Semantic Depth Map Fusion for Moving Vehicle Detection in Aerial Video

Mahdieh Poostchi, Hadi Aliakbarpour, Raphael Viguier, Filiz Bunyak, Kannappan Palaniappan
Computer Science Department, University of Missouri-Columbia
Columbia, MO, USA
mpoostchi@mail.missouri.edu
{aliakbarpourh, rvbb3, bunyak, palaniappan}@missouri.edu

Guna Seetharaman
US Naval Research Laboratory
Washington D.C., USA
guna@ieee.org

Abstract

Wide area motion imagery from an aerial platform offers a compelling advantage in providing a global picture of traffic flows for transportation and urban planning that is complementary to the information from a network of ground-based sensors and instrumented vehicles. We propose an automatic moving vehicle detection system for wide area aerial video based on semantic fusion of motion information with projected building footprint information to significantly reduce the false alarm rate in urban scenes with many tall structures. Motion detections are obtained using the flux tensor and combined with a scene level depth mask to identify tall structures using height information derived from a dense 3D point cloud estimated using multi-view stereo from the same source imagery or a prior model. The trace of the flux tensor provides robust spatio-temporal information of moving edges including the motion of tall structures caused by parallax effects. The parallax induced motions are filtered out by incorporating building depth maps obtained from dense urban 3D point clouds. Using a level-set based geodesic active contours framework, the coarse thresholded tall structures depth masks evolved and stopped at the actual building boundaries. Experiments are carried out on a cropped $2k \times 2k$ region of interest for 200 frames from Albuquerque urban aerial imagery. An average precision of 83% and recall of 76% have been reported using an object-level detection performance evaluation method.

1. Introduction

Aerial imaging provides a global picture of the traffic flow patterns over different time scales that captures large scale activity analysis of vehicles and pedestrians in urban settings. Airborne imagery enables understanding the simultaneous behavior of multiple drivers sharing the same road using multi-object tracking, covers a greater variety of interactions between road-users than would be encountered by any one single user, and facilitates routing around accidents to improve traffic flow [38, 42, 35, 33, 25]. City wide aerial imaging enables the collection of a broad range of road-user interaction behaviors between different categories of vehicles like private vehicles, public transportation vehicles, police cars, rescue vehicles, motorcyclists, construction vehicles, bicyclists, assisted/autopiloted vehicles, autonomous vehicles, pedestrians, animals and others including rare and infrequent behaviors and interactions. The analysis of rare events, accidents and unusual conditions due to weather, construction, public events, law enforcement, ambulances, etc. is facilitated using aerial video analytics of traffic at the individual vehicle level. Combining aerial with ground-based imaging and sensor information would be helpful in the development of rule-based reasoning engines for autopiloted and autonomous driving systems to better anticipate and predict the behavior of other agents in the environment and improve overall safety. Automatic moving object detection and segmentation are fundamental low-level computer vision tasks for many *traffic surveillance* applications including traffic monitoring [25], change detection [42, 27], classification [28, 41, 15], activity and behavior recognition [40, 29, 16, 9], and object tracking [45, 37, 11, 22, 39].

Detection and segmentation of objects in aerial imagery is impacted by many difficulties including small and low resolution targets, large moving object displacement due to low frame rate, congestion and occlusions, motion blur and parallax, camera vibration, camera exposure and varying viewpoints [13, 11, 34, 31] in addition to back-

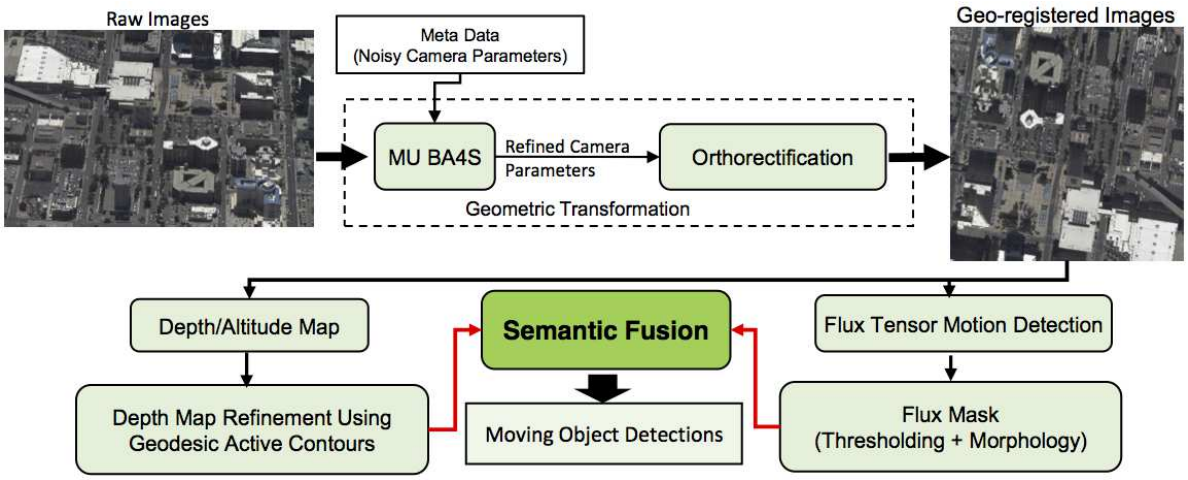


Figure 1. Semantic fusion-based moving vehicle detection system pipeline.

ground variance, illumination changes or shadow interference [24, 32, 36].

A typical moving object detection system follows either an appearance-based or motion-based approach to address these challenges. However, many of the moving vehicle detection algorithms typically focus on motion-based detection methods [7, 14] since appearance-based approaches [4, 39] are mostly computationally expensive especially when applied to large scale aerial imagery. Furthermore, recently different fusion schemes are proposed to combine the strength of each individual detection technique and improve system robustness [24, 21, 14, 43, 23, 17].

