This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2016.2601265

Automatic Registration of Wide Area Motion Imagery to Vector Road Maps by Exploiting Vehicle Detections

Ahmed Elliethy, *Student Member, IEEE* and Gaurav Sharma, *Fellow, IEEE*

***Abstract*—To enrich large scale visual analytics applications enabled by aerial wide area motion imagery (WAMI), we propose a novel methodology for accurately registering a geo-referenced vector roadmap to WAMI by using locations of detected vehicles and determining a parametric transform that aligns these locations with the network of roads in the roadmap. Specifically, the problem is formulated in a probabilistic framework, explicitly allowing for spurious detections that do not correspond to on-road vehicles. The registration is estimated via the EM algorithm as the planar homography that minimizes the sum of weighted squared distances between the homography-mapped detection locations and the corresponding closest point on the road network, where the weights are estimated posterior probabilities of detections being on-road vehicles. The weighted distance minimization is efficiently performed using the distance transform with the Levenberg Marquardt (LM) nonlinear least-squares minimization procedure and the fraction of spurious detections is estimated within the EM framework. The proposed method effectively sidesteps the challenges of feature correspondence estimation, applies directly to different imaging modalities, is robust to spurious detections, and is also more appropriate than feature matching for a planar homography. Results over three WAMI datasets captured by both visual and infra-red sensors indicate the effectiveness of the proposed methodology: both visual comparison and numerical metrics for the registration accuracy are significantly better for the proposed method as compared with existing alternatives.**

***Index Terms*—WAMI, roadmap registration, expectation maximization, geo-registration.**

I. Introduction

RECENT technological advances have made available a number of airborne platforms for capturing imagery [2]–[4]. One of the specific areas of emerging interest for applications is Wide Area Motion Imagery (WAMI) where images at temporal rates of 1–2 frames per-second can be captured for relatively large areas that span substantial parts of a city while maintaining adequate spatial detail to resolve individual vehicles [5]. WAMI platforms are becoming increasingly prevalent and the imagery they generate are also feeding a corresponding boom in large scale visual data analytics. The effectiveness of such analytics can be enhanced by combining the WAMI with alternative sources of rich geo-spatial information such as road maps.

In this paper, we propose a novel iterative framework for registering a vector road network to a WAMI aerial image frame using vehicle detections. Our method is based on the intuitive synergy between the problems of registering of a (vector) roadmap to an image frame and the detection of on-road vehicles in an image. The detection of on-road vehicles in an image allows us to register the image to a vector road map by aligning the detection locations with the roads. Conversely, a roadmap registered with the WAMI image improves the detection of on-road vehicles by allowing off-road detections to be filtered out. To exploit this intuition in an algorithmic framework, we formulate our problem as the minimization of a joint probabilistic objective function that combines (a) the classification of vehicle detections as true on-road vehicles vs. other detections and (b) a penalty for misalignment between the putative on-road vehicle detections and the vector roadmap under a parametric transformation. An explicit algorithm for registration is then developed in an Expectation Maximization (EM) framework that alternates between estimation of posterior probabilities that individual detections of vehicles correspond to on-road vehicles and the minimization of the weighted sum of minimum squared Euclidean distances from detection locations to the corresponding nearest points on the network of roads, where the weights are the estimated posterior probabilities that the detections correspond to on-road vehicles. Efficient computation of the latter metric is accomplished by using the associated distance transform [6], [7]. The parameter that estimates the fraction of detections that correspond to on-road vehicles is itself estimated in the EM framework, providing, as we demonstrate in our results, significant robustness to the quality of the initial detections of vehicle locations.

The problem of registering a captured image to a roadmap, or more generally to geo-referenced coordinates, has been addressed in prior work, although without leveraging vehicular detections. WAMI frames are usually captured from platforms equipped with Global Positioning System (GPS) and Inertial Navigation System (INS) which provide location and orientation information that are usually stored with the aerial

A. Elliethy is with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, USA (e-mail: ahmed.s.elliethy@rochester.edu).

G. Sharma is with the Department of Electrical and Computer Engineering, Department of Computer Science, and Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14627, USA (e-mail: gaurav.sharma@rochester.edu).

A preliminary version of part of the research presented in this paper appears in [1].

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the authors. The material consists of a supplementary document and an animated GIF.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2016.2601265

image as meta-data. This meta-data can be used to align the image with a road network extracted from an external Geographic Information System (GIS) source. However, the accuracy of the meta-data is limited and only provides an approximate registration. For "pixel-accurate" registration that can be exploited in computer vision and image analytics tasks, image-based registration methods are therefore necessary.

Registering an aerial image directly with a geo-referenced vector road map is a challenging task because of the differences in the nature of the data in the two formats: in one case the data consists of image pixel values whereas in the other it is described as lines/curves connecting a series of points. Because of the inherent differences in the data formats, one cannot readily define low/mid-level features that are invariant to the representations that can be used for finding corresponding points with conventional feature detectors, such as SIFT (Scale-Invariant Feature Transform) [8]. For static imagery, methods for aligning vector road maps to aerial imagery have been extensively investigated in the context of *conflation*, which refers to a process that fuses spatial representation from multiple data sources to obtain a new superior representation. In [9]–[11], road vector data are aligned with an aerial image by matching the road intersection points in both representations. The crucial element in these prior works is the detection of road intersections from the aerial image. With the availability of hyper-spectral aerial imagery, spectral properties and contextual analysis are used in [9] to detect these road intersections in the aerial scene. However, road segmentation is not robust, specially when roads in natural scenes are obscured by shadows from trees and nearby buildings. In [10], a Bayes classifier used to classify pixels as on-road or off-road, then a localized template matching used to detect the road intersections. To get a reasonable accuracy with the Bayes classifier, a large number of manually labeled training pixels is required for each dataset. In [11], corner detection is used to detect the road intersections, which is not reliable specially in high resolution aerial images that contain enough wide roads where the simple corner detection fails.

