Geo-spatial Aerial Video Processing for Scene Understanding and Object Tracking

Jiangjian Xiao Hui Cheng Feng Han Harpreet Sawhney
Sarnoff Corporation
(jxiao, hcheng, fhan, hsawhney)@sarnoff.com

Abstract

This paper presents an approach to extracting and using semantic layers from low altitude aerial videos for scene understanding and object tracking. The input video is captured by low flying aerial platforms and typically consists of strong parallax from non-ground-plane structures. A key aspect of our approach is the use of geo-registration of video frames to reference image databases (such as those available from Terraserver and Google satellite imagery) to establish a geo-spatial coordinate system for pixels in the video. Geo-registration enables Euclidean 3D reconstruction with absolute scale unlike traditional monocular structure from motion where continuous scale estimation over long periods of time is an issue. Geo-registration also enables correlation of video data to other stored information sources such as GIS (Geo-spatial Information System) databases. In addition to the geo-registration and 3D reconstruction aspects, the key contributions of this paper include: (1) exploiting appearance and 3D shape constraints derived from geo-registered videos for labeling of structures such as buildings, foliage, and roads for scene understanding, and (2) elimination of moving object detection and tracking errors using 3D parallax constraints and semantic labels derived from geo-registered videos. Experimental results on extended time aerial video data demonstrates the qualitative and quantitative aspects of our work.

1. Introduction & related work

Interpretation of aerial video data for scene and object level understanding is an important problem domain in today's world since large areas of the world are being captured from the air for both commercial and military applications. In particular, videos captured using low-flying aerial platforms consists of strong parallax from non-ground-plane structures [6, 12], which contains rich information about the 3D nature of the scene as well as moving objects. By leveraging the 3D cue implied in the videos, this paper presents an approach to extracting and using semantic layers from low altitude aerial videos for scene understanding and object tracking. In our approach, one key aspect is the use of geo-registration between video frames and reference image to establish a geo-spatial coordinate system for pixels in the video. Geo-registration enables Euclidean 3D reconstruction with absolute scale unlike traditional monocular structure from motion where continuous scale estimation over long periods of time is an issue. Geo-registration also enables correlation of video data to other stored information sources such as GIS databases, which will provide another set of cues for scene segmentation. In addition to the geo-registration and 3D reconstruction aspects, the key contributions of this paper also include: (1) exploiting appearance and 3D shape constraints derived from geo-registered videos for labeling of structures such as buildings, foliage, and roads for scene understanding, and (2) elimination of moving object detection and tracking errors using 3D parallax constraints and semantic labels derived from geo-registered videos.

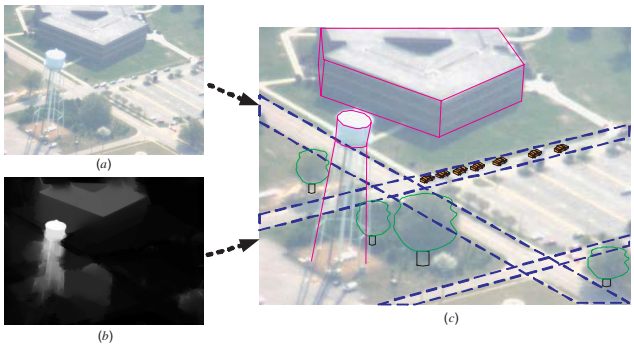


Figure 1. Scene segmentation of aerial video. After combining multi-cue from video, geo-reference image and GIS information, the input video will be partitioned into different meaningful layers, such as building, road, tree, and cars, for video understanding and event analysis. (a) One input frame. (b) Corresponding depth cue estimated from the video. (c) The desired scene segmentation results.

Scene understanding using static image segmentation

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

has been studied extensively and a number of methods have been developed to relate scene structure to semantics [17, 2]. In those approaches, typically manually segmented images are collected as training data to learn a set of specified object clusters, such as trees, grass, roads, buildings, and cars, etc. During the learning process, a range of features, such as color histogram, texture, gradient, lines and curves, are extracted from the images for the model clustering. However, without the help of depth cues, these methods face a serious challenge to infer building structures by either rectangle or parallelogram fitting since a number of man-made ground structures, such as roads, parking lots, and parks have similar geometry as buildings. In photogrammetry community, some researchers combine high resolution DEM (Digital Elevation Model) or LIDAR (Light Detection and Ranging) data with/or without aerial images to recover building structures [4, 3, 18].

Compared to the existing image-based scene segmentation and DEM-based urban reconstruction methods, our approach exploits the information implicit in the aerial video sequence and the associated geo-reference image. We leverage the estimated depth and motion cues with 2D image features to segment video frames into semantic regions. With the help of depth cue, our segmentation approach effectively partitions each video frame into ground and non-ground regions for building and tree detection, and also achieve the consistent scene segmentation results over the frames by enforcing spatiotemporal constraints.

