Detection and Tracking of Moving Objects from a Moving Platform in Presence of Strong Parallax

Jinman Kang, Isaac Cohen, Gérard Medioni, Chang Yuan
Institute of Robotics and Intelligent Systems
University of Southern California, CA 90089, USA
{jinmanka, icohen, medioni, cyuan}@usc.edu

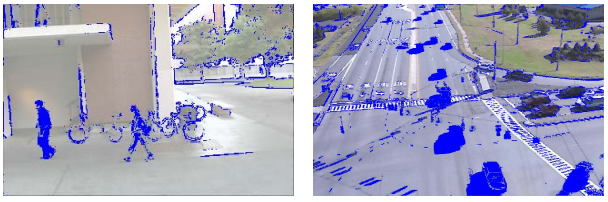
Abstract

We present a novel approach to detect and track independently moving regions in a 3D scene observed by a moving camera in the presence of strong parallax. Detected moving pixels are classified into independently moving regions or parallax regions by analyzing two geometric constraints: the commonly used epipolar constraint, and the structure consistency constraint. The second constraint is implemented within a "Plane+Parallax" framework and represented by a bilinear relationship which relates the image points to their relative depths. This newly derived relationship is related to trilinear tensor, but can be enforced into more than three frames. It does not assume a constant reference plane in the scene and therefore eliminates the need for manual selection of reference plane. Then, a robust parallax filtering scheme is proposed to accumulate the geometric constraint errors within a sliding window and estimate a likelihood map for pixel classification. The likelihood map is integrated into our tracking framework based on the spatio-temporal Joint Probability Data Association Filter (JPDAF). This tracking approach infers the trajectory and bounding box of the moving objects by searching the optimal path with maximum joint probability within a fixed size of buffer. We demonstrate the performance of the proposed approach on real video sequences where parallax effects are significant.

1. Introduction

Detection and tracking of independently moving objects from the scene are important elements in video surveillance. In the case of a moving camera, the detection and tracking problems are inherently more complex, since the camera motion induces a motion in all pixels of the image.

A common approach for detecting moving regions relies on the stabilization due to the camera motion using parametric motion models and defining moving pixels as the ones that have not been stabilized. This can be implemented robustly and efficiently as in [7]. This works well when the scene can be considered planar, or when the motion of the camera is pan/tilt/zoom. Otherwise, 3D depth in the scene produces pixel displacement which cannot be accounted for by the global parametric model, usually termed as *parallax*. The flagged pixels after stabilization correspond to either independently moving regions or parallax regions. The essential issue becomes the pixel classification into planar background, parallax or independent motion. Figure 1 presents examples of the parallax effects from various moving cameras: Figure 1(a) shows the detection results from a ground camera; Figure 1(b) shows the detection results from an airborne camera. As one can see, moving objects (humans and vehicles) are marked (in blue), but so are the parallax pixels.



(a) Ground based camera (b) Airborne camera

Figure 1. Examples of parallax effects

Geometric constraints within multiple frames provide additional cues for this classification task. One natural choice is the epipolar constraint in image pairs. A static 3D point must lie on the epipolar plane which is determined by the point itself and two camera centers. Departure from the epipolar plane characterizes the presence of independent motion. This departure can be measured by the angle between two epipolar lines derived from point correspondence and from fundamental matrix. However, the epipolar constraint is not sufficient, as independently moving objects may also move on the epipolar plane. In other words, if cor-

responding points are not moving along the epipolar lines, they can be conclusively established as independently moving pixels, but the reverse is not always true.

In order to remove the epipolar ambiguity, we utilize a second geometric constraint: the scene structure in the Euclidean world coordinate system remains constant. This structure consistency constraint is implemented within a "Plane+Parallax" framework [6, 10, 12] which represents the scene structure by a 2D residual displacement (parallax) field w.r.t. a 3D reference plane. The magnitude of each parallax displacement is directly related to projective structure, *i.e.* relative depth. If a pair of 2D points correspond to the same static 3D point, their relative depths are constrained by the structure consistency. The deviation from this constraint indicates that the point belongs to an independently moving object. This provides an additional cue for pixel classification.

The structure consistency constraint was also proposed in [5, 11]. They also assume that the reference plane is constant in the scene and consequently the distances of 3D points from the plane are constant. However, this assumption is violated in our case since the reference planes usually change for different pairs of frames. Therefore, these constraints are not applicable in our case.

We derive a new bilinear constraint relating the point pairs to their relative depths. The constraint is a 4×4 matrix which absorbs the camera internal parameters, camera motion and the transformation between reference planes. Thus, this constraint only requires the 2D image points, 2D homographies and epipoles. It does not assume a constant reference plane and consequently eliminates the need for manually identifying a constant reference plane in the video sequence.

The parallax filtering process using these two constraints allows us to identify the presence of parallax in the residual pixels. Instead of filtering out the parallax at the detection step using a threshold, we define probability measures representing the likelihood of a pixel to belong to a moving region. Furthermore, to ensure stability of the proposed approach, we propose to use a temporal sliding window to utilize information from multiple frames. The error measures within this sliding window are accumulated into a likelihood map.

