Moving Object Detection for Vehicle Tracking in Wide Area Motion Imagery Using 4D Filtering

Kannappan Palaniappan[*1], Mahdieh Poostchi[*1], Hadi Aliakbarpour[*1], Raphael Viguier[*1]
Joshua Fraser[*], Filiz Bunyak[*], Arslan Basharat[†], Steve Suddarth[‡]
Erik Blasch[§], Raghuveer M. Rao[¶], Guna Seetharaman[∥]
[*] CIVA Lab, Dept. of Computer Science, University of Missouri, Columbia, MO, 65201
Email: {palaniappank, mpoostchi, aliakbarpourh, rvbb3, jbf8cf, bunyak}@mail.missouri.edu
[†]Kitware, Inc., Clifton Park, NY 12065, Email: arslan.basharat@kitware.com
[‡]Transparent Sky, LLC, Edgewood, NM, 87015, Email: director@transparentsky.net
[§]Air Force Research Lab, Rome, NY, 13441, Email: erik.blasch@rl.af.mil
[¶]Army Research Laboratory, Adelphi, MD, USA, Email: raghuveer.m.rao.civ@mail.mil
[∥]US Naval Research Laboratory, Washington D.C., USA, Email: guna@ieee.org

Abstract—**Most Wide Area Motion Imagery (WAMI) based trackers use motion based cueing for detecting and tracking moving objects. The results are very high false alarm rates in urban environments with tall structures due to parallax effects. This paper proposes an accurate moving object detection method using a precise orthorectification approach for ground stabilization combined with accurate multiview depth maps to reduce the number of false positives induced by parallax effects by 90 percent. Proposed hybrid moving vehicle detection approach for large scale aerial urban imagery is based on fusion of motion detection mask obtained from median-based background subtraction and tall structures height mask provided by image depth map information. Using buildings mask enables us to improve the object level detection accuracy in terms of F-measure by 57 percent from 22.2% to 79.2%.**

I. INTRODUCTION

Automatic moving object detection is often a critical low-level task for many video analysis and tracking applications particularly when applied for large scale aerial imagery. Moving pixels can be further processed for different purposes including urban traffic monitoring [1], [2], object classification [3], registration and tracking [4]–[6].

Background subtraction methods are the most popular moving object detection methods. However, reliable background subtraction based moving object detection depends on several parameters such as illumination compensation, shadow detection, background dynamics and the background model learning rate. In aerial imagery, background modeling is particularly challenging due to small and low resolution targets, camera vibration which results in alignment errors and cause significant drifts in the pixel values close to sharp edges, and strong parallax effects [7]–[11].

Various techniques have been proposed to model the background including probabilistic mixture of Gaussian approach [12], Kalman filter [13], Wallflower [14] and simple median filtering [15]–[17]. Recently context-aware methods [18], [19] are presented that incorporate scene context like tall structure

height mask or road network to reduce false detections and improve system robustness.

In this work, we used spatiotemporal median-based background modeling due to its simplicity, computational efficiency and robustness to noise [1], [15]. Median filtering performs as well as more complicated techniques (i.e., GMM) and avoids creating unrealistic pixel values by blending pixel values [20]. The main contribution of this paper is high precision moving object detection to perform persistent tracking of moving vehicles. Our proposed hybrid moving vehicle detection system for aerial imagery consists of four main steps. In the first step, we applied our state of the art registration algorithm, *MU BA4S*, to orthorectify image sequences into a global reference system and to produce dense 3D point clouds [21], [22]. Then, *depth* or *height* maps are computed by projecting the 3D points into each camera view. In the third step, motion detection masks are obtained using median-based background subtraction. Finally, we fused building masks information extracted from depth maps with motion detection masks information to identify moving objects on the ground from motion induced by parallax effects of tall buildings and reject the false motion responses. We evaluated the performance of our motion detection method using pixel-level and object-level evaluation methods. Figure 1 presents our proposed hybrid moving vehicle detection system pipeline.

Section 2 briefly reviews our fast image registration approach. Hybrid motion detection system using depth map information is discussed in Section 3. Finally, experimental results and conclusion are presented in Section 4 and 5, respectively.