In this paper, we propose a scalable motion-based vehicle detection technique in dense urban scenes which fuses the spatio-temporal flow provided by the trace of the flux tensor with information about tall structures provided by a *depth map* also referred to as a *building* or *altitude map*, to filter out motion parallax induced flow responses and enhance robustness. We improve reliability of the depth mask filtering process

In order to avoid rejecting detected motion blobs associated with moving vehicles that are positioned next to the tall structures (and covered by building mask), the coarse thresholded building masks were guided and stopped at actual building boundaries using a level-set based geodesic active contour framework.

We used a state-of-the-art structure from motion (SfM) and registration algorithm called *MU BA4S* in order to orthorectify image sequences in a global reference system and maintain the relative movement between the moving camera platform and the fixed scene [2, 3]. Figure 1 illustrates our proposed semantic fusion-based moving vehicle detection system pipeline. The ultimate goal of our system is to achieve highly reliable motion detection to perform persistent tracking of moving vehicles over long time frames in large scale urban imagery.

Section 2 briefly reviews the applied orthorectification technique which is used to register the image sequence as well as extract the altitude maps, and describes the moving object detection approach using flux tensor. Section 3 elaborates on the fusion scheme and refinement of tall structures altitude mask. Finally, experimental results and conclusion are discussed in Section 4 and 5, respectively.

2. Moving Vehicle Detection in Aerial Video

2.1. Orthorectification

Videos in aerial imagery are captured on a moving airborne platform. Detection of moving objects, e.g. vehicles, in a scene which is observed by a camera that by itself has large movement and big jitters can be extremely challenging. To address this problem, images from camera planes are orthorectified (registered) in a global reference system to maintain the relative movement between the moving platform and the scene fixed. In this work the registration has been carried out by applying a homography transformation between each image plane and the ground dominant plane of the scene. Such homography transformations are analytically obtained using camera parameters, i.e. their rotation matrices and translation vectors. First the noisy camera parameters (referred to as *metadata*) obtained from platform sensors (i.e. IMU and GPS) are refined by a fast Structure-from-Motion algorithm(BA4S), proposed in [2, 3]. After the refinement process, the homography matrix between the ground plane π and the image plane of camera C is computed as follows:

$$ {}^{c}\mathbf{H}_{\pi} = \mathbf{K}\begin{bmatrix}\mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t}\end{bmatrix} \tag{1} $$

$\mathbf{r}_1$ and $\mathbf{r}_2$ being the first and second columns of the rotation matrix from the world's to the camera local coordinate system, $\mathbf{t}$ is the corresponding translation vector, and $\mathbf{K}$ is the camera calibration matrix. As a result, a 2D homogeneous image point $\tilde{\mathbf{x}}$ from camera C can project back on π as:

$$ {}^{\pi}\tilde{\mathbf{x}} = {}^{c}\mathbf{H}_{\pi}^{-1}\,\tilde{\mathbf{x}} = {}^{\pi}\mathbf{H}_{c}\,\tilde{\mathbf{x}} \tag{2} $$

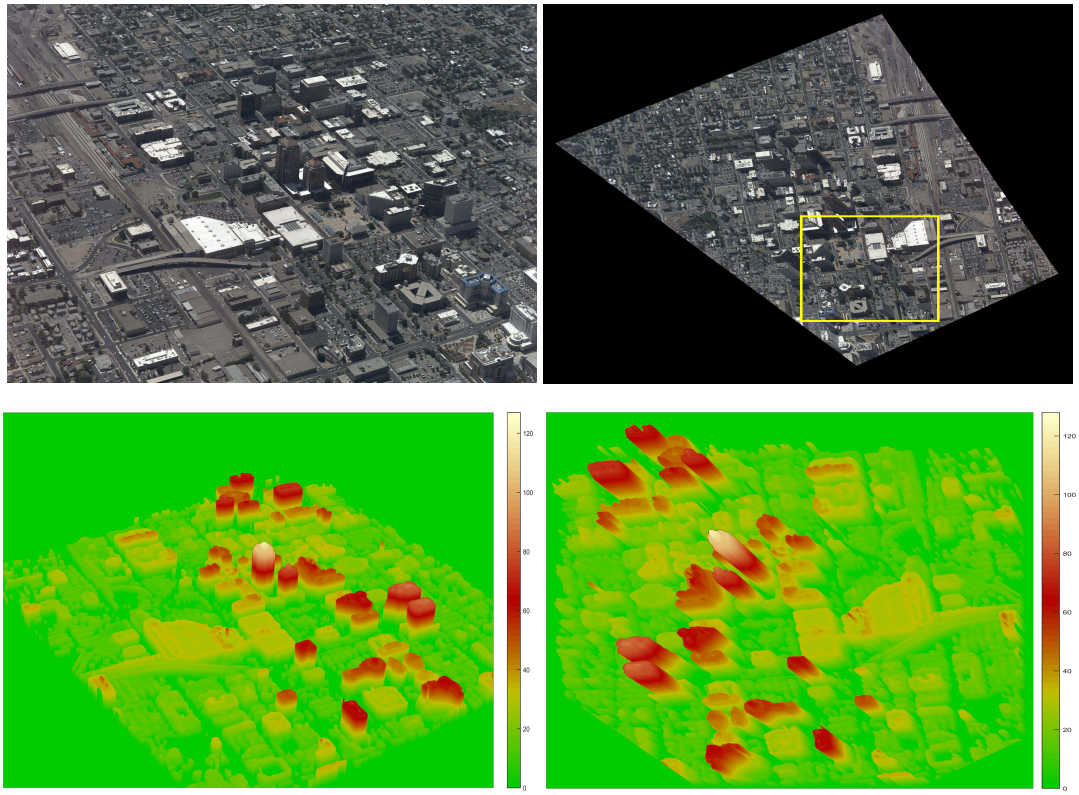


Figure 2. (UL) Illustrates raw ultra high resolution images ($\sim$ 30 MB, each) collected from an airborne platform flying over downtown Albuquerque-NM. (UR) shows the corresponding registered images exploiting *BA4S* orthorectification approach, (LL) corresponding raw image depth mask and (LR) shows the corresponding orthorectified depth mask.

where ${}^{\pi}\mathbf{H}_{c}$ is the inverse of the homography ${}^{c}\mathbf{H}_{\pi}$ in 1. Such a homography transformation is valid to transform all points between the image and ground plane, provided that their corresponding 3D point lies on the ground plane. Otherwise applying this homography transformation for points off the ground plane would cause a phenomenon known as *parallax* (see [20, 46] for more details).