Fig. 1 shows the concept of our geo-based video scene segmentation for both scene understanding and object tracking. Given an input aerial video sequence, we first perform a geo-referenced depth estimation to compute depth for each frame. Then based on the estimated depth, the non-ground regions are segmented and a planar fitting plus depth extension approach is applied to extract the structure of buildings and tree shapes. In our approach, the estimated depth does not only help for the structure detection but also can effectively reduce false alarms in object tracking related to 3D parallax. Moreover, we also integrate GIS information into our framework to detect road network, which further partitions the video frame into blocks and assist tracking association along the road network even when vehicles stopping, occluded or making a turn.

Current approaches to aerial video process are mainly focused on moving object tracking and do not largely exploit scene context from video data [11, 1, 20]. [10] use video scene segmentation to reduce tracking false alarms. These work mostly focus on image based segmentation, where only texture and color information are used for system training. This approach is effective in removing false alarms typically from the trees but unfortunately it ignores the strong 3D cue implied in the video sequence. In the area of aerial video based depth estimation, the most relevant work is [15], where a set of pushbroom mosaics are created to estimate depth. However, the construction of pushbroom mosaic needs a strict requirement on the aerial trajectory and camera setup: the camera should be setup as a top view with a straight flight path as possible. In our case, the aerial platform and the camera can follow an arbitrary trajectory and the camera zoom setting may change over time.

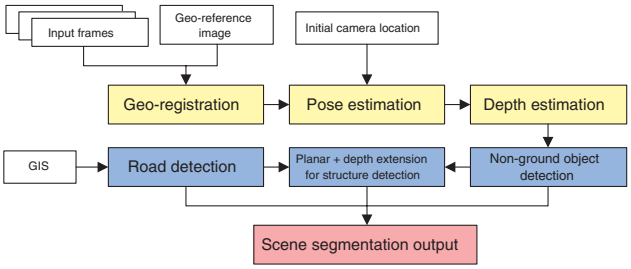


Figure 2. Our two-stage algorithm framework. The first stage is the geo-based depth estimation algorithm, and the second stage is multi-cue scene segmentation algorithm.

The remainder of this paper is organized as follows. Section 2 provides an overview of our algorithmic framework. Section 3 addresses geo-referenced camera pose and depth estimation. In section 4, an approach is presented to extract the structure of buildings, roads and trees with the detected moving objects from the videos by integrating motion, depth, color, texture and GIS information. The experimental results are reported in Section 5.

2. Algorithm overview

Fig. 2 shows the algorithm framework of our approach that includes two main stages: (1) Geo-spatial depth estimation from monocular aerial video, and (2) Multi-cue scene segmentation. In the first stage, we employ the geo-referenced imagery from Terraserver/Google to perform ortho-rectification of video frames and recover metric depth of the scene. The process includes geo-registration, pose estimation and depth recovery. Once depth maps are estimated, the second stage process partitions each frame into ground and non-ground regions based on depth variation. For the non-ground regions, a planar fitting and depth extension approach is designed to identify building structures by integrating depth, color, and texture information and segment trees and other foliage in the video. For the ground regions, additional GIS information is used to extract and refine geo-registered road network. Finally, all extracted regions are combined with object tracking results and visualized as the scene segmentation for video understanding and event analysis.

3. Depth estimation from arial video

Depth estimation from monocular aerial videos is in general a challenging problem as the internal camera parameters can change over time (due to variable zooming), scale needs to be continuously estimated, and the quality of imagery is also highly variable due to blurring and illumination changes. Furthermore, typical imaging scenarios include platforms flying high enough that the camera model often degenerates to an affine camera with the scene being largely planar with some depth variation due to buildings, foliage and terrain. Traditional frame-to-frame structure from motion methods are generally unreliable under these conditions.

We employ geo-referenced reference imagery available from open sources such as Terraserver and Google maps to perform video frame to reference imagery pose estimation. By matching features in video frames to features in reference imagery, metric pose estimation is possible. Since each point in geo-referenced imagery has an associated world coordinate (latitude, longitude), the pixel correlated with a point in the reference imagery inherits that coordinate. This becomes the basis of refining pose estimates for each frame and Euclidean depth estimation by frame-to-frame correlation. The flowchart in Fig. 3 shows the detail steps in our geo-referenced camera pose and depth estimation process.

