 & -\omega_z(t) & -\omega_y(t) \\ \omega_z(t) & 1 & -\omega_x(t) \\ -\omega_y(t) & \omega_x(t) & 1 \end{bmatrix}, \mathbf{\Theta}(t) = \begin{bmatrix} \theta_x(t) \\ \theta_y(t) \\ \theta_z(t) \end{bmatrix} \tag{10}$$

Then equation 2 becomes,

$$\mathbf{x}(t) = \prod_{i=0}^{t}\{\mathbf{\Omega}(i)\}\mathbf{R}(0)\mathbf{X} + \sum_{i=0}^{t}\{\mathbf{\Theta}(i) + \mathbf{T}(0)\} \tag{11}$$

Since we defined world coordinate system in terms of position and orientation of the first instance of the camera center, rotation $\mathbf{R}(0) = \mathbf{I}_{3\times 3}$, and $\mathbf{T}(0)$ is the zero vector, and equation 11 becomes,

$$\mathbf{x}(t) = \prod_{i=0}^{t}\{\mathbf{\Omega}(i)\}\mathbf{X} + \sum_{i=0}^{t}\{\mathbf{\Theta}(i)\}. \tag{12}$$

In this case, the multiframe Essential matrix becomes $\mathbf{E}(t) = (\prod_{i=0}^{t}\{\mathbf{\Omega}(i)\})\mathbf{S}(t)$, where $\mathbf{S}(t)$, is the skew symmetric matrix of $(-(\prod_{i=0}^{t}\{\mathbf{\Omega}(i)\})^\top \sum_{i=0}^{t}\{\mathbf{\Theta}(i)\})$.

Note that this result is similar to the case of intercamera action matching by relating *two* moving cameras via the 'Temporal Fundamental Matrix' described in [23], where it was shown that under assumptions of small camera motion, and polynomial individual components of rotation matrices, and translation vectors, the Temporal Fundamental Matrix itself is a polynomial function. The first assumption is necessary to prevent the order of the final polynomial from exploding when multiplying rotation matrices made up of polynomial functions. Our final result is obtained with fewer matrix multiplications, hence our constraint is more robust to violations of the first assumption, and we can keep the order of the polynomial at a reasonable level. Holding on to the same assumptions, our monocular epipolar constraint becomes,

$$\mathbf{x}'(t)^\top \left(\sum_{i=0}^{k}\mathbf{F}_i t^i\right)\mathbf{x}'(0) = 0, \tag{13}$$

where each $\mathbf{F}_i$ is a $3 \times 3$ matrix of coefficients. The above equation is only valid for images of points that belong to the static scene. Hence, once the coefficients of the Multiframe Monocular Fundamental matrix are computed, using the assumption that the static non-planar scene dominates the frame, trajectories belonging to moving objects can be filtered out. If the point $\mathbf{X}$ belongs to a static object or the background in the world, then equation 4 can now be written as,

$$\Big(\mathbf{X}(t) - \mathbf{C}(t)\Big)^{\top} \mathbf{C}(t) \times \mathbf{X}(0) = 0, \tag{14}$$

where, the position of the 3D point is now a function of time, albeit constant. Notice that there is a degenerate case, where this equation holds even for a *moving* point, which is possible if the point is moving along the epipolar plane defined by $\mathbf{C}(t) \times \mathbf{X}(0)$. Although a moving object cannot be detected under these circumstances, in practice, this scenario is unlikely to be encountered.

In order to estimate the coefficients, we set the degree $k$, of the polynomial equal to 3, and rewrite equation 13 in the form of a linear system,

$$\mathbf{Of} = 0, \tag{15}$$

with observation matrix $\mathbf{O}$, and unknown vector $\mathbf{f}$, where $\mathbf{O} = \left[\mathbf{O}_1^{\top}, \mathbf{O}_2^{\top}, \ldots, \mathbf{O}_p^{\top}\right]^{\top}$, is a $p \times (3 \cdot 3 \cdot (k+1))$ matrix of $p$ correspondences. Each $\mathbf{O}_i$ is generated from one correspondence and is given as,

$$\mathbf{O}_i = \begin{bmatrix} \mathbf{r}_i & \mathbf{r}_i t & \mathbf{r}_i t^2 & \mathbf{r}_i t^3 \end{bmatrix}, \tag{16}$$