Work on registration of (non-static) WAMI frames to geo-referenced vector road maps has received comparatively less attention, even though the capability for performing such registration in a computationally efficient manner is crucial for a number of real/near real-time analysis applications for WAMI, such as large scale vehicle tracking [12]. Some of the prior work on this problem overcomes the problem posed by fundamentally different modalities of the WAMI and vector datasets by using an auxiliary geo-referenced image that is already aligned with the vector road map. The aerial image frames are then aligned to the auxiliary geo-referenced image by using conventional image feature matching methods. For example, in [13], for the purpose of vehicular tracking, the aerial frame is geo-registered with a geo-reference image and then a GIS database is used for road network extraction. This road network is used to regularize the matching of the current vehicle detections to the previous existing vehicular tracks. In an alternative approach that relies on 3D geometry [14], SIFT is used to detect correspondences between the ground features from a small footprint aerial video frame and geo-referenced image. This geo-registration helps to estimate the camera pose and depth map for each frame, and the depth map is used to segment the scene into building, foliage, and roads using a multi-cue segmentation framework. The process is computationally intensive and the use of the auxiliary geo-referenced image is still plagued by problems with identification of corresponding feature points because of the illumination changes, different capturing times, severe view point change in aerial imagery, and occlusion. State of the art feature point detectors and descriptors such as SIFT [8], and SURF (Speeded Up Robust Features) [15], often find many spurious matches that cause robust estimators such as RANSAC [16] to fail when estimating a homography. Also, these methods cannot work directly if the aerial video frames have a different modality (night-time infra-red for example) than the geo-referenced image. Last, but not least, a single homography represents the relation between two images when the scene is close to planar [17]. In WAMI, aerial video frames usually taken from oblique camera array to cover large ground area from moderate height and the scene usually contains non ground objects such as building, trees, and foliage. Thus the planar assumption does not necessarily hold across the entire imagery, although it is not unreasonable for the road network.

Compared with the existing approaches for registration of aerial images to a roadmap, our proposed methodology has several advantages:

- By posing the registration as the problem of aligning vehicle detections with the vector road network, we implicitly transfer both the aerial image and the geo-referenced one to a representation that can be easily matched in a computationally efficient manner using a distance transform [6], [7].
- The proposed methodology applies directly for different modalities of the captured imagery (night-time infra-red for example), where cross-modality image feature matching poses a particularly difficult challenge for image feature matching based methods.
- The use of a single homography is more appropriate as a registration transform in our framework compared with alternatives based on feature matching. This is because in WAMI, aerial video frames are usually captured from an oblique viewpoint to cover large ground area from moderate height and the scene usually contains non ground objects such as buildings, trees, and foliage. Thus a planar assumption that yields the homography as the relation between image coordinates in two images [17] does not necessarily hold across the entire imagery[1], although it is not unreasonable for the road network.
- The incorporation of classifications of vehicle detections as on-road vehicles or others and the EM algorithm for estimation of the classification parameters renders the algorithm robust to significant variations in the quality of the original vehicular detections.

Preliminary results from the research leading to this paper were presented in [1]. Compared with [1] where an ad

[1]In particular, feature points located at the base and near the top of multi-storey buildings egregiously violate the planar assumption.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2016.2601265

hoc algorithm was proposed, the present paper uses a more powerful EM framework that better accommodates inevitable "spurious" detections that do not correspond to on-road vehicles and provides significant robustness to the quality of these individual detections. We note that the framework we propose assumes that the WAMI scene contains a forked road network with vehicles, which is reasonable assumption for urban areas, and also for WAMI that covers a city scale ground area within each frame.

This paper is organized as follows. Section II presents our problem formulation and gives an overview of our proposed algorithm. In Section III, we describe our algorithm in greater depth. Results on aerial images with different modalities and a comparison against alternative methods are presented in Section IV. Section V summarizes our concluding remarks.

II. Problem Formulation and Algorithm Overview

A pictorial representation of our problem is shown in Fig. 1. We are provided with a vector map R_g that identifies the network of roads in a geographic area, where each road is represented as a sequence of spatial locations defined on a 2D Cartesian coordinate system (χ, ζ) derived from a geographical coordinate system (longitude and latitude). Specifically, the k^{th} road r^k is represented as a sequence of spatial locations $({}^r\chi_i^k, {}^r\zeta_i^k)$, $i = 1, 2, \ldots$ in the (χ, ζ) coordinate system. A WAMI frame $I(x, y)$ is obtained by capturing a portion of the same geographic area by an aerial WAMI sensor, where (x, y) are the pixel locations along the native Cartesian coordinates for the image sensor. Our goal is to estimate the parameter vector $\boldsymbol{\beta}$ of a geometric transformation $\mathcal{T}_{\boldsymbol{\beta}} : (x, y) \rightarrow (\chi, \zeta)$, that aligns the WAMI frame I with the road network R_g. The estimated optimal value of the parameters $\boldsymbol{\beta}$ should maximize a measure quantifying the similarity between the aligned WAMI frame and the road network, however, the definition of an appropriate similarity measure is challenging due to the differences in the nature of the data formats between the raster WAMI frame I, and the vector road network R_g.

In this paper, we propose a probabilistic framework that handle the challenges associated with aligning raster to vector data formats effectively by detecting the locations of the on-road vehicles in the WAMI frame I and aligning these detected locations with the network of roads in the vector road map R_g. However, accurate detection of the on-road vehicles is required to estimate an accurate road network alignment, and at the same time, accurate aligned road network can improve the accuracy of the detected on-road vehicles based on the proximity of detections to the aligned road network. As shown in Fig. 2, our proposed probabilistic framework exploits this synergy, and iteratively estimates the alignment of the detected on-road vehicles in the WAMI frame with the vector road network, where this estimated alignment in turn helps to estimate the probability of each detection corresponding to an on-road vehicle. The estimated probabilities are used to weight each detected vehicle appropriately for the alignment estimation process in the next iteration.

Formally, consider for a WAMI frame I, we have available N_v detected vehicles, with the location of the j^{th} detected

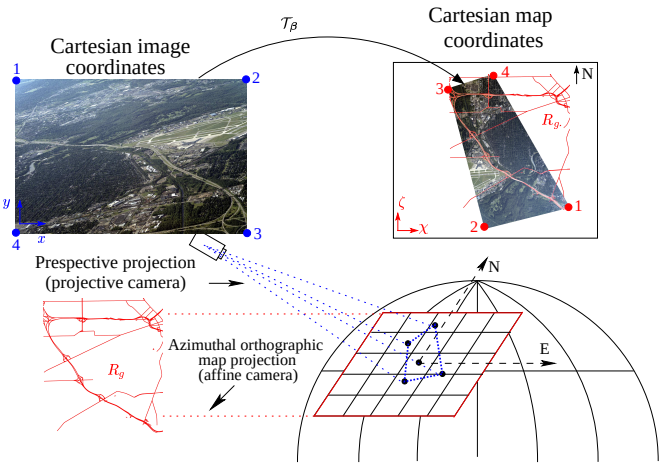


Fig. 1: Problem formulation : Given a vector road map R_g for a geographical area defined on a 2D Cartesian coordinate system (χ, ζ), and a WAMI frame $I(x, y)$ captured for the same area. Our goal is to estimate the geometric transformation parameter $\boldsymbol{\beta}$, that aligns the WAMI frame I with the road network R_g.