II. REGISTRATION

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

---
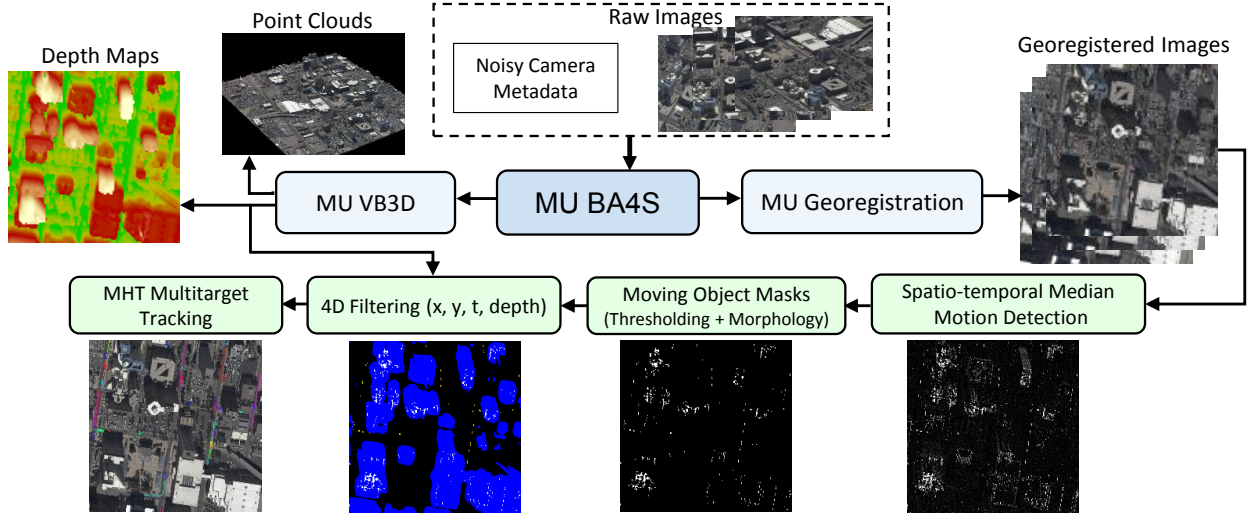
[1]These authors contributed equally to this work.

Fig. 1: Hybrid aerial video-based moving vehicle detection and tracking system pipeline using 4D (x, y, t, depth) filtering optimized for wide area motion imagery of urban scenes with significant parallax effects and small targets. Major modules that are essential for accurate motion detection include the University of Missouri (MU) BA4S [21], [22], video-based 3D reconstruction (VB3D) and orthorectification.

scene fixed. We used our fast *MU BA4S* registration technique in order to orthorectify aerial video and produce the images depth map. We assume that a scene and its dominant ground plane π is observed by an aerial camera C. A 3D scene point $\mathbf{X} = [x \quad y \quad z]^\intercal$ lying on this plane π is imaged as $\mathbf{x}$ by camera C. The imaged point $\mathbf{x}$ can be projected back (*orthorectified*) onto the ground plane π as ${}^\pi\mathbf{x}$ by applying a planar homography transformation.

In *MU BA4S* 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 [21], [22]. After the refinement process, the homography transformation 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 be projected back (*orthorectified*) onto the ground plane π using the following homography transformation:

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

where ${}^\pi\mathbf{H}_c$ is the inverse of the homography ${}^c\mathbf{H}_\pi$ in Eq 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 [8], [23] for more details).

In this paper, we conducted our experiments on ABQ aerial imagery which were collected by TransparentSky [24] over downtown Albuquerque, NM. Figure 2 shows samples of raw and georegistered 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 our collaborator at Kitware (Fig. 2(BR)).

III. SPATIOTEMPORAL MEDIAN MOTION ANALYSIS WITH 3D DEPTH FILTERING

Fast and accurate moving object detection provides primary useful information for a number of video analysis tasks. Background subtraction methods are the most popular methods to estimate the foreground moving objects [25], [26]. Moreover, one of the most commonly used non-recursive technique to obtain the background model is using spatiotemporal median information of a few frames [17], [27]. The most significant features of median background modeling over other complex methods are simplicity, computational efficiency and robustness to noise.