and

$$\begin{aligned} \mathbf{r}_i = \mathbf{[}\ & x_i(0)x_i(t) & x_i(0)y_i(t) & x_i(0) \ldots \\ & y_i(0)x_i(t) & y_i(0)y_i(t) & y_i(0) \ldots \\ & x_i(t) & y_i(t) & 1 \quad \mathbf{]} \end{aligned} \tag{17}$$

while the vector of unknowns $\mathbf{f}$ is a $(3 \cdot 3 \cdot (k+1)) \times 1$ vector of the Multiframe Monocular Fundamental matrix coefficients,

$$\begin{aligned} \mathbf{f} = \mathbf{[}\ & F_{0,1} \mid F_{0,2} \mid F_{0,3} \mid \ldots \\ & F_{1,1} \mid F_{1,2} \mid F_{1,3} \mid \ldots \\ & F_{2,1} \mid F_{2,2} \mid F_{2,3} \mid \ldots \\ & F_{3,1} \mid F_{3,2} \mid F_{3,3} \quad \mathbf{]}^{\top}, \end{aligned} \tag{18}$$

where $F_{i,j}$ refers to the i$^{th}$ coefficient matrix and j$^{th}$ row.

Here $\mathbf{O}$ is a rank deficient matrix of rank 35, i.e., at least 35 point correspondences are required to find a solution for $\mathbf{f}$, which is the unit eigenvector of the covariance matrix $\mathbf{O}^{\top}\mathbf{O}$ corresponding to the smallest eigenvalue.

**Input**: Set of $P$, $N$ points long trajectories, $\{\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_P\}$.
**Output**: Set of pixels, $\mathbf{D}$, classified as foreground.

---

► Compute set of correspondences, $\{\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p, \}$
► Compute set of observation vectors, $\{\mathbf{O}_i^t\}$, $1 \leq i \leq p(N-1)$, $2 \leq t \leq N$.
► $\mathbf{D} \leftarrow \emptyset$

► $q \leftarrow 1$, repeat while $q \leq Q$:

- $\mathbf{O_s} \leftarrow \emptyset$
- $t \leftarrow 2$, repeat while $t < N$:
  - ▲ Randomly choose 3 vectors, $\{\mathbf{O}_{z_1}^t, \ldots, \mathbf{O}_{z_3}^t\}$
  - ▲ $\mathbf{O_s} \leftarrow \mathbf{O_s} \bigcup \{\mathbf{O}_{z_1}^t, \ldots, \mathbf{O}_{z_3}^t\}$
  - ▲ $t \leftarrow t+1$
- Compute $\mathbf{f}_j$ using observation $\mathbf{O_s}$ (eq. 15)
- Compute error $\mathbf{e}_q$ for all $\mathbf{\Phi}$, (eq. 20)
- $q \leftarrow q+1$

► Compute $\mathbf{\Sigma}$, using eq. 21

► $p \leftarrow 1$, repeat while $p \leq P$:

- Compute $\varepsilon_p$, using eq. 22
- If $\varepsilon_p > \tau$, $\mathbf{D} \leftarrow \mathbf{D} \bigcup \{p\}$
- $p \leftarrow p+1$

**Fig. 4.** Algorithmic overview of the proposed framework. See text for details.

## 4 Independently Moving Object Detection

The proposed method for detecting independently moving objects, by finding trajectories of points that are outliers to the previously derived Multiframe Monocular Epipolar constraint consists of the following steps (see Fig. 2 for reference). Given a sequence of frames we first compute optical flow between consecutive frames. Particle advection is then employed to obtain a set of pixel-wise dense trajectories for a subsequence of $N$ frames $\mathbf{\Lambda}_i = [\mathbf{x}_i(0), \mathbf{x}_i(1), \ldots, \mathbf{x}_i(N-1)]$. We normalize the positions of the points within all trajectories to their mean within the entire subsequence. This step is similar to the normalization performed in case of the standard 8-point algorithm [24]. Positions within each trajectory $\mathbf{\Lambda}_i$, provide a set of $N-1$ correspondences, written as,

$$\mathbf{\Phi}_i = \Big[ \left(\mathbf{x}_i(0), \mathbf{x}_i(1)\right), \ldots, \left(\mathbf{x}_i(0), \mathbf{x}_i(N-1)\right) \Big]. \tag{19}$$

Each correspondence $j$ from each trajectory $\mathbf{\Lambda}_i$ provides a single observation vector $\mathbf{O}_{i,j}$. All of these observation vectors are then assembled into the observation matrix $\mathbf{O}$, as described in section 3.

