3D GEOREGISTRATION OF WIDE AREA MOTION IMAGERY BY COMBINING SFM AND CHAMFER ALIGNMENT OF VEHICLE DETECTIONS TO VECTOR ROADMAPS

Li Ding, Ahmed Elliethy, and Gaurav Sharma

Dept. of Electrical and Computer Engineering, University of Rochester, Rochester, NY

ABSTRACT

We propose a novel framework for accurate 3D georegistration of wide area motion imagery (WAMI), which is a challenging problem because parametric transformations are insufficient for aligning WAMI image frames to a georeferenced coordinate system in urban areas containing tall buildings and 3D structures. Using structure from motion (SfM) we estimate a 3D point cloud for the scene. Independently, we also compute a precise alignment between the roads in the WAMI frames and a georeferenced vector roadmap by detecting locations of moving vehicles and aligning these locations with the roads in the vector roadmap via parametric chamfer matching. The aligned vector roadmap then identifies corresponding pixels in the WAMI frames, which can be triangulated using the SfM camera parameters to obtain a set of sparse but georeferenced points in the SfM 3D coordinate frame that directly enable georegistration of the complete 3D scene point cloud via a similarity transform.

The proposed methodology enables 3D georegistration of a sequence of WAMI frames using only georeferenced vector roadmaps, which are readily available, and without requiring independent georeferenced lidar scans that have been used in prior work. Our framework is validated on WAMI dataset including high resolution WAMI frames for the downtown Rochester, NY region. Experimental results demonstrate that the proposed framework produces an accurate georeferenced point cloud representation for the scene.

Index Terms— WAMI, georegistration, structure from motion (SfM), expectation maximization (EM).

1. INTRODUCTION

The past decade has seen the development and deployment of a number of aerial Wide Area Motion Imagery (WAMI) capture systems, such as the CorvusEye [1], Autonomous Real-time Ground Ubiquitous Surveillance Imaging System (ARGUS-IS) [2], Constant Hawk [3], and Hawkeye [4]. These systems capture a sequence of high resolution temporal image frames at 1-3 frames-per-second with a field of view that covers an extended geographic area, typically spanning a few square miles. The rich spatio-temporal information captured in WAMI imagery is increasingly being used for a variety of applications including military operations, surveillance, disaster monitoring/assessment, law enforcement, border enforcement, and urban planning. For such applications, georegistration, i.e. localization of each pixel in 3D space relative to the Earth, enhances the utility of the WAMI data by allowing additional information to be integrated from existing georeferenced sources such as roadmaps, satellite images, and digital elevation maps.

WAMI georegistration is challenging because parametric transformations are insufficient for aligning WAMI image frames to a georeferenced coordinate system, particularly in urban areas containing tall buildings and 3D structures. Methods for accurate georegistration therefore rely on auxiliary information of the scene 3D structure, obtained, for example, from lidar scanning or digital elevation map, to register 2D image [5, 6, 7, 8, 9]. Although these methods are able to register images to georeferenced model, they have limitations: the acquisition of lidar point cloud or DEM can be both expensive and time-consuming and these are not always publicly available. As an alternative, researchers have also used Structure from Motion (SfM) to reconstruct and georegister the 3D point cloud from the scene images. For instance, [10] and [11] choose onboard sensors such as GPS/IMU sensors that record geospatial information. However, the accuracy of the georeferenced point cloud is limited by the low-fidelity geographic information. To improve accuracy, others have explored methods to align the point cloud with 2D georeferenced data. In [12], for example, a 2D map is used for point cloud georegistration that is formulated as aligning 3D planes in point cloud with 2D lines in map. Similarly, recent work in [13] proposes an algorithm to georegister a 3D point cloud from oblique-view video with a cadastral map. The method in [14] leverages three different types of geographic sources, i.e., the noisy meta-data, the Google Street View image, and the 3D model exported from Google Earth. Although this method is able to produce good results, it has limitations: the method uses the Iterated Closest Point (ICP) algorithm [15] to align point cloud with 3D Google Earth model, which highly reliant on finding a good initial guess. In situations where a Google Street View image is not available or where the quality of Google Earth Model is poor, the accuracy is low because the alignment can be trapped in local minima.