Unlike to the image-to-image registration method in [26], in which the aerial images within each dataset could not be registered all together (they were broken to several segments, see Table-I in [26]), our method is able to efficiently register all the frames together with no fragmentations, thanks to the used robust optimization method (BA4S).

In this paper, we conducted our experiments on ABQ aerial imagery which were collected by TransparentSky over downtown Albuquerque, NM and ortho-rectified using *MU BA4S* state-of-the-art registration approach. Figure 2 shows samples of raw and geo-registered ultra high resolution images (6400×4400). However, we worked on a cropped $2k \times 2k$ region of interest for which the ground-truth are provided by Kitware (area bounded by yellow box in Fig. 5(UL)).

2.2. Depth Maps in Orthorectified Projections

The output of *MU BA4S* bundle adjustment includes refined camera parameters and a sparse 3D point cloud [2, 3, 1]. Sparse point clouds can be improved to produce dense point clouds using different multi-view stereo algorithms like PMVS[18], VisualSFM/Bundler[44] or probabilistic volume method described in [12]. The dense 3D point clouds are then used to produce *depth* or *altitude* maps by projecting the 3D points into each camera view. Each point is projected on a grid on the image plane, and after taking occlusion into account, each pixel in the grid is assigned at most one point in the point cloud. The altitude value of this point is used as the intensity for the corresponding pixel. This way we can produce a mask for each view such as the one shown in Figure 2(LL). Then same homography that was applied to original image can be used for the altitude mask to obtain the estimated altitude of every pixel in the orthorectified image, as shown in Figure 2(LR). Frame specific depth maps provide valuable information for distinguishing between motion detections due to motion parallax (motion of tall structures due to viewpoint changes) and moving objects on the ground. The semantic fusion method for combining depth and motion is discussed in more detail in Section 3.

2.3. Detecting Moving Objects and Strong Parallax Regions

The proposed framework uses *flux tensor* to extract motion blobs and to refine building mask. Flux tensor is presented as an extension of 3D structure tensor that allows reliable motion segmentation without expensive eigenvalue decomposition [30]. Following section briefly describes flux tensor structure and computations.

2.3.1 Flux Tensors

The *flux tensor* has been shown to be useful for detecting moving objects in computer vision and biomedical applications[6, 30]. In the current geospatial application we applied the flux tensor for moving object detection but we also found the flux tensor to be very useful for refining the automatically estimated structure of building footprints. The flux tensor was used to detect the parallax induced motion of the building structure that is in many cases more accurate than the multiview stereo based building depth edges.

Under constant illumination model, optic-flow equation of a spatiotemporal image volume $\mathbf{I}(\mathbf{x})$ centered at location $\mathbf{x} = [x, y, t]$ is

$$\begin{aligned} \frac{d\mathbf{I}(\mathbf{x})}{dt} &= \frac{\partial \mathbf{I}(\mathbf{x})}{\partial x}\, v_x + \frac{\partial \mathbf{I}(\mathbf{x})}{\partial y}\, v_y + \frac{\partial \mathbf{I}(\mathbf{x})}{\partial t}\, v_t \\ &= \nabla \mathbf{I}^T(\mathbf{x})\, \mathbf{v}(\mathbf{x}) \end{aligned} \tag{3}$$

taking the derivative of Eq. 3 with respect to t, we obtain Eq. 4

$$\begin{aligned} \frac{\partial}{\partial t}\left(\frac{d\mathbf{I}(\mathbf{x})}{dt}\right) &= \frac{\partial^2 \mathbf{I}(\mathbf{x})}{\partial x \partial t}\, v_x + \frac{\partial^2 \mathbf{I}(\mathbf{x})}{\partial y \partial t}\, v_y + \frac{\partial^2 \mathbf{I}(\mathbf{x})}{\partial t^2}\, v_t \\ &\quad + \frac{\partial \mathbf{I}(\mathbf{x})}{\partial x}\, a_x + \frac{\partial \mathbf{I}(\mathbf{x})}{\partial y}\, a_y + \frac{\partial \mathbf{I}(\mathbf{x})}{\partial t}\, a_t \end{aligned} \tag{4}$$

which can be written in vector notation as,

$$\frac{\partial}{\partial t}(\nabla \mathbf{I}^T(x)\mathbf{v}(\mathbf{x})) = \frac{\partial \nabla \mathbf{I}^T(\mathbf{x})}{\partial t}\, \mathbf{v}(\mathbf{x}) + \nabla \mathbf{I}^T(\mathbf{x})\, \mathbf{a}(\mathbf{x}) \tag{5}$$

where $\mathbf{v}(\mathbf{x}) = [v_x, v_y, v_t]$ is the optic-flow vector and $\mathbf{a}(\mathbf{x}) = [a_x, a_y, a_t]$ is the acceleration of the image brightness located at $\mathbf{x}$. Usually $\mathbf{v}(\mathbf{x})$ is estimated by minimizing Eq. 5 over a local 3D image patch $\mathbf{\Omega}(\mathbf{x}, \mathbf{y})$:

$$\frac{\partial \nabla \mathbf{I}^T(\mathbf{x})}{\partial t}\, \mathbf{v}(\mathbf{x}) + \nabla \mathbf{I}^T(\mathbf{x})\, \mathbf{a}(\mathbf{x}) = 0 \tag{6}$$