Using this matrix we robustly determine which trajectories are outliers by employing a RANSAC based framework. In our implementation we set $N = 15$, which results in 14 correspondences per trajectory. For the Multiframe Fundamental matrix of degree 3, 36 correspondences are sufficient. At each iteration of the algorithm, we randomly select observation vectors corresponding to 3 correspondences from each time instant resulting in $3 \times 14 = 42$ correspondences, for our subsequence of $N$ frames. From these observation vectors, we construct a sample observation matrix $\mathbf{O_s}$, where $\mathbf{s}$ is a vector of indices. We then compute the coefficients of the Multiframe Monocular Fundamental matrix (equation 15), $\mathbf{f}_q$, where $q$ is the current iteration of RANSAC. We then compute a vector of errors for all the correspondences $\mathbf{e}_q$ given as,

$$\mathbf{e}_q = |\mathbf{O}\mathbf{f}_q|, \tag{20}$$

where each element is the Epipolar distance of a single correspondence.

After $Q$, pre-selected number of iterations of the framework, we obtain the average error vector,

$$\mathbf{\Sigma} = \frac{1}{Q} \sum_{q=1}^{Q} \mathbf{e}_q. \tag{21}$$

Finally, for each trajectory $\mathbf{\Lambda}_i$, we obtain its error $\varepsilon_i$, by taking the mean of the errors of its correspondences, by indexing into its corresponding entries in the vector $\mathbf{\Sigma}$, which can be written as,

$$\varepsilon_i = \frac{1}{N-1} \sum_{l=1}^{N-1} \mathbf{\Sigma}(l + (N-1)i) \tag{22}$$

The proposed framework is algorithmically described in figure 4.

## 5 Advantages over Fundamental Matrix

In theory, one can compute a regular fundamental matrix between some frame pairs, compute the error of the points, and average the result for the trajectory. However, in order for fundamental matrix to work properly, the frames must have a wide baseline. Since the motion of the camera is unknown, it is not clear between which frames the fundamental matrix should be computed. Any temporal distance that is selected between the frames will work for some pairs but not for others depending on the motion of the camera. For example one can try to use a floating baseline by computing a fundamental matrix between frames $(1, t)$, for all $2 \leq t \leq N$, where $N$ is the number of frames. However, if the camera is accelerating from low initial velocity, the initial frame pairs (1,2), (1,3), may have a baseline that is too narrow, resulting in poor matrix estimation and motion segmentation. By contrast, the MMFM will properly capture the evolution of the motion of the camera, and be able to correctly segment out the moving objects.

This effect is illustrated in Figure 5. We used synthetic data for this experiment consisting of 1000 static points and 10 randomly moving points. The points are randomly

distributed through space in front of a moving camera, they are imaged, and random noise is added to their image. We then fit a set of fundamental matrices to the static points and compute the error of all points. We do the same for the proposed MMFM. In Figure 5 $(a)$ and $(b)$ the camera is moving with a initial velocity, and then accelerates or decelerates. As can be seen from the figure, when MMFM is used to model the motion of the camera there is a clear separation between the errors of moving points, and the errors of static points for a wide range of noise. By constrast, in the case of regular fundamental matrix, any given threshold works only for a narrow range of noise. What's worse, is that for some levels of noise the error of static points actually becomes greater than the error of moving points making the separation impossible. Results for other noise models, camera motions, and fundamental matrix computation schemes are provided in the supplementary materials.

We have also examined how well our proposed constraint separates the moving points from the static points under degenerate conditions. In the first case, shown in Figure $(c)$, we examined the degeneracy caused by the points moving along the epipolar plane. We computed the proposed constraint for the batch of 14 frames, and then computed the error between the stationary and moving points. The $x$ axis of the graph indicates the level of degeneracy, where the numbers at the bottom correspond to the number of frames (out of 14) for which moving points moved along the epipolar plane. It can be seen from the figure that the sets of static and moving points are still separable using a single threshold unless the points move along the epipolar plane for all the frames used for the MMFM computation. Moreover, if the camera motion is polynomial over time, the chances of a point moving along the epipolar plane for a long time is very remote. The second degenerate case that we tested, (Figure 5 (d)) simulates the scenario when camera is stationary for some of the frames. Once again, the $x$ axis captures the number of frames for which the camera remains stationary. For the simulation of the above two degenerate cases we have used a small random noise $(0.01)$. The noise allows the moving and stationary objects to remain separable even in the case of stationary camera.