3.1. Geo-registration

Geo-reference image databases typically consist of ortho-rectified imagery for the flat ground with each latitude-longitude specified for each point. Optionally it is possible to utilize digital elevation maps (DEMs) but these are not available in public databases such as Terraserver and Google. We use geo-registration to match features on the ground in videos to the reference imagery. No 3D pose estimation is done at this stage. Our approach employs SIFT feature [9] to detect a number of correspondences between the input video frames and geo-reference image for an initial alignment. Then a topology-based bundle adjustment process is applied to refine the registration process by incorporating the frame-to-frame transformation within the sequence [14, 13]. During geo-registration, we also apply a third-order lens distortion model to remove the radial distortion [5] such that

$$\hat{p} = (1 + k_1 r + k_2 r^2 + k_3 r^3)\tilde{p}, \tag{1}$$

where $\tilde{p}$ is the ideal image location of geo-correspondences, $\hat{p}$ is the actual image position, and r is the radial distance. Fig. 4 shows one geo-registration example, where the track and GIS information are also overlapped on the image.

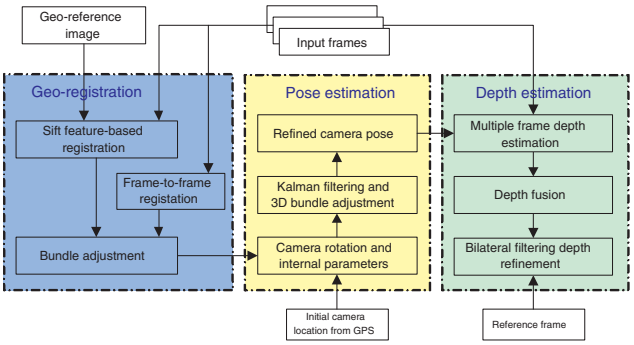


Figure 3. The flow chart of our geo-based depth estimation algorithm, where three major modules are required in the process.

3.2. Camera calibration and pose estimation

After frame-to-reference geo-registration, we re-estimate camera poses based on a set of initial camera locations provided by the on-board GPS sensor. We assume that the geo-reference coordinates are located on the ground plane with height Z_0. Then based on 3D projection, a 3D ground point $P = [X\ Y\ Z_0\ 1]'$ can be projected on frame I_j at $p = [u\ v\ 1]'$ such that

$$\begin{aligned} p &= K_j R_j [I|-C_j] \begin{bmatrix} X \\ Y \\ Z_0 \\ 1 \end{bmatrix} \\ &= K_j R_j \begin{bmatrix} 1 & 0 & -X_j^c \\ 0 & 1 & -Y_j^c \\ 0 & 0 & Z_0 - Z_j^c \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \\ &= K_j R_j A \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}, \end{aligned} \tag{2}$$

where K_j is the camera calibration matrix for frame I_j, R_j is the corresponding rotation matrix, $C_j = [X_j^c\ Y_j^c\ Z_j^c]'$ is the initial camera center given by GPS. Similarly, we can also project the video frame into the geo-coordinates through the projective geo-registration transformation such that

$$p = H_j \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}, \tag{3}$$

where H_j is a projective transformation between frame I_j and geo-reference image. By comparing Eq. 2 with Eq. 3, we obtain

$$\begin{aligned} K_j R_j A &= H_j \\ K_j R_j &= A^{-1} H_j = M, \end{aligned} \tag{4}$$

where M is only depended on H_j and camera center location. Once M is computed, we can apply QR decomposition or SVD decomposition to estimation rotation matrix

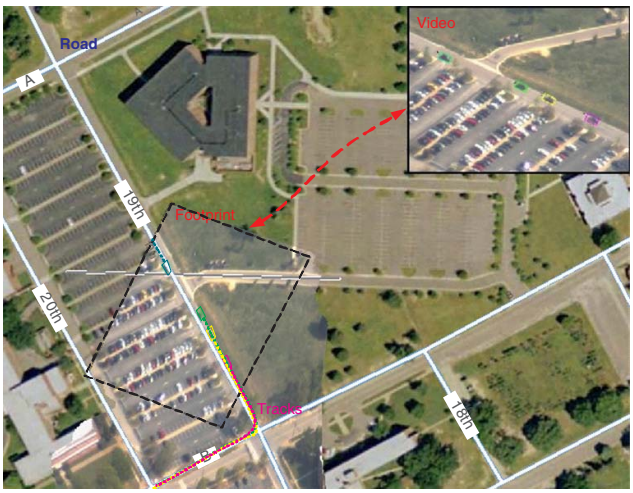


Figure 4. One frame geo-registration with track and GIS information overlapped.

R_j and calibration matrix K_j for each frame I_j. In the experiments, we found that SVD decomposition can tolerate lens distortion and produce better orientation results than QR decomposition that enforces K_j as an upper triangle matrix. The SVD decomposition is given as follows:

$$M = UDV' = (UDU')(UV') = K_j R_j, \tag{5}$$

where $K_j = UDU'$ and $R_j = UV'$. The resulted K_j may not be a strict upper triangle matrix due to the image noise and remaining lens distortion, but it can be further enforced to an upper triangle matrix by parameter fitting[21, 5].