Assuming a constant velocity model subject to the normalization constraint $||\mathbf{v}(\mathbf{x})|| = 1$ and consequently zero acceleration, a least-squares error measure $e_{ls}(\mathbf{x})$ (Eq. 7) is used to minimize the Eq. 6

$$\begin{aligned} e_{ls}(\mathbf{x}) = \int_{\mathbf{\Omega}(\mathbf{x},\mathbf{y})} & \left(\frac{\partial(\nabla \mathbf{I}^T(\mathbf{y})}{\partial t}\, \mathbf{v}(\mathbf{x})\right)^2 d\mathbf{y} \\ & + \lambda\left(1 - \mathbf{v}(\mathbf{x})^T \mathbf{v}(\mathbf{x})\right) \end{aligned} \tag{7}$$

Differentiation of $e_{ls}(\mathbf{x})$ with respect to v, leads to eigenvalue decomposition problem $\mathbf{J_F}(\mathbf{x})\ \hat{\mathbf{v}}(\mathbf{x}) = \lambda\ \hat{\mathbf{v}}(\mathbf{x})$. The 3D flux tensor $\mathbf{J_F}$ for the spatiotemporal volume centered at (x, y) can be written in expanded matrix format as

$$\mathbf{J_F} = \begin{bmatrix} \int_{\mathbf{\Omega}} \left\{\frac{\partial^2 \mathbf{I}}{\partial x \partial t}\right\}^2 d\mathbf{y} & \int_{\mathbf{\Omega}} \frac{\partial^2 \mathbf{I}}{\partial x \partial t} \frac{\partial^2 \mathbf{I}}{\partial y \partial t} d\mathbf{y} & \int_{\mathbf{\Omega}} \frac{\partial^2 \mathbf{I}}{\partial x \partial t} \frac{\partial^2 \mathbf{I}}{\partial t^2} d\mathbf{y} \\ \int_{\mathbf{\Omega}} \frac{\partial^2 \mathbf{I}}{\partial y \partial t} \frac{\partial^2 \mathbf{I}}{\partial x \partial t}\, d\mathbf{y} & \int_{\mathbf{\Omega}} \left\{\frac{\partial^2 \mathbf{I}}{\partial y \partial t}\right\}^2 d\mathbf{y} & \int_{\mathbf{\Omega}} \frac{\partial^2 \mathbf{I}}{\partial y \partial t} \frac{\partial^2 \mathbf{I}}{\partial t^2} d\mathbf{y} \\ \int_{\mathbf{\Omega}} \frac{\partial^2 \mathbf{I}}{\partial t^2} \frac{\partial^2 \mathbf{I}}{\partial x \partial t} d\mathbf{y} & \int_{\mathbf{\Omega}} \frac{\partial^2 \mathbf{I}}{\partial t^2} \frac{\partial^2 \mathbf{I}}{\partial y \partial t} d\mathbf{y} & \int_{\mathbf{\Omega}} \left\{\frac{\partial^2 \mathbf{I}}{\partial t^2}\right\}^2 d\mathbf{y} \end{bmatrix} \tag{8}$$

The elements of the flux tensor incorporate information about temporal gradient changes which leads to efficient discrimination between stationary and moving image features. Thus the trace of the flux tensor matrix which can be compactly written and computed as, $\mathbf{trace}(\mathbf{J_F}) = \int_{\mathbf{\Omega}} ||\frac{\partial}{\partial t} \nabla \mathbf{I}||^2 d\mathbf{y}$ can be directly used to classify moving and non-moving regions without the need for expensive eigenvalue decompositions.

3. Fusion of Flux Tensor and Depth Maps

As described in Section 2.3, each pixel is classified as moving versus stationary by thresholding the trace of the corresponding flux tensor matrix ($\mathbf{trace}(\mathbf{J_F})$) at that pixel location. However, approximately 70% of the detected motions are induced by parallax effects from tall structures as the camera viewpoint changes (Fig. 5). We develop a context-based semantic fusion approach to identify and remove such non-vehicle detections by using the depth map information with an active contour boundary refinement and filtering process. As described in Section 2.2, the accurate height of every pixel in the orthorectified temporal frames can be estimated using 3D point clouds or meshes resulting from dense multiview 3D reconstruction algorithms. In order to produce a frame specific building mask, the 3D point cloud or mesh is projected to produce a depth map that is then thresholded. Image pixels with a height value greater than a threshold value are identified as part of tall structures or buildings which will be used to remove flux tensor motion responses (Fig. 6). Figure 3 illustrates the true motion detection produced by flux tensor (in yellow color) and undesirable moving detection caused by parallax in white color. The areas of tall structures are filtered by building mask and shown in blue. Provided ground-truth bounding-boxes are drawn in red to enable visual evaluation of the detection performance.

2D depth maps are projected from 3D point clouds that are obtained by 3D reconstruction of the scene (Section 2.2). These point clouds have lower resolution compared to the analyzed images. Low resolution combined with 2D projection inaccuracies may result in filtering out correctly detected vehicles positioned close to tall structure (zoomed

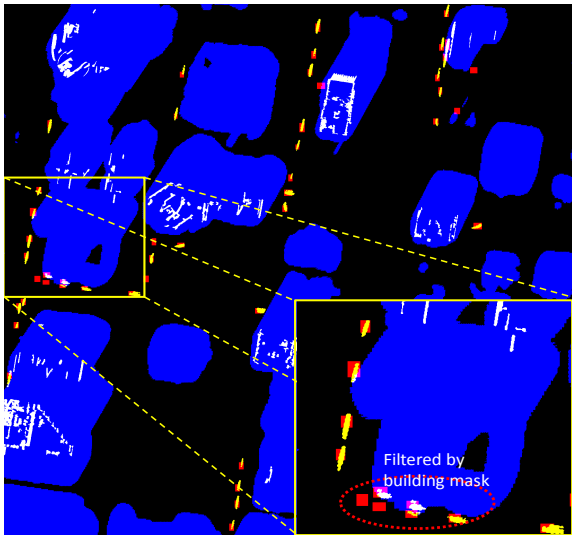


Figure 3. Illustration of motion detection results: true motion detection produced by flux tensor (in yellow color) and false detection caused by parallax in white color. The areas with high altitude are filtered by building mask and are shown in blue. Provided ground-truth bounding-boxes are shown in red.