vehicle represented in the WAMI frame coordinate system by (x_j, y_j). For each detected vehicle, we define d_j as the minimum squared Euclidean distance (MSED) from its mapped location under the geometric transform $\mathcal{T}_{\boldsymbol{\beta}}$ to the road network R_g. Specifically,

$$d_j(\boldsymbol{\beta}) = \min_{i,k} D\left(({}^r\chi_i^k, {}^r\zeta_i^k), \mathcal{T}_{\boldsymbol{\beta}}(x_j, y_j)\right), \tag{1}$$

where $D(\mathbf{a}, \mathbf{b}) \equiv \|\mathbf{a} - \mathbf{b}\|_2^2$ is the squared Euclidean distance, and the minimization on the right hand side finds the point on the road network that is closest to the alignment transform mapped detection location. To account for detections that are spurious, i.e. do not correspond to on-road vehicles, we associate with each vehicle detection a latent variable $z_j \in \{0, 1\}$ that indicates whether the detected vehicle correspond to an on-road vehicle ($z_j = 1$) or not ($z_j = 0$). We model the distribution of this latent variable as a Bernoulli distribution parametrized by an unknown parameter $\gamma = p(z_j = 1)$, while we refer to as the *detection reliability parameter*. Furthermore, to account for the fact that on-road detections must be in close proximity to the (centerline specified in the vector representation of the) road, the conditional distribution of the MSED d_j, when $z_j = 1$, is modeled as an exponential distribution [18] with unknown parameter $\lambda = 1/\mathbb{E}\left[d_j | z_j = 1\right]$, where $\mathbb{E}[.]$ denote the expectation. When $z_j = 0$, the conditional distribution of d_j is modeled as uniform over the extent of the WAMI frame, analogous to the approach adopted in several other robust estimation problems [19]–[21]. We estimate the unknown parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \lambda, \gamma\}$ by maximizing the likelihood function modeled as

$$p(\mathbf{d}|\boldsymbol{\theta}) = \prod_{j=1}^{N_v} p(d_j|\boldsymbol{\theta}), \tag{2}$$

where $\mathbf{d} = [d_1, \ldots, d_{N_v}]^T \in \mathbb{R}^{N_v \times 1}$ is the vector of all

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2016.2601265

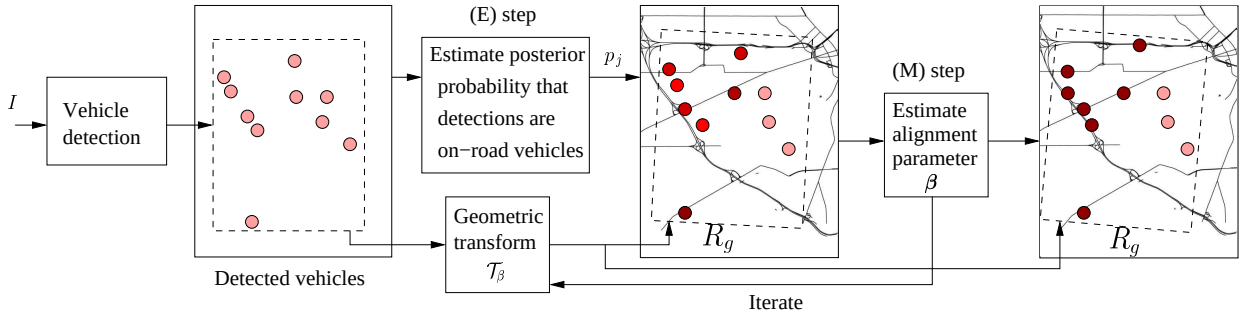


Fig. 2: Schematic of the proposed iterative algorithm showing main steps. The shaded circular dots depict detected vehicle locations, with shading indicating the estimated posterior probability that the dot corresponds to an on-road vehicle. Darker shades represent higher probability values.

MSEDs $\{d_j\}_{j=1}^{N_v}$, and

$$\begin{aligned} p(d_j|\boldsymbol{\theta}) &= \sum_{z_j\in\{0,1\}} p(d_j, z_j|\boldsymbol{\theta}) \\ &= p(z_j=1)p(d_j|\boldsymbol{\theta}, z_j=1) + \\ &\quad p(z_j=0)p(d_j|\boldsymbol{\theta}, z_j=0) \\ &= \gamma\lambda e^{-\lambda d_j} + \frac{(1-\gamma)}{M^2}, \end{aligned} \tag{3}$$

where $M = \sqrt{W^2+H^2}$, with W and H as the width and height, respectively, of the WAMI frame I in pixels.

Our formulation of the problem in terms of the latent variables $z_j, j = 1, 2, \ldots, N_v$ allows us to elegantly obtain the maximum likelihood estimate of the parameters $\boldsymbol{\theta}$ by employing the *Expectation Maximization* (EM) algorithm [22], details of which are outlined next.

The EM algorithm alternates between two steps, which are the Expectation (E) and Maximization (M) steps. In the (E) step, we find the posterior distribution of the latent variables z_j evaluated using the current estimate for the parameters $\boldsymbol{\theta}^t$. That posterior distribution is used to find the expectation of the *complete-data* log likelihood, which is maximized in the (M) step to compute a new estimate of the parameters $\boldsymbol{\theta}^{t+1}$. The expectation of the *complete-data* log likelihood is obtained as shown in Appendix A (up to a constant additive factor) as[2]

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = &\sum_{j=1}^{N_v} p_j \left[\ln(\gamma) + \ln(\lambda) - \lambda d_j\right] + \\ &(1-p_j)\left[\ln\left(1-\gamma\right)\right], \end{aligned} \tag{4}$$

where $p_j = p(z_j = 1|d_j, \boldsymbol{\theta}^t)$ is the posterior probability that the j^{th} detection corresponds to an on-road vehicle, and can be estimated using Bayes rule as

$$\begin{aligned} p_j &= \frac{p(d_j|z_j=1, \boldsymbol{\theta}^t)p(z_j=1|\boldsymbol{\theta}^t)}{p(d_j|\boldsymbol{\theta}^t)} \\ &= \frac{\gamma\lambda e^{-\lambda d_j}}{\gamma\lambda e^{-\lambda d_j} + \frac{(1-\gamma)}{M^2}}. \end{aligned} \tag{5}$$