This likelihood map is integrated to a tracking framework based on the Joint Probability Data Association Filter (JPDAF) [7]. This spatio-temporal joint probability model combines the geometric constraint errors with other cues such as the appearance and motion of the moving objects. The tracking problem is formulated as maximization of the joint probability by searching an optimal path within a fixed size buffer.

The rest of the paper is organized as follows. In the next subsection, the previous work is briefly reviewed. An overview of the proposed approach is given in Section 2. In Section 3, the two geometric constraints are defined. Section 4 describes how the geometric constraint errors are accumulated for parallax filtering. Section 5 presents a spatiotemporal stochastic tracking framework that integrates multiple cues including the geometric constraint errors. Our results on real sequences are presented and discussed in Section 6. Finally, we conclude the paper with a discussion on future work in Section 7.

1.1. Previous Work

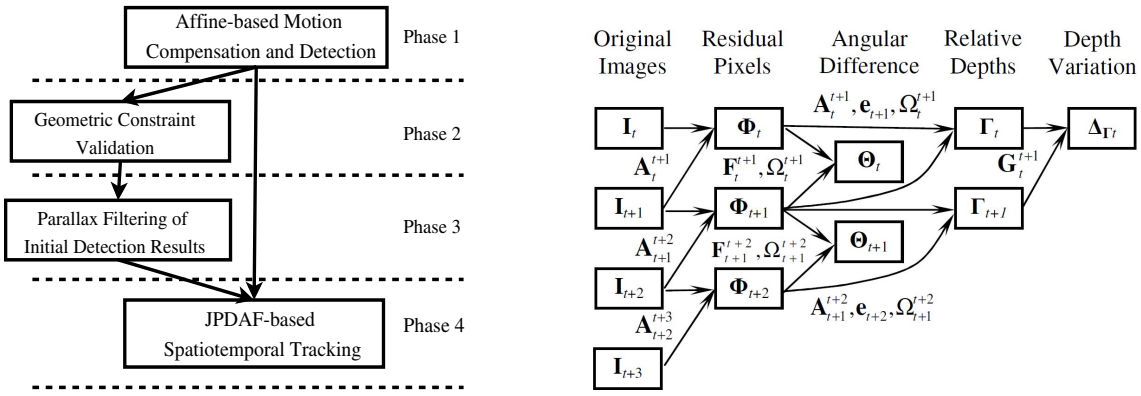
There exists a large body of literature related to our approach. Only the most relevant ones are briefly reported here. Many approaches [3,7] in motion detection and tracking assume that the scene can be approximated by a plane (*i.e.* a flat ground), and image transformation is modeled as a 2D affine or projective transformation. These approaches work well in general, even when the reference plane corresponding to the affine stabilization does not correspond to the ground plane. However, when this assumption is violated due to the presence of tall structures such as buildings, or trees, the observed motion of these stationary structures do not follow the planar approximation of the scene, and these structures are often incorrectly detected as independently moving regions.

In case of strong parallax, geometric and shape constraints are integrated for distinguishing independently moving regions from parallax regions. In [9], the mismatch between a spatiotemporal gradient measurement and projected flow is used as the indicator for independent motion. Whereas, a constant reference plane is assumed and adaptive thresholding is needed for pixel classification. [13] enforces the trifocal constraints in image triplets and clusters the tracked image features (*e.g.* points) into different motion groups. However, the motion segmentation is only applied to sparsely matched features.

Many approaches are based on the "Plane+Parallax" representation [5, 6, 8, 10–12]. [5] proposed a pair-wise rigidity constraint in multiple frames which relates the relative depth and parallax displacements. In [11], a shape constraint for each point triplet is proposed based on the trifocal constraint by eliminating the relative depth. All these methods require that the reference plane is constant and therefore the distances of 3D points to the plane are constant. In contrast, our method is capable of handling different reference planes, and does not require the manual selection of the reference plane.

2. Overview of the Proposed Approach

The pipeline of our approach is shown in Figure 2(a). As the initial phase, an affine motion compensation and



(a) The pipeline

(b) Geometric constraint validation from four frames

Figure 2. Overview of the proposed approach

detection framework is applied to consecutive frames [7]. The residual pixels correspond either to parallax or independently moving regions. In order to identify independent motion in the initial detection results, we estimate the geometric constraint errors after four consecutive frames. During the parallax filtering process, the constraint errors are accumulated within a buffer and represented in probabilistic likelihood models. Multiple cues from appearance, motion of detected blobs and the likelihood maps from parallax filtering are integrated into a JPDAF-based multi-frame tracking model. The approach follows a transition from two-frame processing (*phase* 1) to four-frame processing (*phase* 2) and finally to the multi-frame processing (*phase* 3 and 4).

The affine motion detection framework initially extracts a number of feature points in each frame by using the Harris corner detector. Then the feature points in consecutive frames $\mathbf{I}_t$[1] and $\mathbf{I}_{t+1}$ are matched by evaluating the cross-correlation of local windows around feature points. A 2D affine motion model A_t^{t+1} is robustly estimated by fitting the model to at least three pairs of matched points within a RANSAC-based scheme [2]. This affine model is used not only for motion compensation and detection, but also to estimate the homography matrix for the later "Plane+Parallax" representation in *phase* 2. The affine motion model A_t^{t+1} globally compensates for the motion of pixels from $\mathbf{I}_t$ to $\mathbf{I}_{t+1}$. The pixels that do not satisfy this motion model are classified as residual pixels $\mathbf{\Phi}_t$.