In this paper, we propose a novel framework to accurately georegister a point cloud from WAMI frames by integrating 3D scene structure estimated from SfM with chamfer alignment of vehicle detections with readily available vector roadmaps. Our framework has several advantages. We do not require either (a) prior knowledge from the scene (e.g., known geographic coordinate of landmarks) or human intervention during the processing, or (b) 3D georeferenced models from other sources such as Google Earth, where the quality of information can vary significantly across cities. Instead, we utilize a vector roadmap to provide georeferencing through chamfer alignment with detected vehicle locations in the WAMI frame. An expectation-maximization (EM) formulation of the chamfer alignment makes our method robust to the noisy GPS meta-data and spurious vehicle detections.

The paper is organized as follows. Section 2 sketches out of the proposed framework. We present the experimental results on a real WAMI dataset in Section 3, and conclude the paper in Section 4.

2. GEOREFERENCED POINT CLOUD ESTIMATION

The proposed framework addresses the problem of accurate 3D georegistration of WAMI frames using the pipeline depicted in Fig. 1. Specifically, the 3D point cloud of the scene is reconstructed from WAMI frames using an SfM algorithm. Independently, we register the WAMI frames to the georeferenced vector roadmap by chamfer alignment of detected moving vehicles in the WAMI frames to the roads in the vector roadmap, resulting in a set of georeferenced pixel

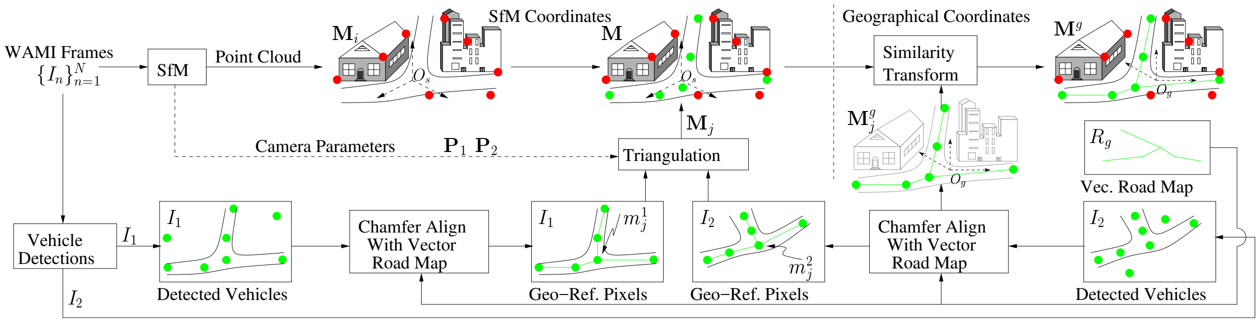

Fig. 1. Proposed methodology 3D georegistration of WAMI using structure from motion (SfM) and chamfer alignment of vehicle detections.

locations on the road network in the WAMI frame. The identified pixel locations can be triangulated to obtain 3D georeferenced points in the SfM coordinate system that directly enable georegistration of the complete 3D scene point cloud via a similarity transform.

2.1. 3D Reconstruction Using SfM

The first stage takes a set of WAMI frames $\mathbf{I} = \{I_n(u^n, v^n)\}_{n=1}^{N}$ as input, where N is the number of images. We perform SfM to simultaneous reconstruct 3D scene points $\mathbf{M}_i = \{M_i\}_{i=1}^{K}$ and estimate projective camera parameters $\mathbf{P}_n$ for each WAMI frame, where K is the number of points and M_i is the 3D coordinate of the point i. The projective camera parameters describe how a 3D point is mapped onto 2D image plane. Among several proposed SfM strategies, incremental SfM [16] has been widely used. The basic idea is that keypoints are detected in each frame and matched between all pairs of frames. Then an iterative procedure is performed to recover camera parameters as well as 3D point locations. In each iteration, only one camera is added for optimization. The reconstructed point cloud is shown as the red dots in the point cloud $\mathbf{M}_i$ in Fig. 1.