Using the estimated p_j, we get the new estimate of the parameters by maximizing (4), i.e. $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$. From the first order optimality conditions from calculus of variations, the optimal parameters for maximizing (4) are obtained as

$$\gamma^* = \frac{\sum_{j=1}^{N_v} p_j}{N_v}, \tag{6}$$

and

$$\lambda^* = \frac{\sum_{j=1}^{N_v} p_j}{\sum_{j=1}^{N_v} p_j d_j}. \tag{7}$$

The optimal parameter $\boldsymbol{\beta}^*$ that maximizes (4), equivalently estimated by minimizing the objective function

$$f(\boldsymbol{\beta}) = \sum_{j=1}^{N_v} p_j d_j(\boldsymbol{\beta}), \tag{8}$$

with respect to $\boldsymbol{\beta}$. This means that the optimal transformation parameter vector $\boldsymbol{\beta}^*$ should map the detected vehicles' locations to be in a close proximity with the road network, and is estimated by minimizing the weighted sum of the squared Euclidean distances between each vehicle detection and the corresponding nearest point on the road network, where the weights corresponds to the posterior probability that a detection correspond to an on-road vehicle. This metric corresponds to a probabilistic formulation of the chamfer distance [24] with the minor modification of that we use squared distance instead of distance magnitude[3], and applies nicely to our problem as it measures how close the vehicle detections are to the road network. We discuss in details the numerical minimization of (8) in Section III-B.

III. Algorithm Details

A. Vehicle detection

As discussed previously, the inputs to our algorithm are vehicle detections in a WAMI frame. Several techniques have been reported in literature for detection of vehicles in aerial images [25]–[30]. One commonly used technique is the compensated frame difference [31], which we adopt here and describe next. We first estimate the geometric transform

[2]We adopt notation from [23].

[3]This modification ensures differentiability which we will exploit in Section III-B.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2016.2601265

(a) I (b) $\tilde{I}_0$ (c) I^d

Fig. 3: Compensated frame difference for successive WAMI images illustrating the two blobs per vehicle that result from the low frame rate in WAMI.

that aligns the WAMI frame with its predecessor I_0 by identifying locations of corresponding feature points in these two images and then estimating a homography [17] that maps the coordinates of feature points in I_0 to corresponding ones in I. Details of this process are sketched in Section S.III in the supplementary material. The estimated transform specifies a corresponding interpolation of I_0 that provides an estimate $\tilde{I}_0$ of the image that would be captured from the same viewpoint at I but at the time instant when I_0 was captured. Moving vehicles in the scene (and other moving objects as well as deviations from the homography model) cause significant differences between I and I_0 at the spatial locations of these objects (at the time instants for either frame but specified in the spatial coordinates of the frame I). We therefore estimate tentative vehicle regions as a binary image $I^d(x,y)$ corresponding to regions of significant deviation, computed as

$$I^d(x,y) = \begin{cases} 1, & \text{if } |I(x,y) - \tilde{I}_0(x,y)| \geq \tau \\ 0, & \text{otherwise} \end{cases}, \qquad (9)$$

where τ is a suitably determined *vehicle detection threshold parameter* that trades-off the sensitivity vs. precision of the detections in the presence of inevitable noise and other sources of variations in the images. Our algorithm is robust to the choice of τ over a wide range because the EM framework we use also estimates the fraction γ of detections that correspond to on-road vehicles (demonstrated in Section IV).

Although other techniques for vehicle detection can be used with our overall algorithm[4], we adopt the compensated frame difference for two reasons. First, as illustrated in the example shown in Fig. 3, due to the low frame rate in WAMI, each moving vehicle results in two blobs in I^d at the vehicle's locations[5] at the time instants when the two frames were captured. One of these blobs can be eliminated using the three frame difference [13]. However, in our case, because both blobs reside on the road network, we use both to our advantage. In other words, I^d contains blobs at locations of the vehicles' in the current frame and in the (compensated) past frame. The number of such blobs approximates two times the number of vehicles in the scene and using both locations helps improve the accuracy of the subsequent alignment between these detections and the road network as required by our algorithm. Second, the compensated frame difference is well suited for online real time applications. Some of the other vehicle detection techniques require learning vehicle models offline [25]–[28] or require complex modeling of the background [29], [30] which is estimated from several frames implying additional delay before the detected vehicle locations are obtained.

[4]Results for another vehicle detection technique are included in Section S.VII in the supplementary material.

[5]The locations of moving vehicles are presumed to correspond to the centroids of the blobs in I^d.

B. Aligning vehicle detections to the road network

Vector road maps typically provide the locations of the road segments in the spherical geographically-referenced coordinate system of latitude and longitude, where the earth is approximated as a sphere. We use the azimuthal orthographic map projection (AOMP) [32] to transform the road network from the geographical coordinate system to our 2D Cartesian map coordinate system (χ, ζ), which is shown in the bottom right corner of Fig. 1. A location on the surface of the earth sphere is chosen as a central point for the AOMP and the plane tangential to the sphere at the central point defines the plane for the map, with (χ, ζ) as orthogonal coordinates for the plane. The points specified by latitude and longitude on the earth sphere are projected orthogonally, i.e. along lines perpendicular to the plane, from the sphere to the tangent plane to obtain corresponding locations in the (χ, ζ) coordinate system. We select the map central point to be approximately the center of the geographical area that is being captured by our WAMI sensor; this approximately minimizes the distortion of distances between the points on the sphere and in the planar representation. The AOMP can be viewed as a virtual affine camera with the camera center located at infinity and the image plane coincident with the AOMP tangent plane (shown in Fig 1). The same scene is projected to the WAMI 2D Cartesian image coordinates (x, y) using the projective camera that is used to capture the scene. Assuming a planar scene, we can relate the (χ, ζ), and the (x, y) coordinate systems through a homography [17]. Thus, the objective function in (8) can be specifically formulated as

$$f(\boldsymbol{\beta}) = \sum_{j=1}^{N_v} p_j \min_{i,k} D\left(\mathbf{p}^r_{i,k}, \mathbf{H}_{\boldsymbol{\beta}} \mathbf{p}^v_j\right), \qquad (10)$$

where $\mathbf{H}_{\boldsymbol{\beta}}$ is a homography defined by the parameter[6] $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_8]^T$, and $\mathbf{p}^v_j = [x_j, y_j, 1]^T$, $\mathbf{p}^r_{i,k} = [{}^r\chi^k_i, {}^r\zeta^k_i, 1]^T$ are the homogeneous coordinates of the j^{th} vehicle detection in the WAMI frame I and the i^{th} location in the k^{th} road in R_g, respectively. To align the vehicle detection locations with the road network R_g, we seek the optimal homography parameter vector $\boldsymbol{\beta}^*$ that minimizes the objective function $f(\boldsymbol{\beta})$.