Before computing the geometric errors, the epipolar geometry is also estimated from the matched feature points in every two consecutive frames. The fundamental matrix F_t^{t+1} is estimated by a RANSAC-based 8-point algorithm [4]. The corresponding epipoles $\mathbf{e}_t$ and $\mathbf{e}_{t+1}$ are obtained as the null vector of the fundamental matrix. As shown in Figure 2(b), the geometric constraint errors are computed on the residual pixels in four consecutive frames. A set of dense point correspondences (optical flow) $\mathbf{\Omega}_t^{t+1}$ is defined between two residual pixel maps $\mathbf{\Phi}_t$ and $\mathbf{\Phi}_{t+1}$, instead of the whole image, as follows:

$$\{\mathbf{p}_t \rightarrow \mathbf{p}_{t+1} : I(\mathbf{p}_t) = I(\mathbf{p}_{t+1}), \mathbf{p}_t \in \mathbf{\Phi}_t, \mathbf{p}_{t+1} \in \mathbf{\Phi}_{t+1}\}$$

where $I(\mathbf{p}_t)$ is the image intensity of point $\mathbf{p}_t$ in $\mathbf{I}_t$. The optical flow is estimated by employing the intensity window matching techniques [1].

Then an angular difference map $\mathbf{\Theta}_t$ is obtained by applying the epipolar constraints and point correspondences to residual pixels. Combining the epipole, optical flow, affine motion and residual pixels, we obtain a relative depth map $\mathbf{\Gamma}_t$ within the "Plane+Parallax" framework. Based on the structure consistency constraint, a bilinear relationship G_t^{t+1} is derived to measure the errors between relative depth pairs and generate a depth variation map $\mathbf{\Delta}_{\Gamma_t}$ between two relative depth maps. Therefore, each four consecutive frames $\mathbf{I}_t, \ldots, \mathbf{I}_{t+3}$ generate two angular difference maps and one depth variation map.

In order to suppress the estimation error in one single frame, we propose to use a sliding window (typically 5 frames) approach for parallax filtering. The angular difference maps and the depth variation maps are accumulated within the window and represented in likelihood functions. The filtering result is a likelihood map, instead of a binary mask image obtained by thresholding.

The JPDAF-based tracking framework infers the motion trajectory and bounding boxes of detected objects. It integrates multiple cues, such as the appearance and motion of detected blobs, and the geometric constraint errors, as observation nodes into a joint probability model. The joint probability is maximized by searching the optimal path across the nodes in a fixed-size buffer.

[1]The bold greek letters refers to the set of geometric elements.

3. Geometric Constraint Errors

Two geometric constraints are proposed to eliminate the false detection of parallax regions. The disparity of residual pixels to the constraints is consequently defined.

3.1. Epipolar Constraint Errors

Let $\mathbf{P}$ denote a static 3D point in the scene and $\mathbf{p}_t$ and $\mathbf{p}_{t+1}$ as its projections in $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$. Let $\mathbf{l}_{t+1}$ denote a line connecting $\mathbf{e}_{t+1}$ and $\mathbf{p}_{t+1}$ in $\mathbf{I}_{t+1}$ and similarly $\mathbf{l}_t$ connecting $\mathbf{e}_t$ and $\mathbf{p}_t$ in $\mathbf{I}_t$. These two epipolar lines are derived from the optical flows (image apparent motion) as follows,

$$\mathbf{l}_{t+1} = \mathbf{e}_{t+1} \times \mathbf{p}_{t+1} \ , \ \mathbf{l}_t = \mathbf{e}_t \times \mathbf{p}_t \tag{1}$$

On the other hand, using the fundamental matrix, we could also obtain two epipolar lines $\mathbf{l}'_{t+1}$ and $\mathbf{l}'_t$:

$$\mathbf{l}'_{t+1} = F_t^{t+1}\mathbf{p}_t \ , \ \mathbf{l}'_t = (F_t^{t+1})^T\mathbf{p}_{t+1} \tag{2}$$

Since static points lie on the epipolar plane, they satisfy both optical flow constraints and epipolar constraints, so that $\mathbf{l}_t \cong \mathbf{l}'_t$ and $\mathbf{l}_{t+1} \cong \mathbf{l}'_{t+1}$ ($\cong$ means equal up to a scale factor). However, the points on moving object do not satisfy this epipolar constraint. Therefore there exists a discrepancy between the lines derived from point correspondence and from fundamental matrix, as illustrated in Figure 3.

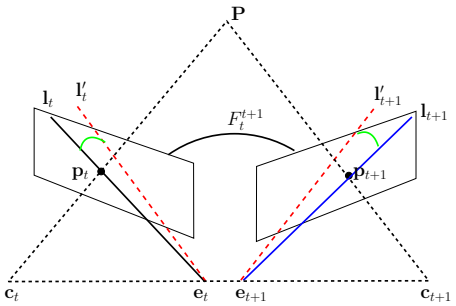


Figure 3. Epipolar constraint

Therefore, identifying moving pixels belonging to independently moving objects in the scene can be performed by evaluating the average angular difference between epipolar lines as

$$\theta_t^i = [\angle(\mathbf{l}_t, \mathbf{l}'_t) + \angle(\mathbf{l}_{t+1}, \mathbf{l}'_{t+1})]/2 \tag{3}$$

where $\angle$ denotes the angle between two 2D lines. Due to the presence of estimation errors of epipolar geometry and image noise, the angular differences for static points are generally not zero. Indeed, the larger the angular difference is, the more likely the 3D point to belong to an independently moving object.