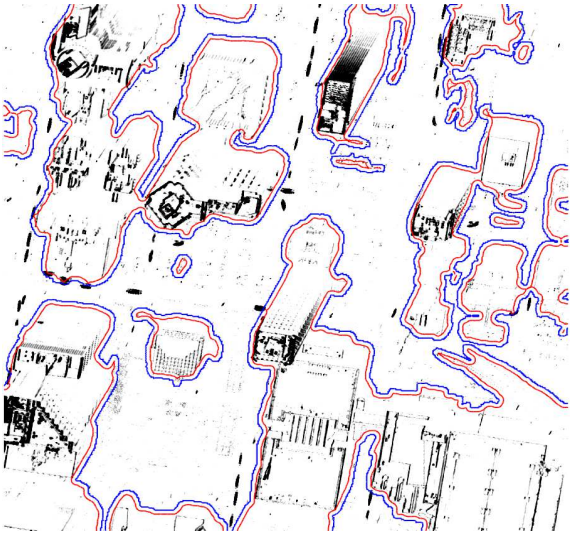


Figure 4. Improved building mask using level-set based geodesic active contours: blue lines are the initial building contours which are evolved and stopped at building actual boundaries (red lines).

in Fig. 3).

In order to refine the coarse building map B_{dmap}, we proposed to fuse the high resolution moving edges information from trace of the flux tensor with B_{dmap} through a level-set based geodesic active contours framework.

The trace of flux tensor is used to construct an edge indicator function g_F which will guide and stop the evolution of the geodesic active contour when it arrives at tall structure boundaries,

$$g_F(\mathbf{trace}(\mathbf{J_F})) = \frac{1}{1+\mathbf{trace}(\mathbf{J_F})} \tag{9}$$

The edge indicator function is a decreasing function of the image gradient that rapidly goes to zero along building edges and holds high values elsewhere.

Active contours evolve a curve $\mathcal{C}$, subject to constraints from a given image. In level set based active contour methods the curve $\mathcal{C}$ is represented implicitly via a Lipschitz function ϕ by $\mathcal{C} = \{(x,y)|\phi(x,y) = 0\}$, and the evolution of the curve is given by the zero-level curve of the function $\phi(t,x,y)$. Evolving $\mathcal{C}$ in a normal direction with speed F amounts to solving the differential equation [10],

$$\frac{\partial \phi}{\partial t} = |\nabla \phi| F; \quad \phi(0,x,y) = \phi_0(x,y) \tag{10}$$

Unlike parametric approaches such as classical snake, level set based approaches ensure topological flexibility since different topologies of zero level-sets are captured implicitly in the topology of the energy function ϕ. Topological flexibility is crucial for our application since we want to guide the coarse thresholded building mask to the actual building contours and reveal the filtered moving vehicles next to buildings. We used geodesic active contours [8] that are effectively tuned to trace of flux tensor edge information. The level set function ϕ is initialized with the signed distance function of the coarse building mask (B_{dmap}) and evolved using the geodesic active contour speed function,

$$\frac{\partial \phi}{\partial t} = g_F(\mathbf{trace}(\mathbf{J_F}))(c+\mathcal{K}(\phi))|\nabla\phi| + \nabla\phi \cdot \nabla g_F(\mathbf{trace}(\mathbf{J_F})) \tag{11}$$

where $g_F(\mathbf{trace}(\mathbf{J_F}))$ is the fused edge stopping function (Eq. 9), c is a constant, and $\mathcal{K}$ is the curvature term,

$$\mathcal{K} = div\left(\frac{\nabla\phi}{|\nabla\phi|}\right) = \frac{\phi_{xx}\phi_y^2 - 2\phi_x\phi_y\phi_{xy} + \phi_{yy}\phi_x^2}{(\phi_x^2+\phi_y^2)^{\frac{3}{2}}} \tag{12}$$

The force $(c+\mathcal{K})$ acts as the internal force in the classical energy based snake model. In this work, the constant velocity c pushes the curve inwards to the tall structures. The regularization term $\mathcal{K}$ ensures boundary smoothness. The external image dependent force $g_F(\mathbf{trace}(\mathbf{J_F}))$ is used to stop the curve evolution at building boundaries edges. The term $\nabla g_F \cdot \nabla\phi$ introduced in [8] is used to increase the basin of attraction for evolving the curve to the boundaries of the objects.

Figure 4 shows the improved building contours results in red. The blue line are the initial building contours which are evolved and stopped at building actual boundaries. As it can be seen, the previously filtered detected cars by initial building mask are revealed and counted as true detections.

4. Experimental Results

In this Section we elaborate and evaluate our proposed vehicle moving object detection results for ABQ aerial urban imagery which were collected by TransparentSky using an aircraft with on-board IMU and GPS sensors flying over downtown Albuquerque, NM. Figure 2 shows samples of raw ultra high resolution images (6400×4400) and the corresponding registered images using *MU BA4S* registration approach which processes the total sequence of 1071 images in very short amount of time. Refinement of camera parameters using *MU BA4S* on this dataset took less than 12 minutes. Inputs to the *MU BA4S* pipeline were the images accompanied by noisy camera parameters measured from onboard sensors. For the camera calibration matrix, a rough initial value of the focal length was fed to the *MU BA4S* with considering the principal points equal to the image centers. After camera parameters refinement, a 3D model (dense point clouds) of the scene was obtained (215 views were used) followed by projecting the elevation maps over each image, which totally took less than two hours. We worked on a cropped $2k \times 2k$ region of interest (ROI) for which the ground-truth are provided by Kitware (Fig. 5).

4.1. Moving Object Segmentation

The first input of our fusion scheme is the trace of flux tensor matrix which provides information about temporal gradient changes or moving edges. Figure 5 shows the original cropped ROI and the trace of flux tensor results. Every pixel is classified as moving versus stationary by thresholding trace of the corresponding flux tensor matrix. However, approximately and in average 70% of the detected motions are posed by parallax of tall structures which significantly degrade the precision of the motion detection results.