2.2. Registration of WAMI With Vector Roadmap

The SfM algorithm only estimates a point cloud in a relative coordinate system rather than in an absolute geographic coordinate system. Therefore it is necessary to obtain geospatial information for the point cloud. Instead of using known geographic coordinates of landmarks, we apply an EM framework [17] to precisely register two WAMI frames to the georeferenced vector roadmap R_g by using locations of moving vehicles detected in the WAMI frames and aligning these locations with vector roadmap. More concretely, we first extract corresponding feature points in two successive frames I_1 and I_2 and estimate a homography H_{21} to transform features in I_2 to the corresponding ones in I_1

$$s_h \tilde{m}^1 = H_{21} \tilde{m}^2, \tag{1}$$

where $\tilde{m}^1$ and $\tilde{m}^2$ are the homogeneous coordinates of feature points in I_1 and I_2, respectively, and s_h is an arbitrary factor. The frame $I_1^{'}$ is estimated based on H_{21} that can be view as an image captured at viewpoint of I_1 but at the time when I_2 is captured. The vehicles are the detected by the compensated frame difference [18]: a pixel location $\tilde{m}_n^1 = \left[u_n^1, v_n^1, 1\right]^T$ where the absolute difference of pixel value between I_1 and $I_1^{'}$ is greater than a threshold is considered part of a vehicle.

The vector roadmap is represented by a sequence of locations of longitude ϕ and latitude θ, which are later mapped to a 2D coordinate system (χ, ζ) using azimuthal orthographic map projection (AOMP) [19]. For each detection $\tilde{m}_n^1$, we compute the minimum distance $d_n(H_\beta)$ from its mapped location under the homography transformation H_β to the road network

$$d_n(H_\beta) = \min_j D\left(\tilde{p}_j, H_\beta \tilde{m}_n^1\right), \tag{2}$$

where $\tilde{p}_j = [\chi_j, \zeta_j, 1]^T$ are the homogeneous coordinates of the vector roadmap and β denotes the homography parameters. The vector road network alignment is formulated as a homography model H_β that minimizes the objective function

$$Q_1 = \sum_{n=1}^{N_v} p_n d_n(H_\beta), \tag{3}$$

where N_v is the number of detections including on-road vehicles and spurious detections, and p_n is the posterior probability that the n-th detection is on-road vehicle. The minimization in (3) can be recognized as a probabilistic formulation of chamfer minimization [20]. We apply the EM algorithm [21] to estimate β. In the E-step, we update the posterior probability p_n of detection reliability based on current parameters for alignment. Given the estimated p_n, we re-estimate the parameters β in the M-step. This process is repeated until convergence. A detailed description can be found in [17]. Once the parameters β is determined, we apply H_β^{-1}, the inverse of the transformation, to the vector road network, which establish a set of correspondences for points $j \in \mathcal{J}$ on the road network in both the WAMI frame pixel coordinates and the georeferenced coordinates.

2.3. Point Cloud Georegistration

The georeferenced pixel locations m_j^1 provide geospatial information for the road network in the scene. We incorporate these pixel locations with the point cloud from SfM for the purpose of point cloud georegistration. Using the previous section's approach, we first identify the georeferenced points m_j^2 in I_2 that correspond to m_j^1. Given a set of point correspondences $m_j^1 \leftrightarrow m_j^2$ and the camera parameters $\mathbf{P}_1$ and $\mathbf{P}_2$ of the two frames, we compute the corresponding 3D positions $\mathbf{M}_j$ by back-projecting rays from the image points. The 3D points $\mathbf{M}_i$ and $\mathbf{M}_j$ are in the same coordinate system because the same camera parameters are used (the red and green dots in point cloud $\mathbf{M}$ in Fig. 1, respectively). Hereafter we refer to $\mathbf{M} = \mathbf{M}_i \cup \mathbf{M}_j$ as the complete points in SfM coordinate system.