In Figure 4(a), both independently moving objects and the parallax from the stationary structures are detected as moving regions. In Figure 4(b), one can see that the angular difference for moving humans is higher (in purple) while that for trees and building edges is lower (in green).

We argue that this angular difference derived from the discrepancy between point correspondence and epipolar constraint is not sufficient. The independently moving objects may possibly moving along the epipolar line or more generally in the epipolar plane, as shown in Figure 1(a). In these cases, the proposed angular difference may not be able to distinguish between parallax regions and independently motion regions. In order to resolve this ambiguity, a second geometric constraint is proposed in the next section.

3.2. Structure Consistency Constraint Errors

The structure consistency constraint is implemented within the "Plane+Parallax" representation which provides a dense estimation of scene structure relative to a reference plane. Given a 2D homography between two frames and the epipoles, the relative depth of a pair of matched points can be estimated as follows [3]:

$$\gamma_t^i = \frac{(A_t^{t+1}\mathbf{p}_t^i \times \mathbf{p}_{t+1}^i)^T(\mathbf{p}_t^i \times \mathbf{e}_{t+1})^T}{\left\|\mathbf{p}_t^i \times \mathbf{e}_{t+1}\right\|^2} \tag{4}$$

where $\mathbf{p}_t^i$, $\mathbf{p}_{t+1}^i$ are the i^{th} matched point pairs between frame $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$, γ_t^i is the estimated relative depth for $\mathbf{p}_t^i$, A_t^{t+1} is the 2D affine motion model (homograpy), and $\mathbf{e}_t$, $\mathbf{e}_{t+1}$ are the epipoles.

A dense relative depth map $\mathbf{\Gamma}_t\{(\mathbf{p}_t^i, \gamma_t^i)\}$ is therefore constructed for the residual pixels in $\mathbf{I}_t$, *i.e.* the map $\mathbf{\Phi}_t$. One single depth map does not provide enough cues for classifying a pixel into parallax or independent motion, as the depth of points belonging to either part is generally non-zero. However, this classification is possible between at least two relative maps. Based on the assumption that the scene structure (parallax regions) remains constant in the Euclidean world coordinate, a bilinear relationship is derived to relate a pair of relative depth maps. If a pixel does not conform to this relationship, it is classified as independent motion. Therefore, the structure consistency constraint provides more cues for pixel classification.

Constraints between the relative depth pairs $(\mathbf{p}_t^i, \gamma)$ and $(\mathbf{p}', \gamma')$[2] were proposed by previous methods [5, 11, 12]. These constraints are based on the assumption that the reference plane is constant and therefore the points' heights are constant. However, the reference planes selected by 2D homographies are generally different as the homographies are not the same. Therefore, these constraints are not applicable in our case.

The bilinear relationship is derived below. Given a static 3D point in the world coordinate, its camera coordinates

[2]The point indices are dropped if no confusion is caused.

in $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$ are respectively denoted by $\mathbf{P}(x, y, z)^T$ and $\mathbf{P}'(x', y', z')^T$. Their 2D projections are respectively $\mathbf{p}(u, v, 1)^T$ and $\mathbf{p}'(u', v', 1)^T$. Integrating the camera motion model $\mathbf{P}' = R\mathbf{P} + \mathbf{t}$ and the perspective projection model $\mathbf{p} = M\mathbf{P}/z$ and $\mathbf{p}' = M'\mathbf{P}'/z'$, we have the relationship between the 2D projections [12],

$$z' M'^{-1}\mathbf{p}' = zRM^{-1}\mathbf{p} + \mathbf{t} \tag{5}$$

where R and $\mathbf{t}$ are the camera rotation and translation, M and M' are the camera internal parameter matrices. This is the mathematical representation of structure consistency.

The geometric definition of the relative depth γ is a ratio of point depth over the distance of the point to the reference plane as follows

$$\gamma = \alpha\frac{H}{z} = \alpha\frac{\mathbf{n}^T\mathbf{P} - d}{z} = d\alpha\frac{\mathbf{v}^T\mathbf{P} - 1}{z} \tag{6}$$

where H is the distance of point $\mathbf{P}$ to the reference plane Π, also called "height". The normal vector of Π is $\mathbf{n}$ and d is the height of the original point. The scaled normal vector is $\mathbf{v} = \mathbf{n}/d$. α is a constant for each frame and after normalizing the depth map, $d\alpha$ could be set to be 1. Thus the relative depth for $\mathbf{p}$ and $\mathbf{p}'$ are simplified as,

$$\gamma = \frac{\mathbf{v}^T\mathbf{P} - 1}{z},\ \gamma' = \frac{\mathbf{v}'^T\mathbf{P}' - 1}{z'} \tag{7}$$

where $\mathbf{v}'$ is the scaled normal vector for Π' in $\mathbf{I}_{t+1}$.