We incorporate the altitude information of tall structure to the flux tensor mask information in order to filter out the detected motion caused by parallax effect of tall buildings. Figure 6 presents the results of trace of flux tensor motion detection filtered by building mask. The left most image in Fig. 6 shows the flux tensor motion detection results in 2 colors; motion detections due to parallax are shown in white color and the rest are in yellow. In order to enable visual evaluation of the results ground truth bounding boxes are overlaid on flux tensor mask in red color. Altitude mask corresponding to the ortho-rectified ROI is shown in the middle. All the pixels with altitude values greater than 20 are considered as tall structures and are shown in blue in the rightmost image. As discussed in Section 3 level-set based geodesic active contours is used to improve the building mask and reveal the filtered moving vehicles positioned next to the buildings. Improved building mask and final motion detection results are shown in Figure 7.

4.2. Evaluation Methodology

Since the ultimate goal of the proposed motion detection system is to enable persistent tracking of moving vehicles, we used an object-level detection performance evaluation method. Associations of the detected moving blobs to ground truth objects is performed using a bidirectional correspondence analysis described in [5, 19]) that handles not only one-to-one matches but also merge and fragmentation cases. Precision and Recall are computed at each stage of fusion as

$$Precision = \frac{N_{TrueDetection}}{N_{Detection}} = \frac{|TP|}{|TP| + |FP|} \quad (13)$$

$$Recall = \frac{N_{TrueDetection}}{N_{GT}} = \frac{|TP|}{|TP| + |FN|} \quad (14)$$

where $N_{TrueDetection}$ or ***TP*** is defined as total number of true one-to-one individual matches plus the number of ground-truth fragmented objects and the number of merged detected objects. $N_{Detection}$ is the cardinality of detected objects and N_{GT} is the total number of moving object bounding boxes presented in ground-truth. We report $F_{measure}$ to evaluate the harmonic mean of recall and precision as well.

$$F_{measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (15)$$

Figure 8 reports the object level performance evaluation results. We improved the low 20% precision of trace of flux tensor only motion detection results to 83% and maintained high recall of 76% by fusing flux tensor motion masks with improved tall structures masks. Table 1 reports the average performance results obtained from each stage of fusion.

Table 1. Average Performance Results

Average	Flux Only	Flux+Depth	Flux+Depth+GAC
GT Objs	8214	8214	8214
TP	6400	6038	6241
FP	25384	928	1223
FN	1814	2176	1973
Precision	0.200	0.867	0.836
Recall	0.779	0.735	0.759
$F_{measure}$	0.318	0.796	0.796

5. Conclusion

We develop a novel context-based semantic fusion approach to detect moving objects in urban aerial imagery. Images were first orthorectified using a fast SfM method. Flux tensor has been used to extract motion blobs. It was shown that using purely conventional motion detection

Figure 5. Illustration of motion detection using trace of flux tensor only: From left to right, cropped ROI of Albuquerque aerial imagery (fr_{100}), the spatio-temporal motion information computed by trace of flux tensor for the selected image, and flux tensor mask in which each pixel is identified as moving or stationary by thresholding the trace of flux tensor. Morphology is applied to improve the result.

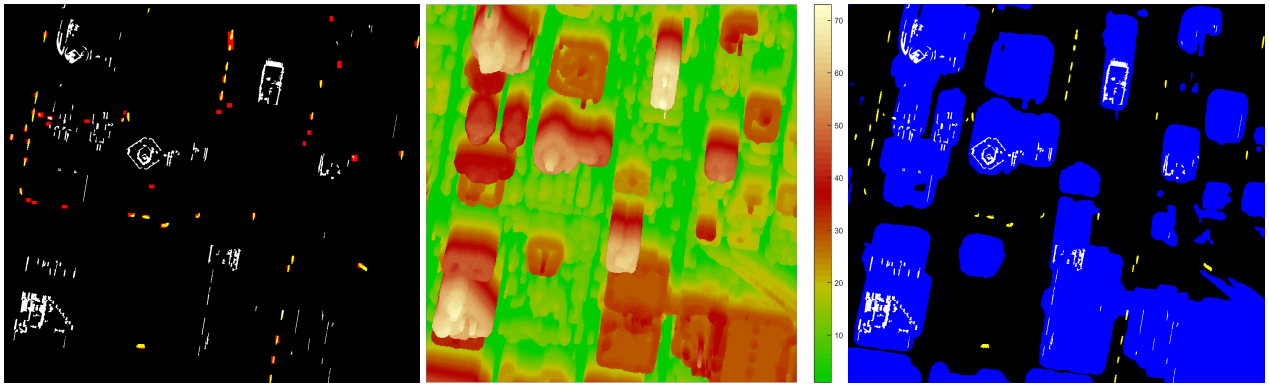


Figure 6. Illustration of motion detection results using trace of flux tensor filtered by depth mask. Left most image presents the motion detection results by thresholding the trace of flux tensor in 2 colors; motion detections due to parallax are shown in white besides other detection results in yellow color. In order to enable visual evaluation of the detection results ground-truth bounding boxes are overlaid on flux tensor mask in red color. Altitude mask corresponding to the ortho-rectified image is shown in the middle. All the pixels with altitude values greater than 20 meters are considered as tall structures and are shown in blue in the rightmost image.