A. Spatiotemporal Median for Moving Blob Extraction

Median computation methods can be grouped as sorting-based or histogram-based approaches. Although the time complexity of histogram based method is much less, but it requires a 3D array of size $w \times h \times b$, where $w \times h$ is the size of the image and b is the number of histogram bins. Since we are working on high resolution large scale WAMI images, memory allocation for such histogram tensor is not feasible. Considering a 3 dimensional kernel with size $m \times n \times T$ around pixel centered at (x, y), the spatiotemporal median value for that pixel is obtained by

$$ M_{3D}(x,y) = \underset{\substack{\Delta t \in [-\frac{T-1}{2}, \frac{T-1}{2}] \\ \Delta y \in [-\frac{n-1}{2}, \frac{n-1}{2}] \\ \Delta x \in [-\frac{m-1}{2}, \frac{m-1}{2}]}}{\text{Median}} I(x + \Delta x, y + \Delta y, t + \Delta t) $$
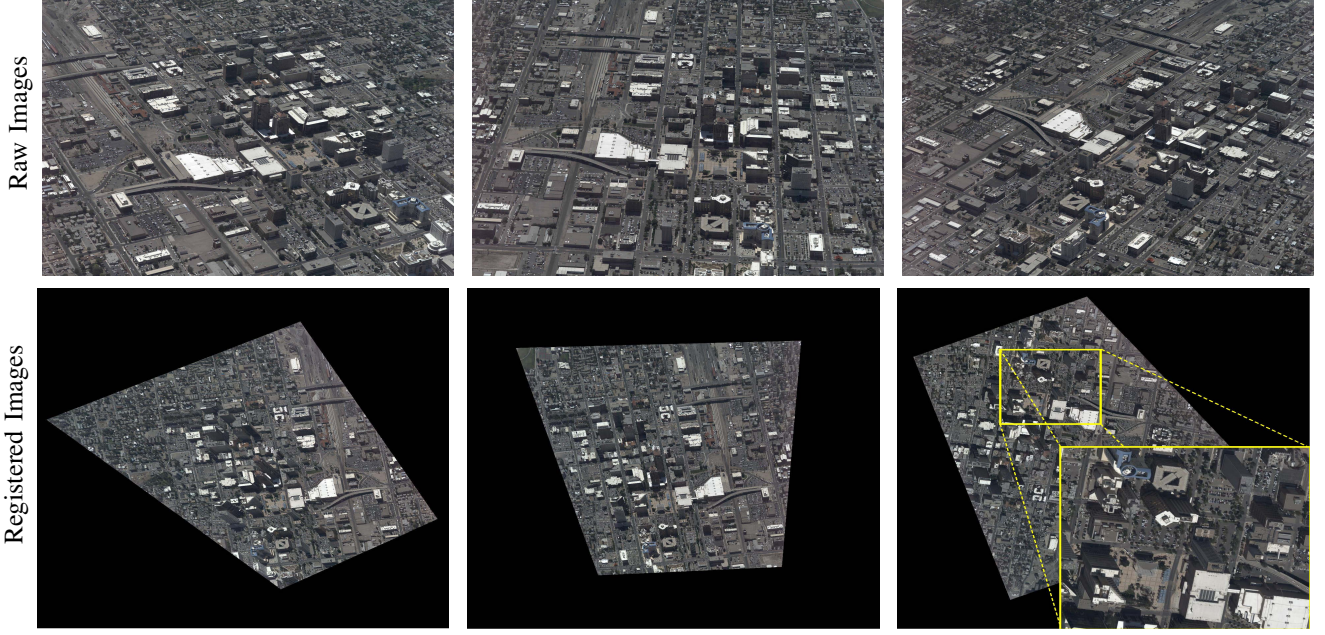
$$ \tag{3} $$

Fig. 2: Top row shows original ultra-high resolution images (6600 × 4400) collected from an airborne WAMI platform flying over downtown Albuquerque, NM. Bottom row shows corresponding registered images (12000 × 12000) using the MU BA4S orthrectification module, an extremely fast bundle adjustment algorithm that avoids RANSAC iterations, and uses less than 12 minutes for 1071 images [21], [22].

which means that for every T frames – considering also a $m \times n$ kernel for spatial median computation within an image – we compute a median background image BG_{Median}, and then perform background subtraction to estimate the moving foreground image $FG(k) = I(k) - BG_{Median}(k)$. Every pixel of estimated foreground is classified as moving versus stationary by thresholding the foreground image. However, as can be seen in the third row of Figure 5, prominent areas of moving object detection masks are undesirable motion responses that are caused by parallax effects of tall structures due to camera viewpoint changes. We proposed to reject such false alarms by building mask filtering. We fuse tall structures height information with estimated moving objects detection mask to suppress undesirable parallax effects.