Let the third row of R be denoted by $\mathbf{r}_3$ and the third component of $\mathbf{t}$ is t_3, the depth z' of point $\mathbf{P}'$ could be derived from (5) as below,

$$z' = z(\mathbf{r}_3 M^{-1}\mathbf{p}) + t_3 \tag{8}$$

z and z' can be represented alternatively from (7),

$$z^{-1} = \mathbf{v}^T M^{-1}\mathbf{p} - \gamma,\ z'^{-1} = \mathbf{v}'^T M'^{-1}\mathbf{p}' - \gamma' \tag{9}$$

By substituting (9) into (8), we have:

$$\begin{aligned}&\mathbf{v}^T M^{-1}\mathbf{p} - \gamma = \\ &\quad(\mathbf{v}'^T M'^{-1}\mathbf{p}' - \gamma')[(\mathbf{r}_3 + t_3\mathbf{v}^T)M^{-1}\mathbf{p} - t_3\gamma]\end{aligned} \tag{10}$$

Derive each side of (10),

$$\text{LHS} = \begin{bmatrix}\mathbf{p}'\\ \gamma'\end{bmatrix}^T \begin{bmatrix}0 & 0 & 1 & 0\end{bmatrix}^T \begin{bmatrix}\mathbf{v}^T M^{-1} & -1\end{bmatrix} \begin{bmatrix}\mathbf{p}\\ \gamma\end{bmatrix} = \begin{bmatrix}\mathbf{p}'\\ \gamma'\end{bmatrix}^T G_1 \begin{bmatrix}\mathbf{p}\\ \gamma\end{bmatrix}$$

$$\begin{aligned}\text{RHS} &= \begin{bmatrix}\mathbf{p}'\\ \gamma'\end{bmatrix}^T \begin{bmatrix}M'^{-T}\mathbf{v}'\\ -1\end{bmatrix} \begin{bmatrix}(\mathbf{r}_3 + t_3\mathbf{v}^T M^{-1}) & -t_3\end{bmatrix} \begin{bmatrix}\mathbf{p}\\ \gamma\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{p}'\\ \gamma'\end{bmatrix}^T G_2 \begin{bmatrix}\mathbf{p}\\ \gamma\end{bmatrix}\end{aligned}$$

Finally we have $\begin{bmatrix}\mathbf{p}'^T & \gamma'\end{bmatrix} G_{4\times4} \begin{bmatrix}\mathbf{p}\\ \gamma\end{bmatrix} = 0$ where $G = G_1 - G_2$. Since $rank(G_1)$=$rank(G_2)$=1, G is generally a $rank$-2 matrix. It is solved by robustly fitting it to at least 15 relative depth pairs selected by a RANSAC scheme [2] and subsequently imposing $rank$-2 constraint onto potential solutions, similar to the algorithm in [4].

This $G_{4\times4}$ matrix is a bilinear constraint which absorbs the plane normal vectors ($\mathbf{n}$, $\mathbf{n}'$), camera internal matrices (M, M'), the third row of camera rotation ($\mathbf{r}_3$) and the third component of camera translation (t_3). It directly relates the 2D point measurements ($\mathbf{p}$, $\mathbf{p}'$) and their relative depth (γ, γ') without knowing the camera configuration and plane position, and furthermore the reference planes are not required to be the same. If $\mathbf{v} = \mathbf{v}'$, then G can be also applied to the cases in [5, 11, 12].

G relates the relative depth map $\mathbf{\Gamma}_t$ derived from $\mathbf{I}_t$, $\mathbf{I}_{t+1}$ and $\mathbf{\Gamma}_s$ derived from $\mathbf{I}_s$, $\mathbf{I}_{s+1}$ as long as the two depth maps share the same scene structure. If $s = t + 1$, G is related to the trifocal tensor [13] which relates the matching points within three frames, $\mathbf{I}_t$, $\mathbf{I}_{t+1}$ and $\mathbf{I}_{t+2}$. In this sense, G can be treated as the combination of trilinear constraint and plane homographies.

The variation between depth maps is defined as the absolute algebraic error w.r.t. G,

$$\delta_G(\mathbf{p}, \gamma, \mathbf{p}', \gamma') = \left|\begin{bmatrix}\mathbf{p}'^T & \gamma'\end{bmatrix} G_{4\times4} \begin{bmatrix}\mathbf{p}\\ \gamma\end{bmatrix}\right| \tag{11}$$

A depth variation map $\mathbf{\Delta}_\Gamma$ is then obtained by computing variation for each residual pixel.

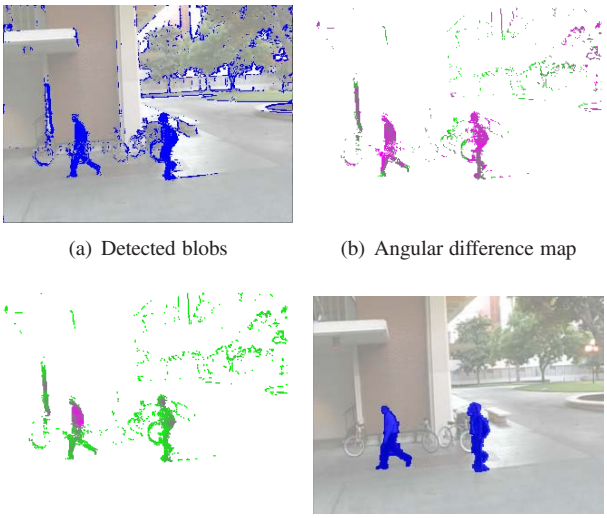
In Figure 4(c), we illustrate the variation of depth maps corresponding to the image as shown in Figure 4(a). The purple pixels indicate large depth variation and the green pixels indicate small depth variation.