Once the 3D points $\mathbf{M}_j$ are found, the transformation that maps $\mathbf{M}_j$ to geographical coordinate system can be estimated. Adopting a similarity transformation which is defined by a unitary rotation matrix $\mathbf{R}_{3\times3}$, a translation vector $\mathbf{t}_{3\times1}$, and a positive scaling factor s, a procedure of transformation estimation can be defined by minimizing the objective function

$$Q_2 = \sum_{j \in \mathcal{J}} \| s\mathbf{R}_{3\times3} M_j + \mathbf{t}_{3\times1} - M_j^g \|^2, \tag{4}$$

where M_j^g is the 3D location of points M_j on a geographical coordinate system derived from the longitude ϕ_j and latitude θ_j (spherical-to-Cartesian transformation). The minimization of objective function Q_2 can be solved using method in [22] that provides a close-form solution for this least-squares problem. The georeferenced point cloud $\mathbf{M}^g$ is then computed by applying the similarity transformation to the complete point cloud $\mathbf{M}$.

3. EXPERIMENTAL RESULTS

We validate our framework by mapping the WAMI frames to a georeferenced satellite image using the estimated point cloud $\mathbf{M}^g$. The satellite image $I_s(\chi, \zeta)$ can be considered as the image of earth surface captured using an affine camera [23]. The spatial location of each pixel (χ, ζ) is assigned with a location of longitude and latitude (ϕ, θ). Since the points in $\mathbf{M}^g$ have georeferenced coordinates, we establish correspondences between these 3D points and their 2D counterparts in the georeferenced satellite image. Given 3D-2D correspondence, the affine camera parameter $\mathbf{P}_a$ can be estimated that maps the 3D reconstructed scene $\mathbf{M}$ onto the satellite image plane, thereby providing transformation between the WAMI frame and the satellite image. We do not consider radial distortion since the WAMI frames and satellite image are corrected to distortion-free images.

We use the CORVUS visible band dataset which was recorded using the CorvusEye 1500 Wide-Area Airborne System [1] for the downtown Rochester, NY region. Each WAMI frame is comprised of a RGB image with resolution 4400×6600 and the associated meta-data that includes the approximate geographical coordinates for the four corners. The vector roadmap is provided by OpenStreetMap (OSM) [24]. For the georeferenced satellite image, we use the Google Static Maps API service [25] to request the map as a satellite image. The satellite image is then manually cropped to cover the similar region to the WAMI frame[1]. The satellite image has a resolution of 1280×1280 and one pixel corresponds to about 0.869m on the ground plane.

3.1. Qualitative and Quantitative Results

In our experiments, we apply the state of the art SfM algorithm [26] on 50 WAMI frames to reconstruct 3D point cloud, and perform the dense reconstruction method PMVS/CMVS [27] to generate a dense point cloud of the scene. We first show the qualitative results of our framework and compare these with two alternative methods: "meta-data based alignment (MBA)" and "road network georegistration based alignment (RN)". The MBA method depends only on the meta-data in the WAMI frame. Specifically, this method estimates the homography between the WAMI frame and satellite image from the correspondences of the locations of the four corner points in the WAMI frame and the map coordinates using DLT. This homography can be interpreted as the transformation used to orthorectify the WAMI frame. The RN method estimates the homography between the WAMI frame and the satellite image by aligning the WAMI frame with the georeferenced vector roadmap, as computed in Section 2.2. These two alternative methods solve the problem of aligning WAMI frames with the satellite image by estimating and applying the 2D homography to all pixels in the WAMI frame.

To visualize the results for each algorithm, we superimpose the transformed WAMI frame and satellite image to compare the difference between these two images. The results of georegistration of one WAMI frame for the MBA, the RN and the proposed method are shown in Fig 4. The satellite image is shown as green while the transformed WAMI frame is represented in magenta. The results of georegistration using MBA method have significant error due to the limited accuracy of on-board WAMI sensor for recording locations. For example, we can readily identify misalignment both for the road network and buildings, as shown in the first row of Fig 4. The RN method is able to enhance the accuracy of georegistration for the road network region, which can be modeled as a planar structure. We show the estimated alignment of the vector roadmap in Fig 2 by superimposing the estimated road locations as highlighted tracks overlaid on the WAMI frame. This method, however, is not able to significantly improve the accuracy in the region where the buildings are located. The proposed method, which employs both 3D structure as well as the georeferenced vector roadmap, offers a significant improvement over the MBA and the RN methods. The buildings in the WAMI frame, for example, which are not properly aligned by the MBA and the RN methods, can be precisely aligned with the satellite image, as shown in the last row of Fig. 4.