B. Depth Map Filtering and Detection Refinement

The output of *MU BA4S* bundle adjustment includes refined camera parameters and a sparse 3D point cloud [21], [22], [28]. Sparse point clouds can be improved to produce dense point clouds using different multi-view stereo algorithms like PMVS [29] or probabilistic volume method described in [30]. The dense 3D point clouds are then used to produce *depth* or *height* maps by projecting the 3D points into each camera view. For each 3D point $[x \; y \; z]^t$, its corresponding 2D projection $[p_x \; p_y]^t$ is given by equation (4), where K, R, and t are respectively the camera calibration, rotation and translation matrices (Fig. 3). If several points projection have the same coordinates $[p_x \; p_y]^t$, the visible point is the one with the minimum depth s.

$$s * \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \qquad (4)$$
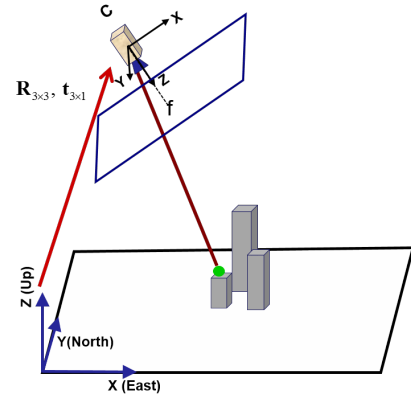

Fig. 3: Projection of a 3D scene onto a 2D camera plane. The calibration matrix **K** is dependent on the focal length **f**. **R** and **t** respectively account for the camera orientation and translation.

Then same homography that was applied to original image can be used for the height mask to obtain the estimated height of every pixel in the orthorectified image, as shown in Figure 4. This information can be used later to identify the moving objects on the ground from motion of tall structure due to viewpoint changes.

IV. 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 [24] using an aircraft with on-board IMU and GPS sensors flying 1.5 km above ground level of downtown Albuquerque, NM on September 3, 2013. Imaging was done at frame rate of 4Hz and 2.6 km orbit radius. Figure 2 shows samples of raw ultra high resolution images (6400 × 4400) with nominal ground
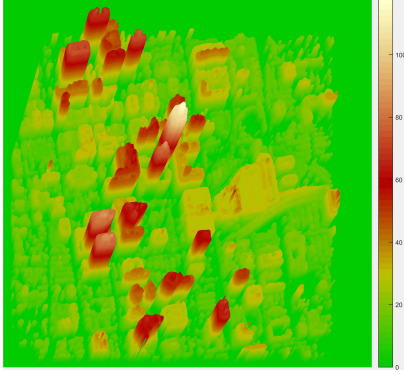
Fig. 4: Estimated height map of the orthorectified image.

TABLE I: Average Moving Object Detection Performance Results

Average	Object Level Evaluation		Pixel Level Evaluation	
	# Objects		Pixel Area	
	Median	Median+Depth	Median	Median+Depth
GT	7,980	7,980	18,412	18,412
Detection	50,090	7,128	127,890	12,150
TP	6,442	6,024	9,644	8,940
FP	43,648	1,104	118,250	3,210
FN	1,538	1,956	8,769	9,472
Precision(%)	12.9	83.9	7.56	73.2
Recall(%)	80.1	75.1	52.8	48.8
$F_{measure}$(%)	22.2	79.2	13.2	58.4

resolution of 25cm and the corresponding registered images using *MU BA4S* registration approach which processes the total sequence of 1071 images in very short amount of time (less than 12 minutes). For evaluation, we carried out the experiments on the first 200, 2000×2000 cropped frames from location $(4761, 5800)$ upper left corner in the $12K \times 12K$ orthorectified images for which the ground-truth are provided by Kitware (Fig. 5). Later, we produced the detection mask for full 1071 images to perform persistent multi-target tracking.

We model the background using median of seventeen images (eight frames before and eight frames after the target image), and perform background subtraction to estimate the moving foreground objects. Every pixel is classified as moving versus stationary by thresholding the estimated foreground image. Contrast enhancement and morphology operations are applied to improve the motion detection results and to filter out the spurious noises caused by illumination changes. Approximately 90.2% of the motion responses are induced by parallax effects of tall structures which significantly degrade the precision of the motion detection results. The low average precision of 12.9% is reported for the first 200 frames of Albuquerque sequence. In order to suppress false alarms due to parallax effects, we generate building masks by thresholding the image height maps. Any pixel with depth value above 20 meters belongs to tall structure areas. Detected Building areas are filtered out from motion masks to reject the motion detection of tall structures due to camera viewpoint changes (blue masks in figure 5, third row). Final moving vehicle detection masks of the region within yellow box are shown in forth row and superimposed on the original image for visual evaluation (last row, Fig. 5).