## 6 Experiments and Results

To validate our proposed method we performed extensive set of experiments on a diverse set of video sequences captured from moving cameras (aerial, and hand held) and in presence of large out of plane objects. We compared the performance of our method with homography based background subtraction method, rank constraint trajectory pruning method of Sheikh [22] et. al, and a method based on the regular fundamental matrix constraints. In the case of homography based motion compensation, in order to perform a maximally fair comparison, we employed the same framework for comparison, where we exploit optical flow based dense frame to frame correspondences to compute homography between two frames. RANSAC is then used to compute the optimal homography matrix and the outliers of RANSAC are detected as moving objects. We also tested the sequences using conventional [9] style homography based image warping and background subtraction framework and similar results were obtained. In order to detect motion based on the regular fundamental matrix we perform particle

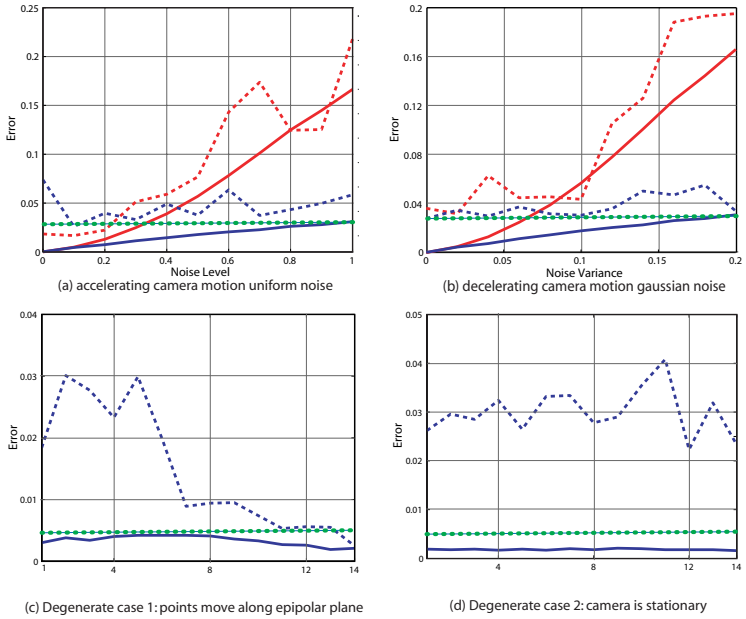


**Fig. 5.** This figure shows error values between moving points (dashed lines) and static points (solid lines) for synthetic data, generated over different levels of noise. Sub-figures (a) and (b) show the scenarios when camera accelerate and decelerate respectively. Sub-figures (c) and (d) respectively illustrates the degenerate cases when points move along the epipolar plane and when camera remains stationary. Red curves are for regular fundamental matrix, while blue curves are for our proposed constraint. The dotted green line indicates an error threshold.


**Fig. 6.** Qualitative comparison on Seq 1 (on the left), JHU cars6 (center), and Seq 4 (right). Red indicates detections. (a) Homography, (b) rank constraint, (c) fundamental matrix, (d) our method.

advection for 13 frames, then we compute fundamental matrices between frames 1 to 7, 2 to 8 and so on to ensure sufficient translation. The matrices are fit using RANSAC, where we accumulate errors for each particle over all iterations. A particles track is selected as belonging to a moving objects if its cumulative error over all iterations of RANSAC and all fundamental matrices in the frame range is above a threshold. We use this particular RANSAC framework to make sure that the differences in performance between the regular fundamental matrix and the proposed multi-frame constraint are due to the geometric model and are not caused by differences in the implementation. This particular RANSAC framework gave better, more consistent results for both methods. Figure 5 shows a qualitative comparison between the four methods. Even though the sequence is aerial in leftmost Figure 5, the scene has a very large out of plane object - the water tower, which causes the orthographic camera assumption to fail. Additionally, parts of the ground-plane are obscured by uneven out-of-plane foliage. These factors cause homography and the rank constraint to fail, resulting in false detections on the out-of-plane tower. In the center and rightmost panels of Figure 5 there are multiple planes present in the scene, and portions of background are at different depths, causing problems for homography and rank constraint, when the camera undergoes translation. As outlined in section 5, motion segmentation based on the fundamental matrix suffers from stability problems related to baseline selection, noise, and motion of the camera.