4. Parallax Filtering within Sliding Windows

We have related the disparity of residual pixels to two geometric constraints, namely the epipolar constraint and structure consistency constraint. The disparity is used to filter out the parallax regions from the initial detection results. Unlike other proposed parallax filtering methods that directly threshold the disparity, we express the disparity values in probabilistic terms, namely likelihood functions. Then the filtering result is formulated as a likelihood map instead of a binary mask image.

The likelihood models of a pixel to belong to an independently moving object based on the geometric constraints are defined as follows:

$$L_\theta(\mathbf{p}_t^i) = \begin{cases}1 - \exp(1 - \alpha_\theta\theta_t^i) & \text{if } \theta_t^i \le \tau_\theta,\\ 1 & \text{if } \theta_t^i > \tau_\theta.\end{cases} \tag{12}$$



(a) Detected blobs (b) Angular difference map

(c) Depth variation map (d) Filtering results

Figure 4. Illustration of geometric constraint validation and parallax filtering process.

$$L_\delta(\mathbf{p}_t^i) = \begin{cases} 1 - \exp(1 - \alpha_\delta \delta_t^i) & \text{if } \delta_t^i \leq \tau_\delta, \\ 1 & \text{if } \delta_t^i > \tau_\delta. \end{cases} \quad (13)$$

where L_θ and L_δ are respectively the likelihood function based on angular difference θ_t^i and depth variation δ_t^i. α_θ and α_δ are positive weight factors. τ_θ is a cut-off threshold (typically $10°$) for angular difference. α_δ is a cut-off threshold for depth variation learned from practical data.

In addition, we propose to accumulate the geometric constraint errors within a sliding window (typically 5 frames) defined around a reference frame. The constraint errors estimated in current window are also utilized in overlapped windows to reduce the computational load.

The accumulated likelihood of point $\mathbf{p}_{t_0}^i$ in the reference frame $\mathbf{I}_{t_0}$ to belong to an independently moving object is defined as:

$$P_{Ind}(\mathbf{p}_{t_0}^i) = \sum_{t=t_0-w}^{t_0+w} e^{-\lambda(t-t_0)} [\eta L_\theta(\mathbf{p}_t^i) + (1-\eta) L_\delta(\mathbf{p}_t^i)] \quad (14)$$

where the corresponding pixels $\mathbf{p}_t^i$ in other frames $\mathbf{I}_t$ ($t \neq t_0$) are obtained by the optical flow mapping $\mathbf{\Omega}_{t_0}^t$. w is the half size of the window. λ is a positive weight factor and $\exp(-\lambda(t-t_0))$ enforces a large influence on the frames closer to t_0. η balances the influence of each likelihood model.

As we are interested in characterizing these moving regions or blobs in the image sequence, we could integrate this pixel-based information to its adjacent moving pixels using 8-connectivity. The likelihood of each detected moving blob B is obtained by integrating the likelihood of all the pixels within this region as follows,

$$P_{Ind}(B_{t_0}^j) = 1 - \exp\left(-\frac{1}{\left|B_{t_0}^j\right|} \sum_{\mathbf{p}_{t_0}^i \in B} P_{Ind}(\mathbf{p}_{t_0}^i) \right) \quad (15)$$

Thresholding the likelihood map is a straightforward method for removing most of parallax pixels and can achieve good results for specific videos, as shown in Figure 4(d). However, thresholding is inherently not flexible since the threshold value needs to be adjusted for each different sequence. Therefore, we propose to directly integrate the likelihood maps into the tracking algorithm as discussed in the next section, instead of the binary (0-1) mask images obtained by thresholding the likelihood maps.

5. Tracking by Spatial-temporal JPDAF

In [7], a JPDAF-based approach was proposed: it formulates the tracking problem as characterizing the position of the moving object that maximizes appearance and motion models. The optimal position at each time step depends on the current appearance observations, as well as the motion estimation obtained at the previous optimal positions. The classical JPDAF-based tracking approach produces local optimal solution since the decision made at time t is based only on current measurement and previous solution at time $t-1$. If a wrong estimation of the position is selected at time t, due to occlusions or to a bad detection, the tracking will not be able to recover the right solution later on.

In this section, we present a novel optimization method for extracting the optimal path by collecting discriminating evidences from past observations in the buffer. However, if the detected moving regions are due to parallax, this parallax information is propagated through tracking step, and it interferes with the extraction of accurate trajectories of the moving objects. In Section 4, we have proposed a refinement of detected moving blobs using the likelihood of a moving pixel to belong to a moving object in the scene or to parallax. We propose to integrate this blobs classification into the tracking algorithm, and show that we can improve the performance of the spatio-temporal JPDAF-based tracking in video sequences with a strong parallax.