After estimating the camera calibration and rotation matrices for each frame, we apply a Kalman filter to smooth the calibration matrix and rotation angles by assuming the camera mechanical change is continuous over consecutive frames. Then, based on the refined K_j, R_j, and the correspondences between each video frame to geo-reference image, the camera location C_j can be re-estimated from Eq. 2 with 3D bundle adjustment[5].

3.3. Depth estimation and depth fusion

In the geo-referenced world coordinates, X and Y axes are along east and north respectively, and Z axis is perpendicular to the ground plane. We quantize depth along the Z axis to represent depth layers along this axis with a total of 50 layers. We apply the multi-frame graph cut algorithm [8, 19] to estimate depth from videos. Due to the varying image quality of video frames, the quality of the estimates depth maps is highly variable as shown in Fig.5.b.

To obtain consistent depth map over the video sequence, we propose a bilateral depth fusion technique to refine the depth map by fusing the low quality depth information from multiple frames. We first project multiple depth maps into the reference frame and apply a weighted average to fuse the

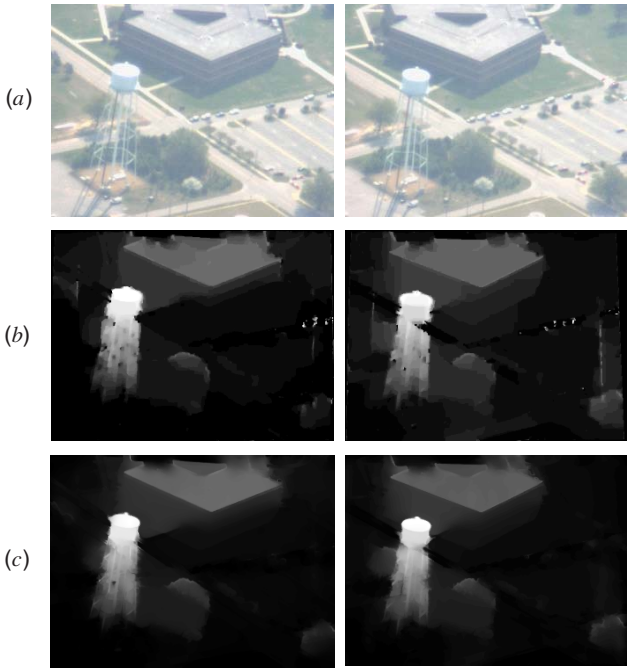


Figure 5. Depth estimation and fusion results. (a) Input video frames. (b) Depth estimation computed by graph cut, which may not be consistent between the frames. (c) Refine depth using bilateral filter.

maps into one depth map $\bar{d}$. Then a color guided bilateral filter [16] is designed to smooth $\bar{d}$ the depth map:

$$d = \frac{1}{k_p} \sum_{q \in \Omega} \bar{d}_q f(\|p - q\|) g(\|I_p - I_q\|), \tag{6}$$

where d is the final output depth map, p and q are pixel locations, f is a spatial filter kernel, g is range filter kernel over image color domain, Ω is the spatial support region of kernel f, and k_p is a normalizing factor. Fig.5.c shows the bilateral fusion results where the smoothness in continuous regions and sharp discontinuities at depth boundaries are preserved.

4. Multi-cue scene segmentation

In this section, we first discuss our planar plus depth extension scene segmentation approach for building and tree segmentation by combining the multiple cues estimated from videos. Then, we present the integration of additional GIS information to identify road network within the segmented ground regions.

4.1. Building and tree detection

Using the estimated depth with a threshold can partition each video frame into ground and non-ground regions as shown in Fig.6.c. After the partition, the segmented non-ground regions include tree, building and some false alarms