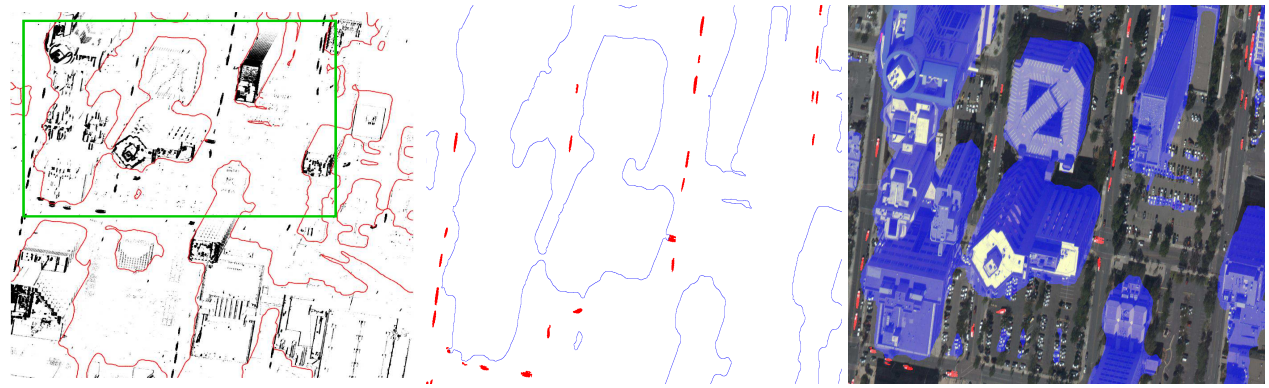
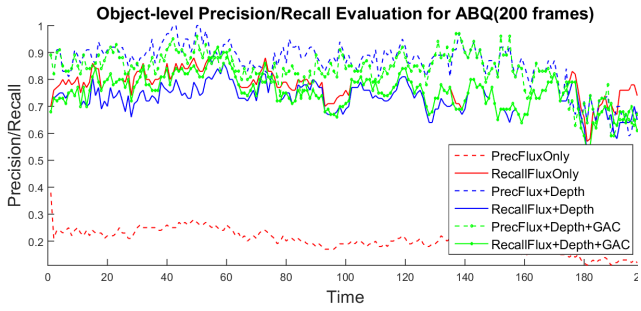
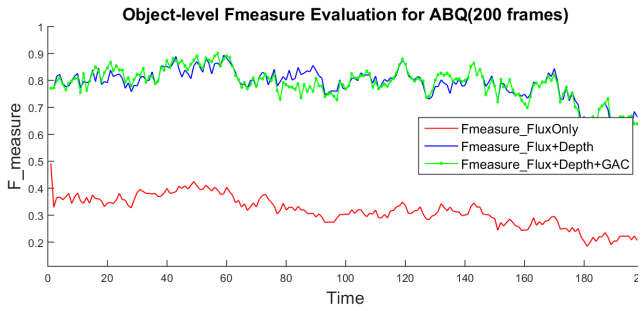


Figure 7. Moving object detection results using the proposed semantic fusion-based approach. The evolved building contours are shown in red in the left most image. The final moving object detection results of the region bounded in green box are shown in middle in red and final building contours in blue. Results are superimposed on the original image where building masks are shown in blue.

methods would not be sufficient for a wide area aerial imagery in which there are strong traces of parallax induced by tall buildings. In order to reject undesirable detections due to tall structures we used depth map information – obtained from the fast SfM followed by a dense 3D point clouds algorithm – in a boundary refinement and filtering processing stage. Using the proposed approach a high average precision of 83% and average recall of 76% have been achieved which promises a reliable persistent tracking.



(a) Precision and Recall



(b) $F_{measure}$

Figure 8. Object-Level performance evaluation of proposed fused moving object detection method.

Acknowledgments

This research was partially supported by the U.S. Air Force Research Laboratory under agreement AFRL FA875014-2-0072. Aerial WAMI of Albuquerque (ABQ) was collected by Transparent Sky, LLC in Edgewood, NM and provided by Steve Suddarth. Ground-truth for the 200 frames of the ABQ dataset was provided by Arslan Basharat at Kitware. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of AFRL, NRL, or the U.S. Government.

References

[1] H. Aliakbarpour, K. Palaniappan, and J. Dias. Geometric exploration of virtual planes in a fusion-based 3d registration framework. In *Proc. SPIE Conf. Geospatial InfoFusion III (Defense, Security and Sensing: Sensor Data and Information Exploitation)*, volume 8747, page 87470C, 2013.

[2] H. AliAkbarpour, K. Palaniappan, and G. Seetharaman. Fast structure from motion for sequential and wide area motion imagery. *Proc. IEEE International Conference on Computer Vision Workshop (ICCVW) Video Summarization for Large-scale Analytics Workshop*, Dec 2015.

[3] H. Aliakbarpour, K. Palaniappan, and G. Seetharaman. Robust camera pose refinement and rapid SfM for multiview aerial imagerywithout ransac. *IEEE Geoscience and Remote Sensing Letters*, 12(11):2203–2207, Nov 2015.

[4] A. Basharat, M. Turek, Y. Xu, C. Atkins, D. Stoup, K. Fieldhouse, P. Tunison, and A. Hoogs. Real-time multi-target tracking at 210 megapixels/second in wide area motion imagery. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–846, 2014.

[5] F. Bunyak, A. Hafiane, and K. Palaniappan. Histopathology tissue segmentation by combining fuzzy clustering with multiphase vector level sets. In *Software tools and algorithms for biological systems*, pages 413–424. Springer, 2011.

[6] F. Bunyak, K. Palaniappan, S. K. Nath, and G. Seetharaman. Flux tensor constrained geodesic active contours with sensor fusion for persistent object tracking. *J. Multimedia*, 2(4):20–33, Aug 2007.

[7] V. Carletti, P. Foggia, A. Greco, A. Saggese, and M. Vento. Automatic detection of long term parked cars. In *12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2015.

[8] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International journal of computer vision*, 22(1):61–79, 1997.

[9] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzàlez. Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, 116(3):396–410, 2012.

[10] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE transactions on Image processing*, 10(2):266–277, 2001.

[11] B.-J. Chen and G. Medioni. Motion propagation detection association for multi-target tracking in wide area aerial surveillance. In *IEEE Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2015.

[12] D. Crispell, J. Mundy, and G. Taubin. Parallax-free registration of aerial video. In *Proceedings of the British Machine Vision Conference*, pages 73.1–73.10. BMVA Press, 2008. doi:10.5244/C.22.73.

[13] I. Ersoy, K. Palaniappan, and G. Seetharaman. Visual tracking with robust target localization. In *IEEE Int. Conf. Image Processing*, pages 1365–1368, 2012.