For better illustration of the performance of our method, we provide quantitative comparison of the results. First we obtained the ground truth for all sequences used in our experiments by manually selecting a silhouette around each moving object of each frame. We then compared the detection results of the methods mentioned above, as well as the proposed method with the ground truth. In order to quantify the performance, we used measures similar to VACE performance measures [25], which are area based measures which penalize false detections as well as missed detections. Accuracy of detection, called frame detection accuracy ($FDA$), is estimated as: $FDA(t) =$ Overlap Ratio$/N_G^{(t)}$, where Overlap Ratio is,$\sum_{i=1}^{N_G^{(t)}} (|G_i^{(t)} \bigcap D_i^{(t)}|)/(|G_i^{(t)} \bigcup D_i^{(t)}|)$. $N_G^{(t)}$ is the number of ground truthed objects in frame *t*, $G_i(t)$ and $D_i(t)$ are the sets of pixels belongs to the object number *i* of frame number *t*, in the ground truth and computed detection respectively, where a set of connected pixels $D_j$ of the detection, referred to as a blob, is mapped to the $i^{th}$ ground truth object, if $G_i^{(t)} \bigcap D_j^{(t)} \neq \emptyset$. All the detected blobs in a frame which cannot be mapped with any of the groundtruth objects are classified as false detections and a measure of per frame false detection called Frame False Detection Ratio ($FFDR$) is computed as: $FFDR(t) = FD(t)/TD(t)$, where, $FD(t)$ is the combined area of all false detections, and $TD(t)$ is the combined area of all detections.

It is therefore ensured that $FFDR$ always remains between 0 to 1 and depends on the size of the false detection. For the whole sequence, the performance measures are simply calculated as Sequence Frame Detection Accuracy, $SFDA = \sum_{t=1}^{N} FDA(t)/N$, and Sequence False Detection Ratio, $SFFDR = \sum_{t=1}^{N} FFDR(t)/N$

Quantitative results of comparison of our approach with Homography based detection, as well as [22] are reported in table 1. On sequences 1, 2, and 4 our method obtained superior results both in terms of true detections, and false positives. In sequences 3 and 5, we achieve much better false positive rate, but our detection is slightly

| | Frame Count | Homography [9] | | Rank Constraint [22] | | Fundamental Matrix | | Our Method | |
|---|---|---|---|---|---|---|---|---|---|
| | | $SFDA$ | $SFFDR$ | $SFDA$ | $SFFDR$ | $SFDA$ | $SFFDR$ | $SFDA$ | $SFFDR$ |
| **Seq1** | 390 | 0.3844 | 0.0536 | 0.5627 | 0.0419 | 0.4262 | 0.0011 | 0.7611 | 0.00092 |
| **Seq2** | 650 | 0.5957 | 0.0477 | 0.6494 | 0.0610 | 0.7619 | 0.0212 | 0.8664 | 0.0339 |
| **Seq3** | 400 | 0.8184 | 0.2683 | 0.8425 | 0.1949 | 0.6557 | 0.0961 | 0.7202 | 0.0713 |
| **Seq4** | 280 | 0.3117 | 0.1433 | 0.3517 | 0.1224 | 0.3833 | 0.0640 | 0.5311 | 0.0594 |
| **Seq5** | 130 | 0.6738 | 0.08 | 0.4583 | 0.0545 | 0.4358 | 0.0553 | 0.6268 | 0.003 |
| **cars3** | 20 | 0.9122 | 0.0274 | 0.9412 | 0.0053 | 0.5414 | 0.0187 | 1.0 | 0.0042 |
| **cars4** | 54 | 0.8406 | 0.0332 | 0.6933 | 0.1556 | 0.6419 | 0.1626 | 0.8375 | 0.0323 |
| **cars5** | 37 | 0.8128 | 0.0110 | 0.9351 | 0.0191 | 0.7672 | 0.0376 | 0.9464 | 0.0081 |
| **cars6** | 31 | 1 | 0.0592 | 1 | 0.0049 | 1 | 0.0099 | 1 | 0.0023 |
| **cars7** | 25 | 1 | 0.0929 | 1 | 0.3550 | 0.7751 | 0.2785 | 1 | 0.0141 |
| **cars9** | 61 | 0.4452 | 0.0711 | 0.3603 | 0.0256 | 0.3137 | 0.0935 | 0.5165 | 0.0164 |

**Table 1.** Shows a comprehensive analysis of the performance of the four methods. $SFDA$ is the overall detection accuracy and $SFFDR$ is the overall measure of false detection for a sequence. Sequences 1, 2, and 5 are aerial from VIVID 3 and 2 datasets. Sequences 3 and 4 are hand-held collected by us. The cars sequences are from the JHU155 dataset.

lower than homography in sequence 5, and both homography and rank in sequence 3. Performance on the JHU 155 dataset is similar for all four methods. This is because the sequences are very short, and the camera does not have enough time to undergo large translation to cause severe parallax distortion.