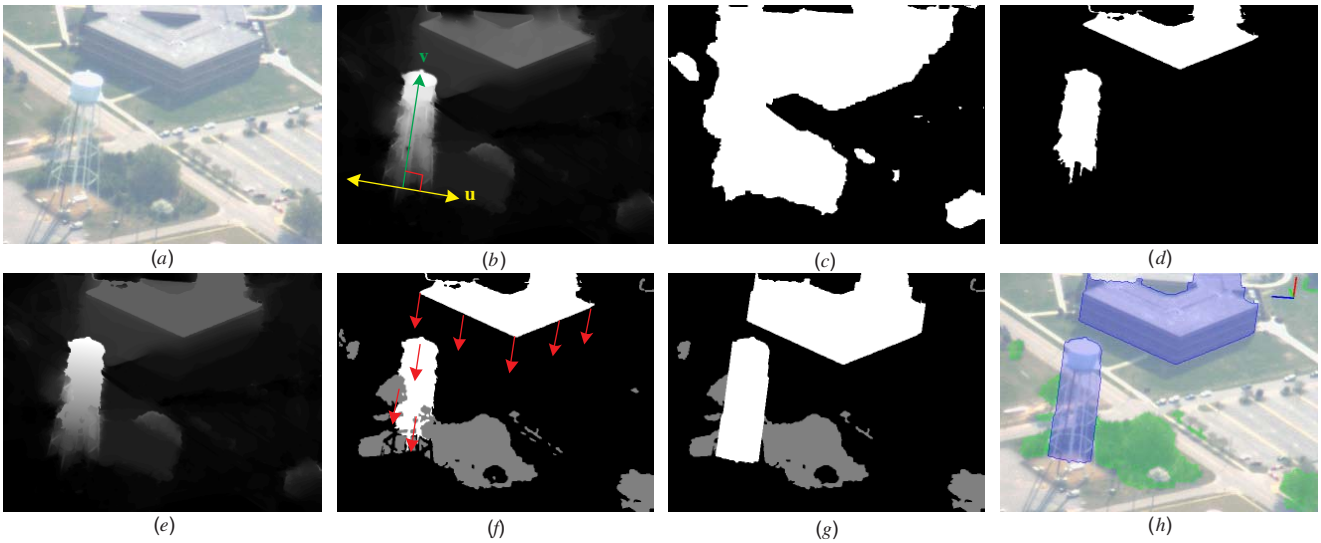


Figure 6. Multi-cue building and tree detection and segmentation. (a) Input video frame. (b) Corresponding depth map. (c) Non-ground region detection using a low depth threshold, which may include tree, building and some false alarm regions. (d) Roof detection with a high depth threshold and non-tree filtering. (e) Refine depth for roof regions by two-dimension plane fitting (Eq.9). (f) Tree region detection (gray pixels) using a mix-gaussian model with depth, color, and texture. (g) Region extension from the roof to ground along depth direction as shown in (f). (h) Final result of building and tree detection and segmentation.

due to the imperfect depth estimation. Then, we raise the depth threshold to detect the roof segments as shown in Fig.6.d. For each roof region, a common plane model can be used to refine the depth and identify the roof categories, such as flat roof, slant roof, or a tall sidewall like water tower. The plane mode is given as

$$f_1 x + f_2 y + f_3 z + f_4 = 0, \tag{7}$$

where f_is are the plane parameters, x and y are image coordinates, and z is estimated depth. However, this model is for genetic plane fitting and does not fully exploit 3D camera pose constraint implied in the video.

Given a fixed depth z, the plane in 3D world will intersect with the viewing plane as a line as shown in Fig.6.b. Along this intersection line, the depth would be invariant. For different depthes, the planes will have different intersection lines in the image. Since the camera in the UVA video is a weak perspective camera due to large flying height, this set of lines will approximately parallel to a direction, $\mathbf{u}$, and depth change along the gradient descent direction, $\mathbf{v}$, which would be perpendicular to $\mathbf{u}$. Therefore, the freedom of plane fitting model is reduced and a two-dimension function is good enough to estimate the parameters for the plane such that

$$f_5 w + f_3 z + f_4 = 0, \tag{8}$$

where w is the coordinate along v direction and $w = [x\ y]' \cdot \mathbf{v}/\|\mathbf{v}\|$. Fig.7 and Fig.6.$e-f$ compare the results using different plane fitting schemes. The approach using two-dimension function (Eq.9) can greatly improve imprecise depth estimation and achieve more reasonable results for scene structure extraction.

Once the depth gradient descent direction, $\mathbf{v}$, is determined, we can recover the whole building segment by extending the roof pixels along this direction to the corresponding pixels on the ground with depth $z = Z_0$ as shown in Fig.6.$f-g$. Using this way, the structures of the buildings are fully recovered and the height of each building is also determined. Our approach can effectively tolerate depth quality variations and does not require highly precise depth estimation for building detection and segmentation.

For tree detection, a multi-cue gaussian mixture model is built from a set of sample images with the estimated depth map such that

$$p(x; a_k, S_k, \pi_k) = \sum_{k=1}^{m} \pi_k p_k(x), \tag{9}$$

where π_k is the weight of k-th mixture and $\sum_{k=1}^{m} \pi_k = 1$, a_k and S_k are the corresponding mean and covariance matrix. In our case, the feature space dimension, m, is nine, which is composed by depth, color, and texture. After the training stage, we apply the model to classify image pixels for the tree detection and segmentation. Using our model, we can reject most false alarms from either grass or specular lighting regions as shown in gray color Fig.6.f. Some false alarms will be further reduced by building extension and region size constraints. The final segmentation result is illustrated in Fig.6.h.

Figure 7. The results using genetic plane fitting. The refined depth may have some change along depth invariant direction $\mathbf{u}$, which may cause building distortion as shown in the right side image.

4.2. GIS guided road detection

Without any knowledge, detecting road from single image or video sequence is still a challenging problem. In this paper, we leverage the correlated GIS information to relax the difficulty for road detection, which is then used for vehicle tracking and scene segmentation.