In the proposed JPDAF, the optimal position of the moving object at each time step depends on several cues: the current appearance observation, the motion estimation and the blob's probability to belong to a moving object or parallax. Each cue is associated with a probability measure. We propose to define a joint probability reflecting current and past observations and define the appropriate data association (*i.e.* tracking) by maximizing this joint probability by collecting discriminating evidences in the sliding buffer.

The joint probability of a given position and bounding box at time t is given by:

$$\begin{aligned} & P(A^t, X^t, B^t, \cdots, A^0, \hat{X}^0, \hat{B}^0) \\ & = P(A^t|X^t)P(X^t|\hat{X}^{t-1}, \cdots, \hat{X}^0) \\ & \quad \cdot P(B^t|X^t, \hat{X}^{t-1})P_{total}(\hat{X}^{t-1}, \hat{B}^{t-1}) \\ & = P_{app}(A^t)P_{motion}(X^t)P_{Ind}(B^t)P_{total}(\hat{X}^{t-1}, \hat{B}^{t-1}) \end{aligned} \quad (16)$$

where A^t denotes the appearance of the moving object, X^t and $\hat{X}^t$ respectively denote the current and optimal positions of the moving object, and B_t and $\hat{B}^t$ denote respectively the current and optimal bounding boxes of the moving objects. The selection of the optimal path is guided by the following equation:

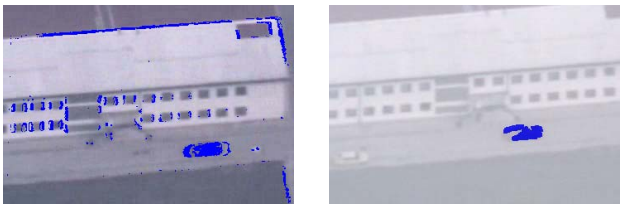
$$\begin{aligned} (\hat{\xi}, \hat{\psi}) = \arg\max(\mu \log P_{total}(\hat{X}^t, \hat{B}^t), \\ (1-\mu) \sum_{(\xi,\psi)} \log[\tilde{P}_{app}(\xi)\tilde{P}_{motion}(\xi)\tilde{P}_{Ind}(\xi)]) \end{aligned} \quad (17)$$

where $0 < \mu < 1$ is a weight factor used for increasing the confidence level of the refined path, $\hat{\xi}$ is the refined optimal path, ξ is the selected possible sub-optimal path, $\hat{\psi}$ is the refined set of bounding boxes, ψ is the selected possible sub-optimal bounding boxes, $\tilde{P}_{app}(\xi)$ and $\tilde{P}_{motion}(\xi)$ correspond respectively to the appearance, and motion probability along the path ξ, and $\tilde{P_{Ind}}(\psi)$ corresponds to the independence probability within the bounding box ψ. For further details, interested readers may refer to [7].

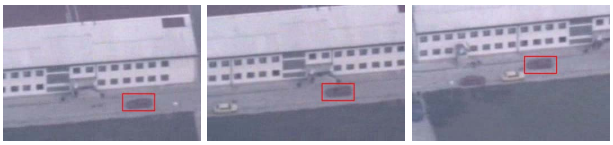
6. Experimental results

We present in this section some results obtained on real video sequences for illustrating the proposed detection and tracking in presence of strong parallax. The three video sequences include one video shot by ground-based cameras and two by airborne cameras. In Figure 5, we show a sequence of frames containing very strong parallax due to the large proximity of the camera to large structures in the scene. An initial detection result is presented in Figure 5(a), filtered detection result is presented in Figure 5(b), and the tracking results after removing parallax is presented in Figure 5(c). The parallax regions in the video, *i.e.* the building facade and static vehicles, are successfully removed.

Figure 6 presents another example which also contains a large amount of parallax caused by an UAV flying at very low altitude. In Figure 6(a), an initial detection result is presented, where the signs and marker lines on the road are incorrectly detected as motion. The filtered detection result is presented in Figure 6(b), and the tracking result after removing parallax is presented in Figure 6(c). As one can see, the proposed approach successfully filters regions due to parallax in both examples. Particularly, in Figure 6(c), many of small objects, which are near the epipole, are also successfully tracked although the variation of the depth maps is not significant around the epipoles by measuring the amplitude of the depth maps. Note that in this video, the camera is moving forward while the vehicles are also moving forward along the road. The epipolar constraint is not applicable in this situation and yet the structure consistency constraint still works.