For performing the minimization in (10), we adopt the Levenberg-Marquardt (LM) [33] non-linear least squares iterative optimization algorithm. In each iteration, the LM algorithm estimates the parameter update vector $\boldsymbol{\delta} \in \mathbb{R}^{8\times 1}$ such that the value of the objective function is reduced when moving from $\boldsymbol{\beta}$ to $\boldsymbol{\beta} + \boldsymbol{\delta}$, with the parameters converging to a minima of the objective function with the progression

[6]Appendix B gives the detailed expression for the homography transformation in terms of the parameter vector $\boldsymbol{\beta}$

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2016.2601265

of iterations. The parameter update vector $\boldsymbol{\delta}$ is obtained by solving the linear system of equations

$$(\mathbf{A} + \eta \mathbf{I})\boldsymbol{\delta} = -\mathbf{b}, \tag{11}$$

where η is a non-negative damping parameter automatically adjusted at each step [33] to determine the step size, $\mathbf{I}$ is the identity matrix, and $\mathbf{b} \in \mathbb{R}^{8\times 1}$ is the gradient of $f(\boldsymbol{\beta})$, computed as

$$\mathbf{b} = \frac{\partial f}{\partial \boldsymbol{\beta}} = -2\sum_{j=1}^{N_v} p_j \mathbf{J}_j^T \mathbf{r}_j, \tag{12}$$

with $\mathbf{r}_j = \left(\min_{i,k}(\mathbf{p}_{i,k}^r - \mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v)\right)$ defined as the residual vector; $\mathbf{J}_j \in \mathbb{R}^{3\times 8}$ is the Jacobian matrix computed at the transformed point $\mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v$, computed as

$$\mathbf{J}_j = \frac{\partial \mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v}{\partial \boldsymbol{\beta}} = \left[\frac{\partial \mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v}{\partial \beta_1}, \ldots, \frac{\partial \mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v}{\partial \beta_8}\right]; \tag{13}$$

and $\mathbf{A} \in \mathbb{R}^{8\times 8}$ is the approximation of the Hessian matrix, obtained as

$$\mathbf{A} = \sum_{j=1}^{N_v} \mathbf{J}_j^T \mathbf{J}_j. \tag{14}$$

Explicit expressions for the required derivatives for the proceeding steps are provided in Appendix B. The objective function in (10) and the required derivatives for the LM algorithm can be efficiently calculated using a distance transform [6], [7]. To compute the residual vector $\mathbf{r}_j$ in (12), we augment the distance transform computation to additionally provide the coordinates $({}^r\chi_i^k, {}^r\zeta_i^k)$ of the location in R_g that achieves the minimum for the terms in (10).

Because the LM algorithm converges to a local minima, it is important to initialize the algorithm with a good starting point. The approximate geographical coordinates of the four corners of the WAMI frame are included in the meta-data, from which corresponding locations in the coordinate system (χ, ζ) for the geo-referenced road map R_g can be calculated. From the correspondences of the locations of these non-collinear corner points between the (x, y) and the (χ, ζ) coordinates, we employ the direct linear transformation algorithm (DLT) [17] to estimate the initial solution $\boldsymbol{\beta}^0$. Starting from this initialization, at iteration n, the homography parameter vector is updated using the LM iteration $\boldsymbol{\beta}^n = \boldsymbol{\beta}^{n-1} + \boldsymbol{\delta}$, and the process is continued until convergence[7]. The LM optimization is integrated into the overall EM iterations described in Section II. The resulting overall algorithm is shown in Algorithm 1.

As noted in the introduction, our formulation has several advantageous features. First, by exploiting the vehicle detections in aerial WAMI frames, we implicitly transfer the WAMI image to a representation that can be efficiently aligned with the vector road network using the distance transform. Second, our formulation does not depend on the specific type of imaging sensor used to capture the WAMI frame and can be readily applied to different image modalities, for example for night-time infra-red (IR) imagery. Third, the use of a single homography is more appropriate as a registration transform in our framework compared with alternatives based on feature matching, because planar assumption for the road network is very reasonable, whereas the planar assumption does not necessarily hold across the entire imagery. Finally, the use of the EM algorithm and introduction of the latent variables $\{z_j\}$ in our formulation make the method robust to spurious detections.

[7]An animated GIF that shows a visual representative example of the LM iteration process is provided in the supplementary material included with this paper.

Algorithm 1: Proposed algorithm for registering a WAMI frame to a vector road map

Input : Roadmap R_g, WAMI frame I, and immediately proceeding frame I_0
Output: Parameters for registration and detection reliability $\boldsymbol{\theta}^* = \{\boldsymbol{\beta}^*, \gamma^*, \lambda^*\}$

Vehicle detections:
Estimate homography that aligns I_0 to I, and compute putative vehicle locations in I using (9);
Initialization:
Estimate initial value $\boldsymbol{\beta}^0$ of registration parameter vector using meta-data for frame I;
$t \leftarrow 0$; $\boldsymbol{\theta}^t \leftarrow \{\boldsymbol{\beta}^t, \lambda^t, \gamma^t\}$;
repeat */*EM iterations*/*
  E step:
  Compute posterior probabilities of detection reliability $\{p_j\}_{j=1}^{N_v}$ using (5) with $\boldsymbol{\theta}^t$;
  M step:
  $n \leftarrow 0$; $\tilde{\boldsymbol{\beta}}^n \leftarrow \boldsymbol{\beta}^t$;
  repeat */*LM iterations*/*
    Estimate LM update $\boldsymbol{\delta}$ for registration parameter vector using (11);
    $n \leftarrow n + 1$; $\tilde{\boldsymbol{\beta}}^n \leftarrow \tilde{\boldsymbol{\beta}}^{n-1} + \boldsymbol{\delta}$;
  until $(\|\boldsymbol{\delta}\|_2 \leq \epsilon)$ *or* $(n >$ *max iterations*$)$;
  $\boldsymbol{\beta}^* \leftarrow \tilde{\boldsymbol{\beta}}^n$; Compute γ^* using (6); Compute λ^* using (7);
  Update:
  $t \leftarrow t + 1$; $\boldsymbol{\theta}^t \stackrel{\text{def}}{=} \{\boldsymbol{\beta}^t, \lambda^t, \gamma^t\} \leftarrow \{\boldsymbol{\beta}^*, \lambda^*, \gamma^*\}$;
until $(\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^{t-1}\|_2 \leq \epsilon_1)$ *and* $(|\lambda^t - \lambda^{t-1}| \leq \epsilon_2)$ *and* $(|\gamma^t - \gamma^{t-1}| \leq \epsilon_3)$ *//Convergence of parameters*;
Output current parameter estimate $\boldsymbol{\theta}^t$ as $\boldsymbol{\theta}^*$;