Given a road network, the image can be easily partitioned into different blocks, which may be associated with certain semantic meanings, such as parking lot, foliage, business zone, or residential zone. However, the GIS metadata related to geo-reference image may not be precisely aligned with the image and does not provide road width information for each road. Fig.8.a shows one example of the original road metadata projected on the geo-reference image, where a few pixels offset between the real road and the GIS road.

In order to obtain better road network, we combine road appearance with parallel lines detection to refine the road. Our approach includes two steps: model training and detection. In the training step, a set of road patches are sampled from the images based on the GIS road information. After aligning these patches along the road direction, we extract color and gradient features from the patches. These features are projected on the vertical direction of the road and form a combined histogram to represent the road patch. A gaussian mixture model is then created to model the distribution of the feature vectors.

In the detection stage, we also align the extracted road along the road direction (y axis) as shown in Fig.8.c. Next, we shift the histogram along x axis to identify the best correlation between the gaussian mixture model and the input feature histogram. Fig.8.d shows the correlation result. Once the road center is determined (green lines in Fig.8.c and 8.e), we can estimate the road width (bounded by pink lines) by peak detection on gradient histogram with symmetrical constraint as shown in Fig.8.e. Fig.8.e shows the final results for the road refinement.

The estimated road network is also very useful for vehicle tracking and depth correction. For example, with the road knowledge, the tracking process can effective handle occlusion issue along the road direction.

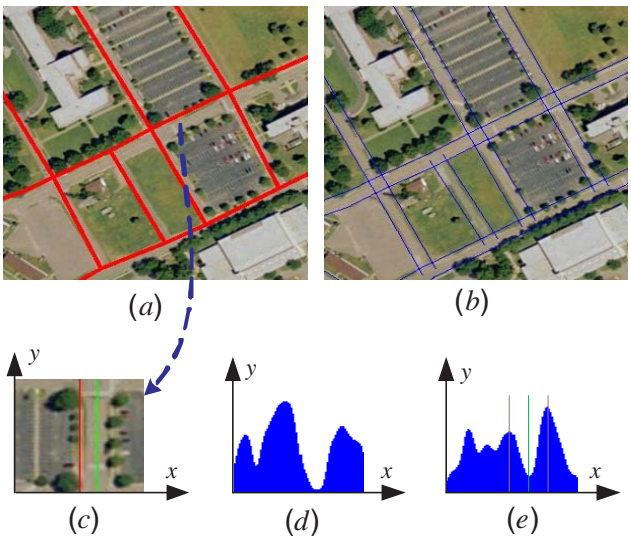


Figure 8. (a) The road network from GIS is overlapped on the geo-reference image, where a few pixels may be off from the real road. (b) The corresponding refined road network. (c) One road patch is extracted and aligned along the road direction, where red line is the road center provided from GIS and green line is the refined road center by our approach. (d) The histogram correlation results using our gaussian mixture model. (e) The projected gradient histogram for road width estimation.

5. Experimental results

In this section, we report the experimental results on VIVID2 aerial video data set. These videos are captured by an aerial platform with flying height about 1000 meters for semi-urban area vehicle tracking. In our experiment, we select several video sequences from DLTV (daylight TV sensor), and each sequence is about four minutes with more than 7000 frames. After preforming the geo-spatial aerial video process, we generate a series of results including geo-registration, camera pose, depth map, road map, scene segmentation and moving object tracking for each input video, which provide fundamental primitives for high level event analysis and scene exploitation.

Since low flying aerial video consists of strong parallax when the camera viewing a 3D scene, it would be challenging to detect and track moving objects using 2D frame-to-frame stabilization. Using epipolar constraint or 3D shape constraint is the most popular way to remove false alarms whose motion is consistent with 3D geometry. This paper also employs the 3D shape constraint to remove false alarms [12] since we have explicitly estimated the depth information for each frame. In our approach, we combine stabilization, optical flow warping and depth warping to detect moving blobs for tracking. Fig.9 compares vehicle tracking results with/without using depth integration. With the depth estimation, false alarms around tall static objects, such as water tower or tree tips, can be significantly reduced. To