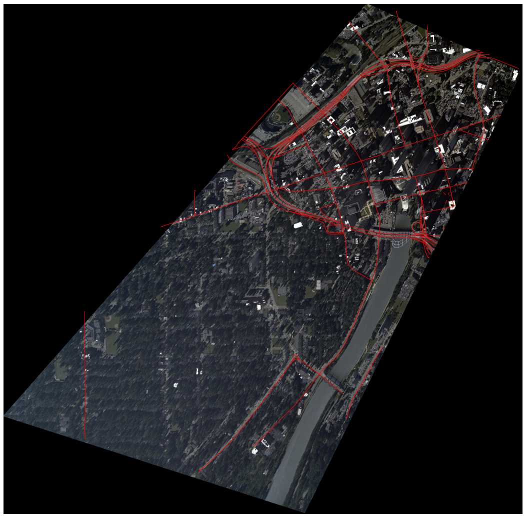

Fig. 2. Sample results of road network alignment. The red line segments represent vector roadmap that are superimposed on the transformed WAMI frame.

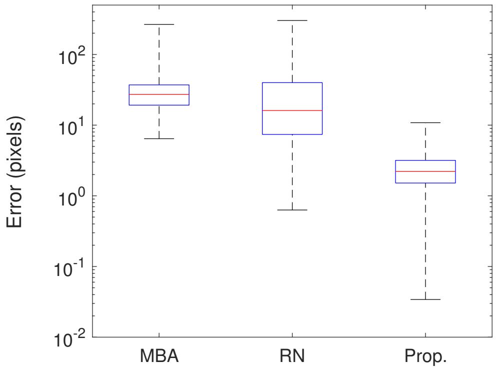

Fig. 3. Box plot of reprojection error between ground truth points and the mapped points estimated from the MBA, the RN, and the proposed method. Note the logarithmic scale for the y-axis

[1] Code for downloading from Google Maps is provided by Zohar Bar-Yehuda: https://www.mathworks.com/matlabcentral/fileexchange/27627-zoharby-plot-google-map