IV. Results

We evaluated our algorithm on three WAMI datasets captured over different geo-graphical areas, and containing images that are sensed in either the visible spectral bands (red, green, blue) or mid-wave infra red-band (single band). Specifically the datasets are: (1) the CORVUS(V) visible band dataset, which was recorded using the CorvusEye 1500 Wide-Area Airborne System [5] for the Rochester, NY region, (2) the CORVUS(IR) mid-wave infra-red band dataset recorded at night with the same system for the Lakeland, FL region, (3) the Wright-Patterson Air Force Base (WPAFB) 2009 visible band dataset [34], which was recorded over the WPAFB, OH region.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2016.2601265

The WAMI frames provided by the three datasets are stored in the NITF 2.1 format [35], that contains both the imagery itself and meta-data information that includes the approximate geographical coordinates for the four corners associated with each aerial WAMI frame. Additional details regarding the specific geo-graphical coordinates for the captured regions and the encoding format for the WAMI datasets are provided in the supplementary material in Sections S.I and S.II. For the vector road map, we used OpenStreetMap (OSM) [36], which is a collaborative project that uses free data sources, such as Volunteered Geographic Information (VGI) [37], to create a free editable map of the world. In our experiments, the EM algorithm uses the initial parameter values $\lambda^0 = 10^{-5}$ and $\gamma^0 = 0.5$. These values are chosen so that the initial posterior probabilities p_j in (5) fall-off relatively slowly with increasing values of the distance d_j, which models the intuition that the initial alignment provides little discrimination between on-road vehicles and other detections. The threshold parameter in (9) is set empirically as $\tau = 0.15$ (for images on a $0-1$ scale) and an initial value of $\eta = 0.01$ is used for the LM algorithm in (11).

We compare our proposed method with three alternative methods which we will refer to as "Meta-data Based Alignment (MBA)", "SIFT matching with auxiliary geo-referenced image (SBA)", and the method in [1], which is a preliminary version of the method presented here. The MBA method simply uses the aerial WAMI frame meta-data to estimate the alignment between that frame and the road network R_g from the correspondences of the locations of the corner points in the image and the map coordinates using the DLT algorithm [17] as discussed previously. The alignment obtained from the MBA method approximately ortho-rectifies the WAMI frame, since the road network R_g is already defined on the orthographic coordinate system (χ, ζ). The SBA method refines the estimated alignment from the MBA method by matching SIFT features between the approximately ortho-rectified WAMI frame obtained using the MBA method and an auxiliary geo-referenced image taken from Google Maps. To improve SIFT based matching, for each SIFT feature point in one image, we search for a prospective matched feature point in the other image only within a circle with radius r that corresponds to the uncertainty of the estimated alignment using the MBA method. We set the radius r by determining the maximum spatial error for the alignment provided by the MBA method. After obtaining these putative correspondences, we use RANSAC [16] to filter out the incorrect matches and to estimate the final transformation between the geo-referenced image and the aerial WAMI frame.

Results for the proposed method and for the MBA and SBA[8] methods are presented in Fig. 7 in an image format that facilitates visual evaluation of the methods. In the images, the estimated alignment of the road network is visualized[9] by superimposing the estimated road locations as transparent highlighted tracks overlaid on the WAMI frame with different color highlights for the different methods. Additional results in the same visual format are presented in Fig. S.1, Fig. S.2, and Fig. S.3 in the supplementary material. From these images, we can see that the proposed method offers a significant enhancement over MBA which depends only on the meta-data to get an aligned road network and over SBA which uses SIFT and auxiliary geo-referenced Google map image. The MBA method has significant errors because of the inaccuracy of the meta-data parameters due to the limited accuracy of on-board devices for recording location and orientation. The SBA method does not improve significantly because of spurious correspondences found by the SIFT matching between the aerial image and the Google map image. Even though these images are for the same region and already approximately geo-registered, they have significant differences due to severe view point change, different illumination, and different capturing times. Moreover, when applied to the infra-red frames, the SBA method yields a less accurate result even compared with the MBA method because SIFT is not invariant to the different modalities between the night-time IR WAMI frames and the visible-bands Google map images. This mismatch in sensing modalities, causes most matches to be spurious, resulting in a poor estimated alignment. Our proposed method does not encounter the challenges associated with aligning images captured under these different conditions because it aligns vehicle detections to the road network by minimizing the distances between them, and thus provides accurate alignments for both visual and infra-red modalities.

To provide quantitative comparison between the methods, we label the ground truth road network for few frames for different test areas[10] by manually identifying the road segments within each frame. For each identified road segment, we store its start, end points, and its road width in meters which are obtained from measuring the actual road width in Google Map. Our quantitative analysis uses three metrics: *chamfer distance*, *relative positional accuracy*, and *precision-recall*.

The chamfer distance is computed in pixel units as the mean value of the Euclidean distances between each point in the estimated aligned road network and its closest point in the ground truth road network. Table I shows the chamfer distance between the ground truth road network and the aligned road network generated from our proposed method, the MBA method, the SBA method, and the preliminary version of our work presented in [1]. In our chamfer distance computation, we represent the ground truth roads by the roads' actual widths, while the roads in the aligned road network are represented by its center line. Table I highlights three important points. First, it reinforce the conclusions seen from the visual images. The proposed method has a much lower value for the chamfer distance highlighting the fact that the proposed method offers a significant improvement over both the MBA and SBA methods for both visual and infra-red frames. Second, compared to the preliminary version of our work presented in [1], the more complete framework that we

[8]All results of SBA method, are reported using the radius r that gives the best result.

[9]We present our visual results in the coordinate system of the WAMI frame by applying the inverse of the estimated transformation to the vector road network

[10]We generate the ground truth for few test areas in each dataset because it is very tedious to manually extract roads in WAMI image.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2016.2601265

present here provides better accuracy than [1], as it takes into consideration the reliability of the vehicle detector and weights each detected vehicle appropriately before minimizing the distance between the detected vehicles and the road network. Finally, the SBA method provides little enhancement over the MBA method for visual frames, and performs much worse than the MBA method in the case of infra-red frames, which indicates the challenges associated with SIFT as a feature matching technique when dealing with these different imaging conditions.

Dataset	Test area	MBA	SBA	method in [1]	Proposed method
CORVUS (V)	Area 1	28.22	17.1	6.36	**3.95**
	Area 2	122.28	83.09	9.30	**2.07**
	Area 3	36.95	26.49	8.69	**3.45**
	Area 4	87.35	87.29	6.68	**5.21**
CORVUS (IR)	Area 5	450.19	462.76	3.15	**2.13**
	Area 6	104.28	387.84	4.25	**2.14**
	Area 7	179.13	266.85	5.12	**3.11**
	Area 8	81.38	116.37	17.94	**11.34**
WPAFB	Area 9	14.19	11.87	9.04	**3.15**
	Area 10	16.03	14.23	6.15	**4.12**
	Area 11	13.09	10.84	8.28	**3.36**
	Area 12	13.90	10.18	8.86	**4.43**

TABLE I: Chamfer distance (in pixels) between the ground truth road network and the road network generated using the MBA method, the SBA method, the preliminary version of our work presented in [1], and our proposed method. The test areas for the datasets are specified in Table S.I in the supplementary material.