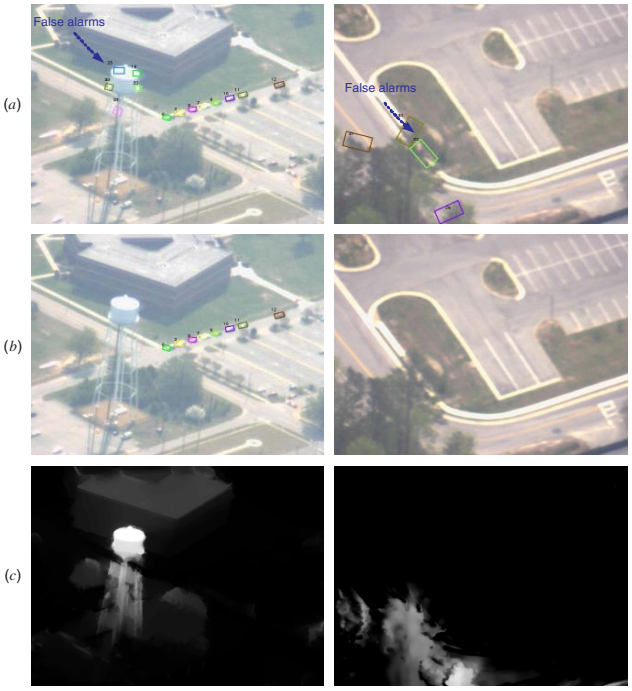


Figure 9. Elimination of tracking false due to 3D parallax. (a) The tracking results without using 3D shape constraint. (b) The tracking results with 3D shape constraint, where the false alarms are reduced significantly around the water tower or tree tip. (c) The corresponding depth maps.

quantitatively evaluate our tracking performance, we manually generate ground truth for some sequences and then employ the performance metrics from standard evaluation program [7] to test our algorithm. One critical metric in [7] is Multiple Object Tracking Accuracy (MOTA) defined as

$$MOTA = 1 - \frac{\sum_{t=1}^{N}(c_f + c_m + \log(c_s))}{\sum_{t=1}^{N}(c_g)}, \qquad (10)$$

where c_f and c_m are the false acceptance and false rejection counts, c_s is total number of incorrect identity switches made by the system, and c_g is the ground truth object counts. The perfect tracking result will be given a score equal to 1, and the tracking performance become worse when the score becomes smaller or negative. After we tested the sequences, the average MOTA is improved from 0.740 to 0.851 (15% improvement) and false alarm rate is dropped from 0.190 to 0.072 (62% improvement).

Fig.11 illustrates the final scene segmentation results for different frames by combining vehicle tracking, building, foliage, and road detection, where the corresponding depth maps are provided in Fig.10. After applying our multi-cue scene segmentation approach, the input aerial video is then able to be interpreted as a text message or an event report. For example, by exploiting the spatiotemporal relationship among the segmented layers and moving vehicles, the image in Fig.11.a then can be encoded as a report for event analysis, such as "eight cars are driving along 19th street and stop at intersection between 19th street and A avenue", "building 0 is located at the right side of the vehicles", etc.

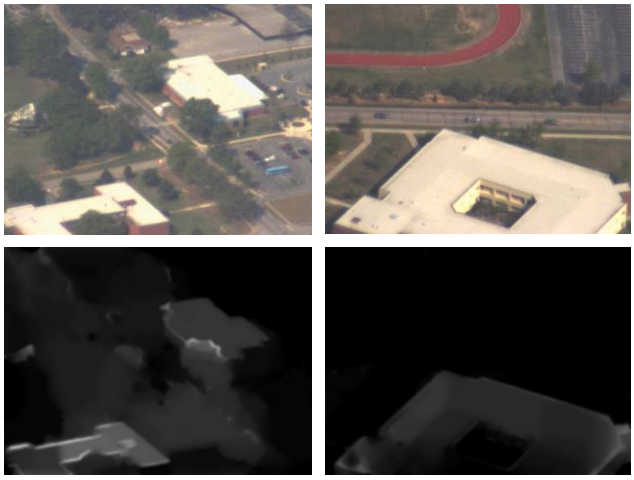
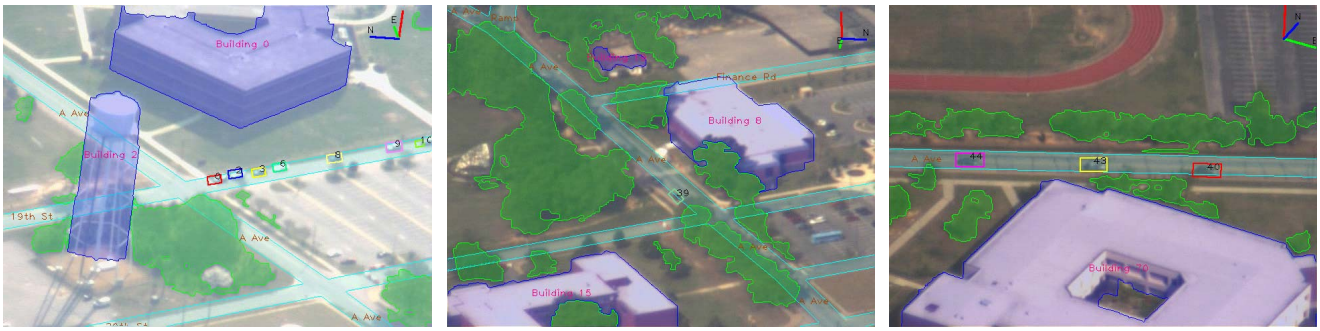


Figure 10. The corresponding original images and depth maps of Fig.11.b and 11.c.