## 7 Conclusion

We developed a novel method for detecting objects that are moving in a general 3D scene, in video captured by a moving camera, while avoiding motion registration, assumptions about the planarity of the scene, and the use of metadata. To do so we proposed a monocular multi-frame epipolar constraint, which we derived from an evolving epipolar plane defined by the motion of the center of the camera, and a scene point. We parameterized it as a polynomial function of time, in order to estimate fewer unknowns. We showed comparative qualitative and quantitative results on real and synthetic sequences, and analyzed the results.

## Acknowledgments

This research was funded by Harris Corporation.

## References

1. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: CVPR. (1999)
2. Jain, R., Nagel, H.: On the analysis of accumulative difference pictures from image sequences of real world scenes. In: PAMI. (1979)

3. Javed, O., Rasheed, Z., Alatas, O., Shah, M.: A real time surveillance system for multiple overlapping and non-overlapping cameras. In: ICME. (2003)
4. Zhong, J., S., S.: Segmenting foreground objects from a dynamic textured background via a robust kalman filter. In: ICCV. (2003)
5. Hayman, E., Eklundh, J.: Statistical background subtraction for a mobile observer. In: ICCV. (2003)
6. Middal, A., Paragios, N.: Motion-based background subtraction using adaptive kernel density estimation. In: CVPR. (2004)
7. Friedman, N., Russel, S.: Image segmentation in video sequences: A probabilistic approach. In: UAI. (2000)
8. Haritaoglu, I. Harwood, D.D.L.: W4: Real-time surveillance of people and their activities. In: PAMI. (2000)
9. Ali, S., Shah, M.: Cocoa: tracking in aerial imagery. Volume 6209., SPIE (2006)
10. Kaucic, R., Perera, A., Brooksby, G., Kaufhold, J., Hoogs, A.: A unified framework for tracking through occlusions and across sensor gaps. In: CVPR. (2005)
11. Kang, J., Cohen, I., Yuan, C.: Detection and tracking of moving objects from a moving platform in presence of strong parallax. In: ICCV. (2005)
12. Irani, M., Anandan, P.: A unified approach to moving object detection in 2d and 3d scenes. In: PAMI. Volume 20. (1998)
13. Sawhney, S.H., Guo, Y., Kumar, R.: Independent motion detection in 3d scenes. In: PAMI. Volume 22. (2000)
14. Cheng, H., Butler, D., Basu, C.: ViTex: Video to tex and its application in aerial video surveillance. In: CVPR. (2006)
15. Xiao, J., Cheng, H., Han, F., Sawhney, H.: Geo-spatial aerial video processing for scene understanding and object tracking. In: CVPR. (2008)
16. Pollard, T., Mundy, J.L.: Change detection in a 3-d world. In: CVPR. (2007)
17. Wang, J., Adelson, E.: Representing moving images with layers. In: TIP. (1994)
18. Tao, H., Sawhney, H.S., Kumar, R.: Object tracking with bayesian estimation of dynamic layer representations. In: TPAMI. Volume 24. (2002)
19. Ke, Q., Kanade, T.: A subspace approach to layer extraction. In: CVPR. (2001)
20. Xiao, J., Shah, M.: Motion layer extraction in the presence of occlusion using graph cuts. In: TPAMI. Volume 27. (2005)
21. Yuxin Jin, e.a.: Background modeling from a free-moving camera by multi-layer homography algorithm. In: ICIP. (2008)
22. Yaser Sheikh, Omar Javed, T.K.: Background subtraction for freely moving cameras. In: ICCV. (2009)
23. Yilmaz, A., Shah, M.: Matching actions in presensce of camera motion. In: CVIU. Volume 105. (2006)
24. Hartley, R.I.: In defense of 8-point algorithm. In: TPAMI. Volume 19. (1997)
25. Kasturi, R.e.a.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. PAMI **31**(2) (2009) 319 –336