A. Evaluation Methodology

The general requirement for moving detection algorithms is providing reasonable precision in terms of the number of true detected objects as well as high recall in objects contour detection [7]. In this work, we improved the precision of the motion detection results while maintaining high recall by fusing median motion mask information and building mask information. The performance of motion detection results are evaluated after each stage of the fusion by computing the

spatial precision and recall (pixel-level evaluation) as

$$Precision = \frac{\sum_{i=1}^{N_D} |G_i \cap D_i|}{\sum_{i=1}^{N_D} |D_i|} = \frac{|TP|}{|TP| + |FP|} \quad (5)$$

$$Recall = \frac{\sum_{i=1}^{N_D} |G_i \cap D_i|}{\sum_{i=1}^{N_D} |G_i|} = \frac{|TP|}{|TP| + |FN|} \quad (6)$$

where G_i is the moving object bounding box presents in Ground Truth and D_i is the segmented moving object obtained by motion detection algorithm. N_D is the cardinality of the detected objects. Figure 6 reports the computed measures for first 200 frames from Albuquerque sequence. The average precision of 73.2% and recall of 48.8% is reported.

Since the ultimate goal of the proposed motion detection system is to perform persistent tracking of moving vehicles, we have evaluated detection performance using object level measures as well. Associations of the detected moving blobs to ground truth objects is performed using a bidirectional correspondence analysis described in [31], [32] that handles not only one-to-one matches but also merge and fragmentation cases. Average precision of 83.9% and recall of 75.1% is achieved. Table I reports the average of computed statistics for 200 frames. Figure 6 presents the performance evaluation results using pixel-level and object-level methods. It can be seen that precision has been drastically increased while recall remained almost the same using both evaluation methods.