Another benefit of our geo-based approach is that we can provide absolute scale in the video for the photogrammetry purpose including height of building, driving distance or vehicle speed, while the traditional monocular structure from motion has a problem to estimate such accurate scale over long periods of time. For example, in Fig.11.a, the heights of the water tower and building are measured as 50 meters and 15 meters respectively.

6. Conclusion

In this paper, we have presented an approach to extract semantic layers from low attitude aerial videos for scene understanding and object tracking. Our main contributions consist of: (1) We provide a reliable geo-based solution to estimate camera pose for depth estimation of an aerial video. (2) Using the estimated depth cue combining with other image features, we propose a planar plus depth extension approach to preform scene segmentation for both building and foliage. (3) Elimination of false tracking alarms of moving objects using 3D parallax constraints and semantic labels derived from geo-registered videos. After applying our approach on a set of aerial video sequences captured by VIVID program, the experiment results illustrate that our method can produce reasonable depth map over a long period time and significantly reduce the false alarms for vehicle tracking due to 3D parallax. Even for such low quality, monocular video sequences, our approach still generates promising scene segmentation results for video-based semantic scene exploitation and event analysis.



(*a*) (*b*) (*c*)

Figure 11. The final scene segmentation results. For each frame, a 3D coordinate system is drawn on the top-right corner, where red line is Z axis, green line points to east, and blue line points to north. The road name, building ID, and tracking ID are marked on each object in the image.

7. Acknowledgements

This work was supported by the IARPA/ODNI.

References

[1] S. Ali, V. Reilly, and M. Shah. Motion and appearance contexts for tracking and re-acquiring targets in aerial videos. In *Computer Vision and Pattern Recognition*, 2007. 2

[2] C. Baillard, C. Schmid, A. Zisserman, and A. Fitzgibbon. Automatic line matching and 3d reconstruction of buildings from multiple views. In *ISPRGIS*, 1999. 2

[3] T. Belli, M. Cord, and M. Jordan. 3d data reconstruction and modeling for urban scene analysis. In *Workshop of Automatic Extraction of Man-Made Objects from Aerial and Space Images (III)*, 2001. 2

[4] N. Haala and C. Brenner. Extraction of buildings and trees in urban environments. *ISPRS Journal of Photogrammetry and Remote Sensing*, 54:130–137, 1999. 2

[5] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. 3, 4

[6] J. Kang, I. Cohen, G. Medioni, and C. Yuan. Detection and tracking of moving objects from a moving platform in presence of strong parallax. In *IEEE International Conference on Computer Vision*, 2005. 1

[7] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, V. Manohar, M. Boonstra, and V. Korzhova. Performance evaluation protocal for face, person and vehicle detection & tracking in video analysis and content extraction. In *Workshop of Classification of Events, Activties and Relationships*, 2006. 7

[8] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cut. In *Proceedings of ECCV*, 2002. 4

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 3

[10] A. Perera, G. Brooksby, A. Hoogs, and G. Doretto. Moving object segmentation using scene understanding. In *Computer Vision and Pattern Recognition*, 2006. 2

[11] A. Perera, C. Srinivas, G. Hoogs, A. Brooksby, and W. Hu. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In *Computer Vision and Pattern Recognition*, 2006. 2

[12] H. Sawhney, Y. Guo, and R. Kumar. Independent motion detection in 3d scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22:1191–1199, 2000. 1, 6

[13] H. Sawhney, S. Hsu, and R. Kumar. Robust video mosaicing through topology inference and local to global alignment. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22:1191–1199, 2000. 3

[14] H. Shum and R. Szeliski. Construction of panoramic image mosaics with global and local alignment. *International Journal on Computer Vision*, 16(1):63–84, 2000. 3

[15] H. Tang, Z. Zhu, G. Wolberg, and J. Layne. Dynamic 3d urban scene modeling using multiple pushbroom mosaics. In *the Third International Symposium on 3D Data Processing, Visualization and Transmission*, 2006. 2

[16] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *International Conference on Computer Vision*, pages 839–846, 1998. 4

[17] Z. Tu and S. Zhu. Image segmentation by data-driven markov chain monte carlo. *IEEE Trans. Pattern Anal. Mach. Intell*, 24:657–673, 2002. 2

[18] V. Verma, R. Kumar, and S. Hsu. 3d building detection and modeling from aerail lidar data. In *Computer Vision and Pattern Recognition*, 2006. 2

[19] J. Xiao and M. Shah. Motion layer extraction in the presence of occlusion using graph cut. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10), 2005. 4

[20] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Journal of Computing Surveys*, 38, 2006. 2

[21] Z. Zhang. A flexible new technique for camera calibration. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(11), 2001. 4