The second metric which is the relative positional accuracy, measures the fraction of aligned road pixels that are within a certain threshold distance from the ground truth roads' center line. This threshold distance is set as a factor κ of each road's width from the ground truth road as shown in Fig. 4 (a) for $\kappa = 0.5$. We vary the factor κ and plot the corresponding relative positional accuracy for Area 1 and Area 2 in Fig. 5 (a) and (b), respectively, for our proposed method, the MBA method, the SBA method, and the method in [1]. Our method provides the largest area under the curve (AUC) for the relative positional accuracy plot compared to the other methods, which highlights the improvement of the road alignment accuracy of our proposed method compared to the other methods.

Finally, we present the precision-recall performance for the methods compared. Figure 4 (b) defines how we estimate the true positives (TP), the false positives (FP), and the false negatives (FN), which are used to compute precision and recall. For our precision-recall computation, we represent the ground truth roads using their actual widths and represent each estimated aligned road by its center line. Then we progressively increase the width of the estimated aligned roads using the morphological dilation operation and record both precision and recall as the dilation amount is varied. Precision-recall plot for Test Area 1 is shown in Fig. 5 (c). Once again, the improvement offered by the proposed method over the other methods is apparent from the plot.

Both the preliminary version of our work presented in [1], and our proposed method in this paper rely on detected vehicles to obtain the aligned road network, and the accuracy of

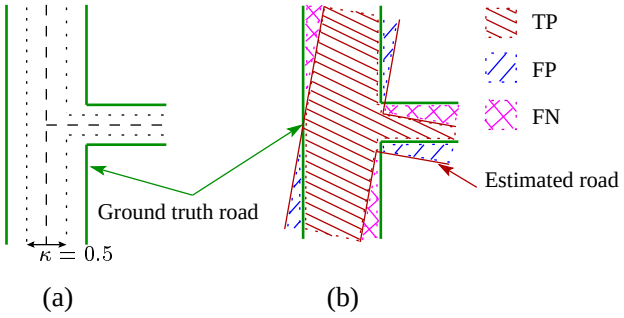


Fig. 4: Illustration showing the computation of the relative positional accuracy and the precision-recall quantitative measures. (a) shows an example of a threshold distance from the ground truth roads' center line with a width $\kappa = 0.5$ of the actual road width. (b) defines the true positive (TP), false positive (FP), and false negative (FN) regions used to calculate precision and recall.

the alignment depends on the reliability of the vehicle detector. As described in Section III-A, we adopt the compensated frame difference technique with a threshold parameter τ to detect tentative locations for vehicles. In order to compare the robustness of our proposed method in this paper with its preliminary version presented in [1], we vary the vehicle detection threshold parameter τ and compute the chamfer distance between the ground truth road network and the road network obtained from each method. A plot of the computed chamfer distance against τ is shown in Fig. 6. Compared with the method in [1], our proposed method provides an accurate aligned road network across the wide range of different vehicle detection thresholds compared with the method in [1]. This robustness of the proposed method arises specifically from the objective function formulation of (8) where the distance penalty to be minimized for each detected vehicle is weighted by the estimated posterior that it is an on-road vehicle, which minimizes the impact of spuriously detected vehicle on the estimated alignment. In particular, this highlights the advantage of the EM framework we introduce in this paper.

As already noted in the introduction, our methodology relies on having a forked network of roads in the WAMI scene with an adequate number of moving vehicles. These requirements are usually met over the relatively large geographic areas covered by WAMI imagery, particularly for applications involving traffic analysis in urban WAMI scenes. To assess the impact of the number of vehicles on the accuracy of registration, we also performed a semi-synthetic experiment in which we compared the registration accuracy as a function of the number of vehicle detections, where the number of vehicle detections was reduced by random subsampling of our initial set of detections used for the results presented in Fig. 7 (a). The results summarized in the supplementary data indicate that the accuracy of the registration is maintained even when the number of detections is a relatively small fraction of the total detections in our original experiments. The results in Fig. S.4 show that there is minimal degradation in the accuracy of the registration as the number of detections is subsampled to include as few as 20% of the original detections, although

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2016.2601265

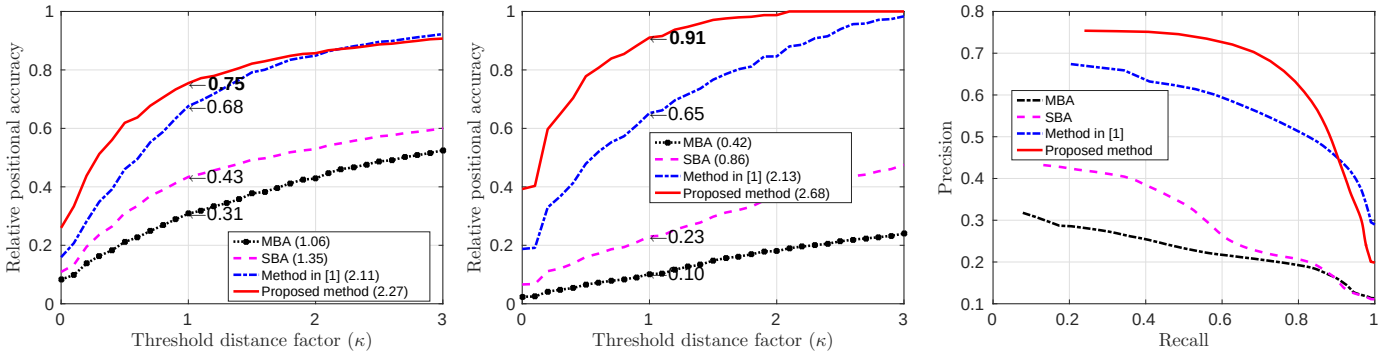


Fig. 5: The relative positional accuracy and precision-recall quantitative metrics for the road network generated using the MBA method, the SBA method, the preliminary version of our work presented in [1], and our proposed method here. The relative positional accuracy for test areas 1 and 2 are shown in (a), (b), respectively, while the precision-recall plot for test area 1 is shown in (c).

when the number of detections is further reduced to 15%, the registration error increases rapidly. While further exploration that also examines the dependence on the distribution of vehicles is of interest, it requires a significant number of additional WAMI datasets to meaningfully represent practical situations and is beyond the scope of the present work.