V. CONCLUSION

We have developed a multi-component framework that fuses motion information with estimated 3D structure to reliably detect moving objects in aerial imagery of urban scenes. Moving object detection in wide-area aerial imagery is very challenging since fast camera motion prevents direct use of background subtraction methods and strong parallax induced by tall structures in the scene causes excessive false detections. The proposed framework first orthorectifies the images using our fast SfM method. Then, it extracts the motion blobs using robust 3D median-based background subtraction. Finally, it suppresses the false detections caused by parallax using projected depth map information obtained from our fast SfM followed by a dense 3D point clouds algorithm. The proposed framework has been tested on ABQ aerial urban imagery. We were able to improve the object level detection accuracy in terms of F-measure by 57 percent from 22.2% to 79.2%. These detection results are very promising for subsequent persistent object tracking in aerial imagery.
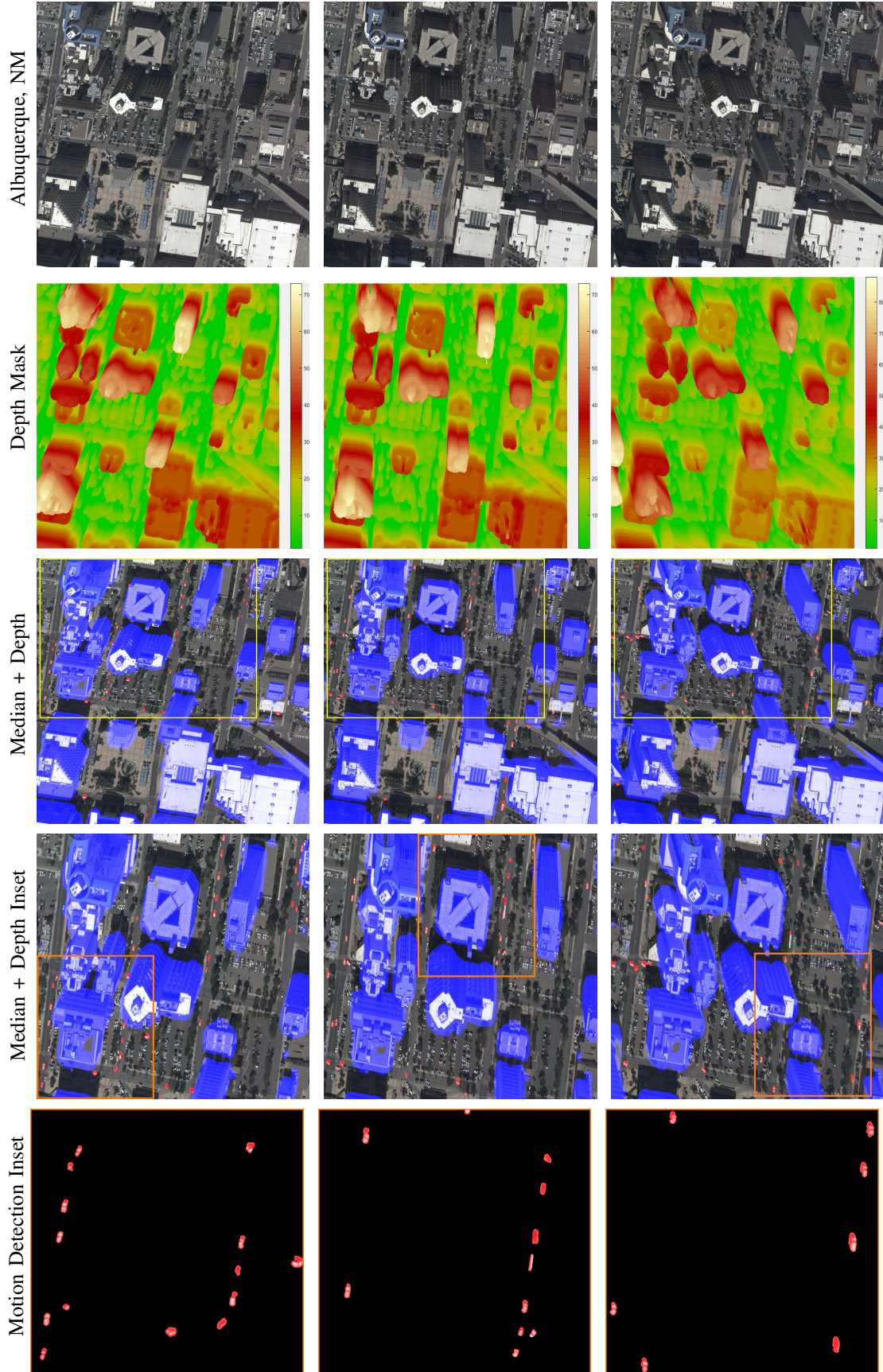
Fig. 5: Illustration of motion detection results: First row shows the registered $2k \times 2k$ cropped frames (57, 123, 182) from Albuquerque aerial imagery. Second row shows the height mask for the corresponding frames. Third row presents the motion detection masks obtained from median background subtraction and the building masks in blue. The final motion detection results are shown in forth row and superimposed on the original image (region within yellow box) for visual evaluation.
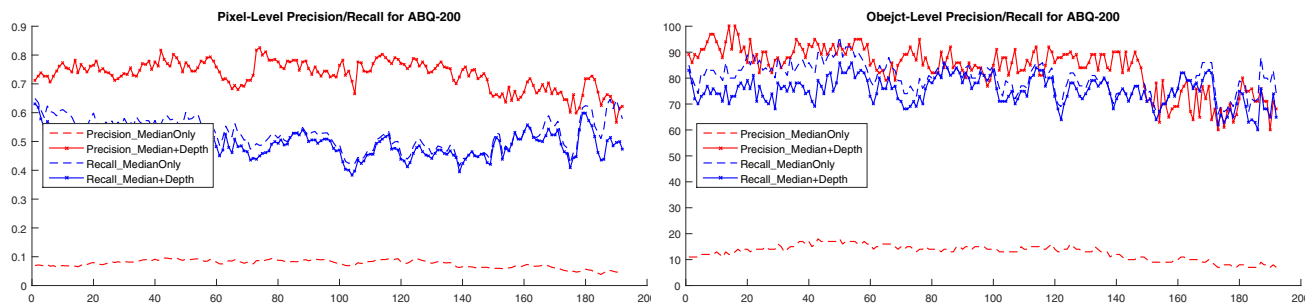
Fig. 6: Performance evaluation of our proposed fused motion detection method

REFERENCES

[1] S. C. Sen-Ching and C. Kamath, "Robust techniques for background subtraction in urban traffic video," in *Proc. SPIE Visual Communications and Image Processing*, vol. 5308, 2004, pp. 881–892.