[14] M. E. Farmer, X. Lu, H. Chen, and A. K. Jain. Robust motion-based image segmentation using fusion. In *IEEE International Conference on Image Processing*, volume 5, pages 3375–3378, 2004.

[15] R. Feris, R. Bobbitt, S. Pankanti, and M.-T. Sun. Efficient 24/7 object detection in surveillance videos. In *IEEE Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2015.

[16] P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, and N. Petkov. Car crashes detection by audio analysis in crowded roads. In *12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2015.

[17] K. Fragkiadaki, P. A. Arbeláez, P. Felsen, and J. Malik. Spatio-temporal moving object proposals. *arXiv preprint arXiv:1412.6504*, 2014.

[18] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010.

[19] A. Hafiane, F. Bunyak, and K. Palaniappan. Fuzzy clustering and active contours for histopathology image segmentation and nuclei detection. In *Advanced concepts for intelligent vision systems*, pages 903–914. Springer, 2008.

[20] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[21] P.-M. Jodoin and M. Mignotte. Motion segmentation using a k-nearest-neighbor-based fusion procedure of spatial and temporal label cues. In *Image Analysis and Recognition*, pages 778–788. Springer, 2005.

[22] M. Keck, L. Galup, and C. Stauffer. Real-time tracking of low-resolution vehicles for wide-area persistent surveillance. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 441–448, 2013.

[23] B. Lee, K. Yun, J. Choi, and J. Y. Choi. Robust pan-tilt-zoom tracking via optimization combining motion features and appearance correlations. In *IEEE Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2015.

[24] L. Li, A. Ellis, and J. Ferryman. On fusion for robust motion segmentation. In *Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2015.

[25] R. Lin, X. Cao, Y. Xu, C. Wu, and H. Qiao. Airborne moving vehicle detection for video surveillance of urban traffic. In *IEEE Intelligent Vehicles Symposium*, pages 203–208, 2009.

[26] M. E. Linger and a. A. Goshtasby. Aerial image registration for tracking. *IEEE Trans. Geosci. Remote Sens.*, 53(4):2137–2145, apr 2015.

[27] B. Morris and M. Trivedi. Robust classification and tracking of vehicles in traffic video streams. In *IEEE Intelligent Transportation Systems Conference*, pages 1078–1083, 2006.

[28] B. T. Morris and M. M. Trivedi. Learning, modeling, and classification of vehicle track patterns from live video. *IEEE Transactions on Intelligent Transportation Systems*, 9(3):425–437, 2008.

[29] B. T. Morris and M. M. Trivedi. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2287–2301, 2011.

[30] K. Palaniappan, I. Ersoy, and S. K. Nath. Moving object segmentation using the flux tensor for biological video microscopy. *Lecture Notes in Computer Science (PCM)*, 4810:483–493, 2007.

[31] K. Palaniappan, R. Rao, and G. Seetharaman. Wide-area persistent airborne video: Architecture and challenges. In *Distributed Video Sensor Networks: Research Challenges and Future Directions*, pages 349–371. Springer, 2011.

[32] D. H. Parks and S. S. Fels. Evaluation of background subtraction algorithms with post-processing. In *IEEE Advanced Video and Signal Based Surveillance*, pages 192–199, 2008.

[33] R. Pelapur, F. Bunyak, G. Seetharaman, and K. Palaniappan. Vehicle detection and orientation estimation using the radon transform. In *Proc. SPIE Conf. Geospatial InfoFusion III (Defense, Security and Sensing: Sensor Data and Information Exploitation)*, volume 8747, page 87470I, 2013.

[34] R. Pelapur, S. Candemir, F. Bunyak, M. Poostchi, G. Seetharaman, and K. Palaniappan. Persistent target tracking using likelihood fusion in wide-area and full motion video sequences. In *15th Int. Conf. Information Fusion*, pages 2420–2427, 2012.

[35] M. Poostchi, F. Bunyak, and K. Palaniappan. Feature selection for appearance-based vehicle tracking in geospatial video. In *Proc. SPIE Conf. Geospatial InfoFusion III (Defense, Security and Sensing: Sensor Data and Information Exploitation)*, volume 8747, page 87470G, 2013.

[36] M. Poostchi, K. Palaniappan, F. Bunyak, M. Becchi, and G. Seetharaman. Realtime motion detection based on the spatio-temporal median filter using gpu integral histograms. In *8th Indian Conference on Computer Vision, Graphics and Image Processing*, 2012.

[37] A. Prioletti, A. Mogelmose, P. Grisleri, M. M. Trivedi, A. Broggi, and T. B. Moeslund. Part-based pedestrian detection and feature-based tracking for driver assistance: real-time, robust algorithms, and evaluation. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1346–1359, 2013.

[38] F. Schubert and K. Mikolajczyk. Robust registration and filtering for moving object detection in aerial videos. In *Proceedings of the 2014 22Nd International Conference on Pattern Recognition*, ICPR '14, pages 2808–2813, Washington, DC, USA, 2014.

[39] S. Sivaraman and M. M. Trivedi. Integrated lane and vehicle detection, localization, and tracking: A synergistic approach. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):906–917, 2013.

[40] S. Sivaraman and M. M. Trivedi. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1773–1795, 2013.

[41] S. Sivaraman and M. M. Trivedi. Active learning for on-road vehicle detection: A comparative study. *Machine vision and applications*, 25(3):599–611, 2014.

[42] Z. H. Sun, M. Leotta, A. Hoogs, R. Blue, R. Neuroth, J. Vasquez, A. Perera, M. Turek, and E. Blasch. Vehicle change detection from aerial imagery using detection response maps. In *SPIE Defense+ Security*, pages 908906–908906. International Society for Optics and Photonics, 2014.

[43] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan. Static and moving object detection using flux tensor with split gaussian models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 414–418, 2014.

[44] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3057–3064, 2011.

[45] J. Xiao, H. Cheng, H. Sawhney, and F. Han. Vehicle detection and tracking in wide field-of-view aerial video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 679–684, 2010.

[46] C. Yuan, G. Medioni, J. Kang, and I. Cohen. Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1627–1641, 2007.