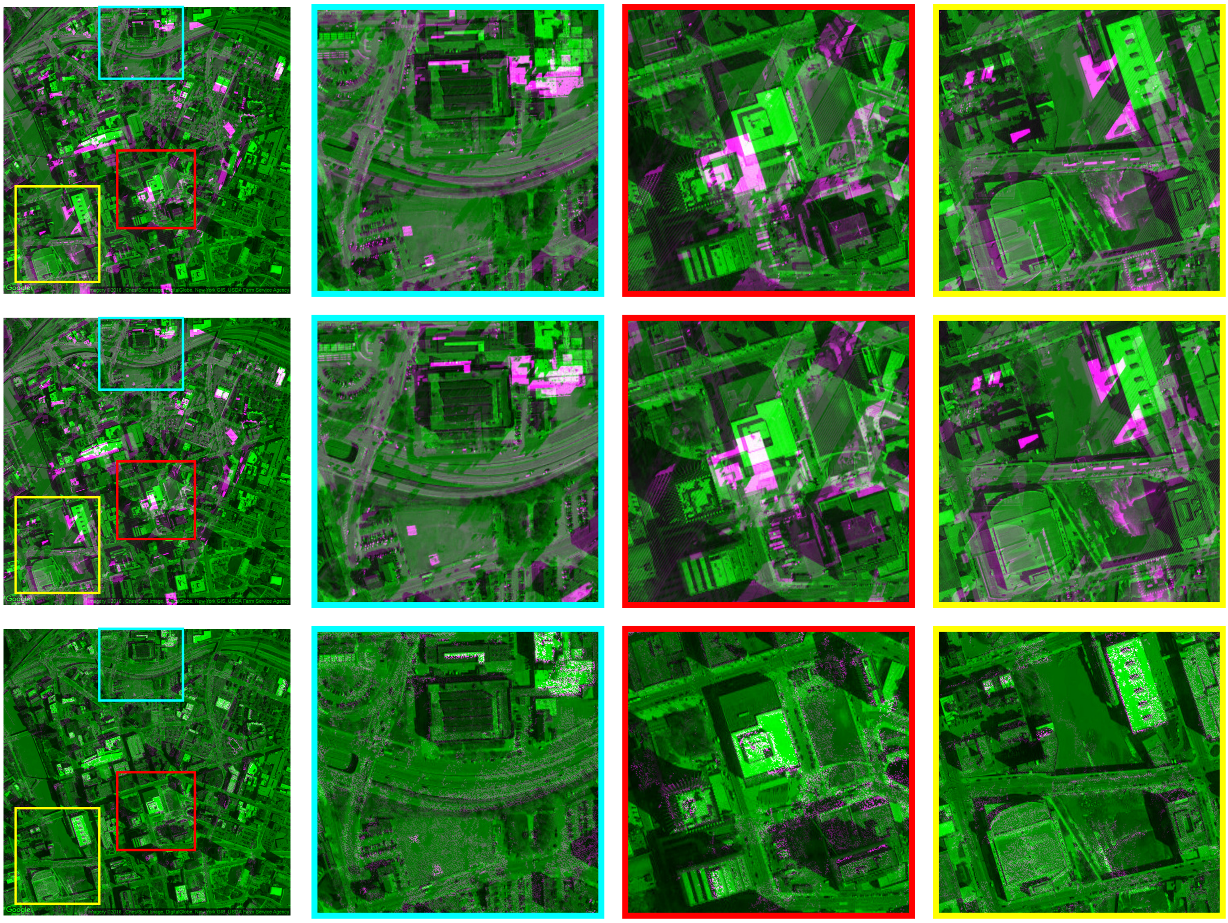

Fig. 4. Sample results of visual comparison of WAMI georegistration using different methods. Rows from top to bottom: the MBA method, the RN method, and the proposed method. The transformed WAMI frame, which appears in magenta, is superimposed with the georeferenced satellite image shown in green. The first column is the full satellite images, while the second to the last columns shows the close-ups that marked on the corresponding full satellite images by the same colors.

To provide quantitative comparison, we carefully select several 3D points obtained from SfM and manually identify their counterparts in the 2D satellite image to obtain the ground truth camera (GTC) parameters. From SfM, we also have the parameters of each individual scene cameras (SCs) that captured the WAMI frames. Combining the parameters of SC and GTC, we built a 2D-3D-2D ground truth correspondence among pixels in WAMI frame, points in 3D point cloud, and pixels in satellite image. Before discussing how the proposed method performs, it is worth emphasizing that here we assume the parameters of each scene camera are accurate. This assumption holds because we found the average reprojection error is relatively small, e.g., one pixel in the WAMI frame.

For each reconstructed pixel in the WAMI frame, we apply the MBA, the RN method, and the proposed method to determine its pixel location in satellite image. Figure 3 shows the displacement error, in pixel units, between the ground truth points and the transformed points in the satellite image, for each aforementioned method. Several observations can be made: (1) The MBA method has noticeable error due to limited accuracy of on-board sensors. The error ranges from 6.4242 to 266.1876 pixels, with the mean of 31.6572 pixels which equals a distance of 27.51m. and (2) the RN method can improve the accuracy of georegistration (the minimum error is 0.63 pixels), but still has a large mean and maximum error (28.613 and 302.4708 pixels, respectively). This is not surprising, however, because the homography model used in the MBA and the RN method is valid only for the planar structure. The proposed method, which employs both 3D structure and georeferenced vector roadmap, offers a significant enhancement over other two methods. For instance, the error ranges from 0.0341 to 10.8202 pixels, with the mean of 2.7021 pixels. The results of quantitative comparison are in accordance with that of visual examination.

4. CONCLUSION

The proposed framework provides an accurate methodology for 3D georegistration of WAMI frames by integrating SfM and chamfer alignment of detected vehicles with vector roadmaps. Our approach relies only on georeferenced vector roadmaps, which are readily available. Both qualitative and quantitative results indicate that the method achieves high accuracy of 3D WAMI georegistration.

5. ACKNOWLEDGMENT

We thank Bernard Brower of Harris Corporation for providing CorvusEye [1] datasets, and the Center for Integrated Research Computing, University of Rochester, for providing access to computational resources.