[2] R. Feris, R. Bobbitt, S. Pankanti, and M.-T. Sun, "Efficient 24/7 object detection in surveillance videos," in *IEEE AVSS*, 2015, pp. 1–6.

[3] S. Sivaraman and M. M. Trivedi, "Active learning for on-road vehicle detection: A comparative study," *Machine Vision and Applications*, vol. 25, no. 3, pp. 599–611, 2014.

[4] F. Schubert and K. Mikolajczyk, "Robust registration and filtering for moving object detection in aerial videos," in *IEEE ICPR*, 2014, pp. 2808–2813.

[5] W. R. Thissell, R. Czajkowski, F. Schrenk, T. Selway, A. J. Ries, S. Patel, P. L. McDermott, R. Moten, R. Rudnicki, G. Seetharaman, I. Ersoy, and K. Palaniappan, "A scalable architecture for operational FMV exploitation (AVAA)," *IEEE ICCV Video Summarization for Large-scale Analytics*, 2015.

[6] R. Gargees, B. Morago, R. Pelapur, D. Chemodanov, P. Calyam, Z. Oraibi, Y. Duan, G. Seetharaman, and K. Palaniappan, "Incident-supporting visual cloud computing utilizing software-defined networking," *IEEE Trans. on Circuits and Systems for Video Technology*, 2016.

[7] S. Bhattacharya, H. Idrees, I. Saleemi, S. Ali, and M. , "Moving object detection and tracking in forward looking infra-red aerial imagery," in *Machine Vision Beyond Visible Spectrum*. Springer, 2011, pp. 221–252.

[8] 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 Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1627–1641, 2007.

[9] 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*, 2012, pp. 2420–2427.

[10] 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*. Springer, 2011, pp. 349–371.

[11] M. Poostchi, F. Bunyak, K. Palaniappan, and G. Seetharaman, "Feature selection for appearance-based vehicle tracking in geospatial video," in *SPIE Defense, Security, and Sensing*, 2013, pp. 87 470G–87 470G.

[12] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE CVPR*, vol. 2, 1999.

[13] J. Zhong and S. Sclaroff, "Segmenting foreground objects from a dynamic textured background via a robust kalman filter," in *IEEE ICCV*, 2003, pp. 44–50.

[14] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *IEEE CVPR*, vol. 1, 1999, pp. 255–261.

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

[16] 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.

[17] V. Reilly, H. Idrees, and M. Shah, "Detection and tracking of large number of targets in wide area surveillance," in *IEEE ECCV*, 2010, pp. 186–199.

[18] M. Poostchi, H. Aliakbarpour, R. Viguier, F. Bunyak, K. Palaniappan, and G. Seetharaman, "Semantic depth map fusion for moving vehicle detection in aerial video," *Proc. IEEE CVPR Workshop on Automatic Traffic Surveillance*, pp. 32–40, 2016.

[19] X. Shi, H. Ling, E. Blasch, and W. Hu, "Context-driven moving vehicle detection in wide area motion imagery," in *IEEE ICPR*, 2012, pp. 2512–2515.

[20] D. Gutchess, M. Trajkovics, E. Cohen-Solal, D. Lyons, and A. Jain, "A background model initialization algorithm for video surveillance," in *IEEE ICCV*, vol. 1, 2001, pp. 733–740.

[21] H. AliAkbarpour, K. Palaniappan, and G. Seetharaman, "Fast structure from motion for sequential and wide area motion imagery," *IEEE ICCV Video Summarization for Large-scale Analytics Workshop*, 2015.

[22] 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*, vol. 12, no. 11, pp. 2203–2207, Nov 2015.

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

[24] "http://www.transparentsky.net."

[25] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proceedings of the IEEE*, vol. 90, no. 7, pp. 1151–1163, 2002.

[26] E. Maggio and A. Cavallaro, *Video Tracking: Theory and Practice*. Wiley, 2011.

[27] S. Cheung and C. Kamath, "Robust techniques for background subtraction in urban traffic video," in *Proc. SPIE Visual Communications and Image Processing*, vol. 5308, no. 1, 2004, pp. 881–892.

[28] 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)*, vol. 8747, 2013, p. 87470C.

[29] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.

[30] D. Crispell, J. Mundy, and G. Taubin, "Parallax-free registration of aerial video," in *BMVC*, 2008, pp. 73.1–73.10.

[31] 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*, 2011, pp. 413–424.

[32] 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*, 2008, pp. 903–914.