(a) Initially detected blobs (b) Filtered blobs

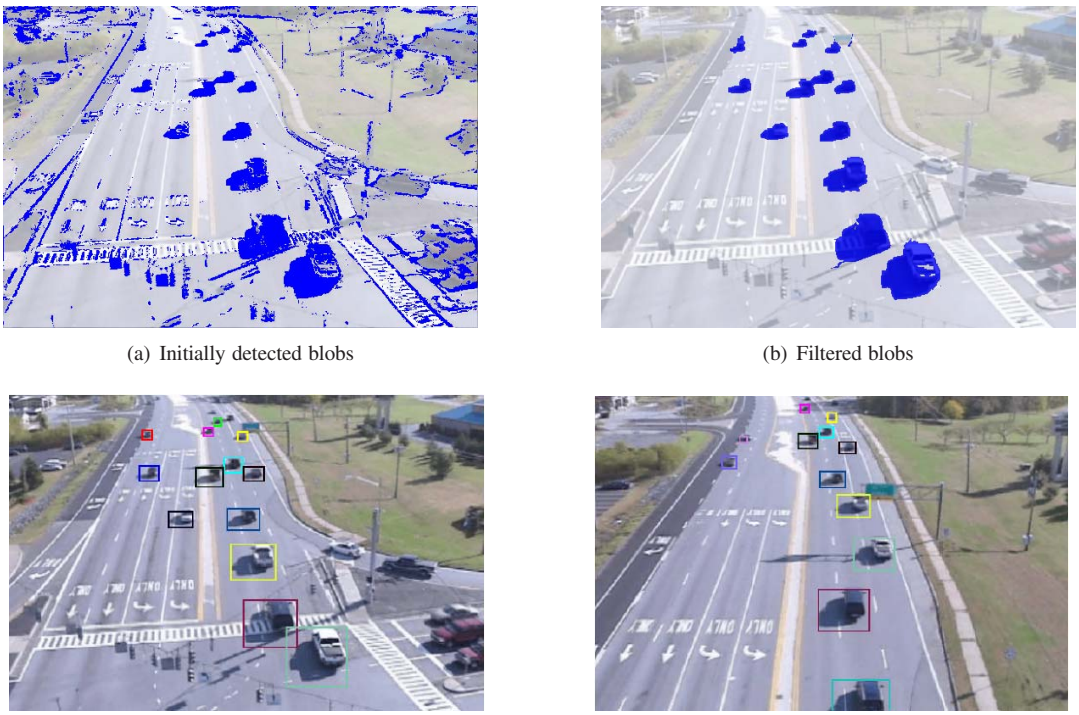


(c) Tracking results

Figure 5. Tracking in presence of strong parallax (with large structure)

7. Conclusion

We have presented a novel approach for detecting and tracking independent moving regions in presence of strong parallax. An affine motion compensation and detection framework generates the initial detected blobs which correspond to either parallax regions or independent motion. The proposed method filters out the parallax regions by analyzing two geometric constraints. The first constraint is the epipolar constraint represented by the angular difference between epipolar lines. The second constraint is the structure consistency constraint implemented as the variation between relative depth maps within "Plane+Parallax" framework. The variation of relative depth is obtained by a newly derived bilinear relationship. The disparity to the geometric constraints are represented in likelihood models and accumulated within a number of frames during parallax filtering. The accumulated likelihood maps are then integrated into the tracking framework. The use of a spatio-temporal JPDAF allows us to track moving objects in cases where the tracking provided by classical JPDAF was not optimal due to the lack of discriminating evidences. The defined JPDAF maximizes the joint probability by searching the optimal path across the nodes within a chosen buffer.



(a) Initially detected blobs

(b) Filtered blobs

(c) Tracking results

Figure 6. Tracking in presence of strong parallax (low altitude UAV)

The proposed approach does not fully employ 3D information (*e.g.* 3D motion trajectory) and we believe that integrating 3D information whenever available will improve tracking performance. Moreover, an in-depth analysis of the bilinear relationship between relative depths and a stable estimation method are of future interest to us.

Acknowledgements

This research was funded, in part, by the Advanced Research and Development Activity of the U.S. Government under contract # MDA-904-03-C-1786.

References

[1] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *IJCV*, 12(1):43–77, 1994.

[2] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *CACM*, 24:381–395, 1981.

[3] A. Fusiello, S. Caldrer, S. Ceglie, N. Mattern, and V. Murino. View synthesis from uncalibrated images using parallax. In *Proc. of ICIAP*, pp. 146–151, 2003.

[4] R. Hartley. In defence of the 8-point algorithm. In *Proc. of ICCV*, pp. 1064–1070, 1995.

[5] M. Irani and P. Anandan. A unified approach to moving object detection in 2d and 3d scenes. *IEEE-PAMI*, 20(6):577–589, 1998.

[6] M. Irani, P. Anandan, and M. Cohen. Direct recovery of planar-parallax from multiple frames. *IEEE-PAMI*, 24(11):1528–1534, 2002.

[7] J. Kang, I. Cohen, and G. Medioni. Continuous tracking within and across camera streams. In *Proc. of IEEE CVPR*, pp. 267–272, 2003.

[8] R. Kumar, P. Anandan, and K. Hanna. Direct recovery of shape from multiple views: A parallax based approach. In *Proc. of ICPR*, pp. 685–688, 1994.

[9] R. Pless, T. Brodsky, and Y. Aloimonos. Detecting independent motion: The statistics of temporal continuity. *IEEE-PAMI*, 22(8):768–773, 2000.

[10] H. Sawhney. 3d geometry from planar parallax. In *Proc. of IEEE CVPR*, pp. 929–934, 1994.

[11] H. S. Sawhney, Y. Guo, J. Asmuth, and R. Kumar. Independent motion detection in 3d scenes. *IEEE-PAMI*, 22(10):1191–1199, 2000.

[12] A. Shashua and N. Navab. Relative affine structure: Canonical model for 3d from 2d geometry and applications. *IEEE-PAMI*, 18(9):873–883, 1996.

[13] P. Torr. Geometric motion segmentation and model selection. *Philosophical Trans.: Mathematical, Physical and Engineering Sciences*, 356(1740):1321–1340, 1998.