6. REFERENCES

[1] "Corvuseye 1500 data sheet." [Online]. Available: http://www.exelisinc.com/solutions/corvuseye1500/Document/CorvusEye500DataSheetAUG14.pdf

[2] "Autonomous real-time ground ubiquitous surveillance imaging system," Defense Advanced Research Projects Agency, http://www.darpa.mil.

[3] "Constant hawk," Gobal Security, globalsecurity.org/intell/systems/constant-hawk.htm.

[4] "Wide-area aerial surveillance systems," Persistent Surveillance Systems, http://www.persistentsurveillance.com.

[5] A. P. Brown, M. J. Sheffler, and K. E. Dunn, "Persistent electro-optical/infrared wide-area sensor exploitation," in *Proc. SPIE*, vol. 8402, 2012, pp. 840 206–840 206–9.

[6] M. D. Pritt and K. J. LaTourette, "Georegistration of motion imagery with error propagation," in *Proc. SPIE*, vol. 8386, 2012, pp. 838 606–838 606–12.

[7] ——, "Automated georegistration of motion imagery," in *2011 IEEE Applied Imagery Pattern Recognition Wksp*, Oct 2011, pp. 1–6.

[8] C. Bodensteiner, S. Bullinger, S. Lemaire, and M. Arens, "Single frame based video geo-localisation using structure projection," in *IEEE Intl. Conf. Comp. Vision Wksp.*, Dec 2015, pp. 1036–1043.

[9] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz, "Accurate geo-registration by ground-to-aerial image matching," in *Intl. Conf. 3D Vision*, vol. 1, Dec 2014, pp. 525–532.

[10] S. Ruano, G. Gallego, C. Cuevas, and N. Garca, "Aerial video georegistration using terrain models from dense and coherent stereo matching," in *Proc. SPIE*, vol. 9089, 2014, pp. 90 890V–90 890V–10.

[11] D. J. Walvoord, A. J. Rossi, B. D. Paul, B. Brower, and M. F. Pellechia, "Geoaccurate three-dimensional reconstruction via image-based geometry," in *Proc. SPIE*, vol. 8747, 2013, pp. 874 706–874 706–13.

[12] K. Ni, N. Armstrong-Crews, and S. Sawyer, "Geo-registering 3D point clouds to 2D maps with scan matching and the Hough transform," in *IEEE Intl. Conf. Acoust., Speech, and Signal Proc.*, May 2013, pp. 1864–1868.

[13] Y. Li and T. K. Ng, "Where is that pixel in the oblique-view video?" in *IEEE Workshop on Appl. of Comp. Vision.*, March 2016, pp. 1–8.

[14] C. P. Wang, K. Wilson, and N. Snavely, "Accurate georegistration of point clouds using geographic data," in *Intl. Conf. 3D Vision*, June 2013, pp. 33–40.

[15] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 14, no. 2, pp. 239–256, Feb 1992.

[16] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *Intl. J. Computer Vision*, vol. 80, no. 2, pp. 189–210, 2008.

[17] A. Elliethy and G. Sharma, "Automatic registration of wide area motion imagery to vector road maps by exploiting vehicle detections," *IEEE Trans. Image Proc.*, vol. 25, no. 11, pp. 5304–5315, 2016.

[18] A. M. Tekalp, Ed., *Digital Video Processing*. Upper Saddle River, NJ: Prentice Hall, 1995.

[19] J. P. Snyder, *Map projections–A working manual*. US Government Printing Office, 1987, vol. 1395.

[20] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching," in *Proc. Int. Joint Conf. Artificial Intell.*, 1977, pp. 659–663.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B (methodol.)*, pp. 1–38, 1977.

[22] B. K. Horn, H. M. Hilden, and S. Negahdaripour, "Closed-form solution of absolute orientation using orthonormal matrices," *J. Optical Society of America A*, vol. 5, no. 7, pp. 1127–1135, 1988.

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

[24] "Open street map." [Online]. Available: http://www.openstreetmap.org

[25] "Google static maps API." [Online]. Available: https://developers.google.com/maps/documentation/static-maps/intro

[26] C. Wu, "Towards linear-time incremental structure from motion," in *Intl. Conf. 3D Vision*, 2013, pp. 127–134.

[27] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 32, no. 8, pp. 1362–1376, 2010.