In some applications, accurate geo-registration is desired for an entire sequence of WAMI frames and not just for a single frame. In these situations, it is beneficial to decompose the alignment task into two subtasks: Alignment of occasional "key" frames can be obtained using the methodology proposed in this paper and the alignment of remaining frames can be obtained by estimating the geometric transforms for registering sequential frames from one key frame to the next. The latter task is readily accomplished by using conventional feature matching based estimation of a global registration homography between successive frames. SIFT and SURF feature matching works quite well in this scenario because the difference in viewpoint and illumination between successive frames is quite small. Note that the process of estimation of global frame-to-frame motion as a homography is already a part of the method we use in this paper for detecting vehicles and in our recent work [12], a similar decomposition has been effectively incorporated into the problem formulation for tracking.

V. Conclusion

The EM framework proposed in this paper offers a novel methodology for accurately registering vector road maps to wide area motion imagery (WAMI) by exploiting the locations of on-road vehicles detected in the WAMI frame. Compared with alternative approaches, the framework has the advantages that it eliminates the need for feature matching, is applicable across different imaging modalities (e.g. visible/IR), is better matched with the use of a planar homography as the registration transform, and is robust to spurious detections through the estimation of a detection reliability parameter within the EM iterations. Results obtained for test datasets captured using both visual and infra-red sensors, show the effectiveness of the proposed methodology. Both visually and in terms of numerical metrics for alignment accuracy, the proposed method offers a significant improvement over available alternatives.

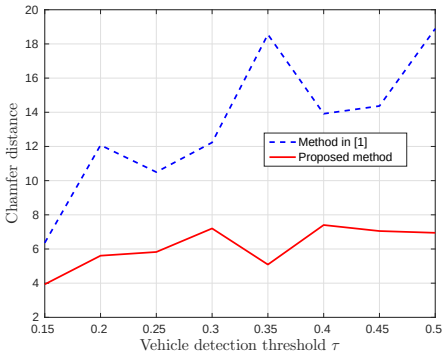


Fig. 6: Chamfer distance between the ground truth road network and the road network for the method in [1] and for the proposed method as a function of different values of the vehicle detection threshold parameter τ.

Appendix A

The complete data likelihood is $p(\mathbf{z}, \mathbf{d}|\boldsymbol{\theta}) = \prod_{j=1}^{N_v} p(z_j, d_j|\boldsymbol{\theta})$, where $\mathbf{z} = [z_1, \ldots, z_{N_v}]^T$, and

$$p(z_j, d_j|\boldsymbol{\theta}) = \begin{cases} \gamma\lambda e^{-\lambda d_j}, & \text{if } z_j = 1 \\ \frac{(1-\gamma)}{M^2}, & \text{if } z_j = 0 \end{cases}$$

The complete data log likelihood is defined as

$$\ln \prod_{j=1}^{N_v} p(z_j, d_j|\boldsymbol{\theta}) = \sum_{j=1}^{N_v} \ln \left(p(z_j, d_j|\boldsymbol{\theta})\right). \qquad (15)$$

The expectation of the complete data log likelihood is evaluated using the posterior probability computed using the current

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2016.2601265

estimate of the parameters $\boldsymbol{\theta}^t$, as

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^t) &= \sum_{j=1}^{N_v}\sum_{z_j\in\{0,1\}} p(z_j|d_j,\boldsymbol{\theta}^t)\ \ln\left(p(z_j,d_j|\boldsymbol{\theta})\right)\\ &= \sum_{j=1}^{N_v} p_j\left[\ \ln(\gamma) + \ \ln(\lambda) - \lambda d_j\right] + \\ &\quad (1-p_j)\left[\ \ln(1-\gamma) - 2\ \ln(M)\right] \end{aligned} \tag{16}$$

which reduces to (4) after dropping the constant term $\ln(M)$.

APPENDIX B

The homography $\mathbf{H}_{\boldsymbol{\beta}}$ maps a 2D location (x_j, y_j) in the WAMI frame I to the 2D location (χ_j, ζ_j) in the R_g coordinate system, given by

$$\chi_j = \frac{\beta_1 x_j + \beta_2 y_j + \beta_3}{\beta_7 x_j + \beta_8 y_j + 1}, \quad \zeta_j = \frac{\beta_4 x_j + \beta_5 y_j + \beta_6}{\beta_7 x_j + \beta_8 y_j + 1}.$$

In homogeneous coordinates [17], these equations are equivalently represented as the matrix multiplication

$$w_j \begin{pmatrix}\chi_j\\ \zeta_j\\ 1\end{pmatrix} = \mathbf{H}_{\boldsymbol{\beta}} \begin{pmatrix}x_j\\ y_j\\ 1\end{pmatrix} = \begin{pmatrix}\beta_1 & \beta_2 & \beta_3\\ \beta_4 & \beta_5 & \beta_6\\ \beta_7 & \beta_8 & 1\end{pmatrix} \begin{pmatrix}x_j\\ y_j\\ 1\end{pmatrix}, \tag{17}$$

where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_8]^T$ is the transformation parameters and $w_j = x_j\beta_7 + y_j\beta_8 + 1$ is the scaling factor for the homogeneous coordinates on the left-hand-side [17]. The derivatives in (13) are then obtained as

$$\begin{aligned}
\frac{\partial \mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v}{\partial\beta_1} &= \left[\frac{x_j}{w_j}, 0, 0\right]^T, \quad \frac{\partial \mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v}{\partial\beta_2} = \left[\frac{y_j}{w_j}, 0, 0\right]^T,\\
\frac{\partial \mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v}{\partial\beta_3} &= \left[\frac{1}{w_j}, 0, 0\right]^T, \quad \frac{\partial \mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v}{\partial\beta_4} = \left[0, \frac{x_j}{w_j}, 0\right]^T,\\
\frac{\partial \mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v}{\partial\beta_5} &= \left[0, \frac{y_j}{w_j}, 0\right]^T, \quad \frac{\partial \mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v}{\partial\beta_6} = \left[0, \frac{1}{w_j}, 0\right]^T,\\
\frac{\partial \mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v}{\partial\beta_7} &= \frac{-x_j}{w_j}\left[\chi_j, \zeta_j, 0\right]^T,\\
\frac{\partial \mathbf{H}_{\boldsymbol{\beta}}\mathbf{p}_j^v}{\partial\beta_8} &= \frac{-y_j}{w_j}\left[\chi_j, \zeta_j, 0\right]^T.
\end{aligned}$$

ACKNOWLEDGEMENT

We thank Bernard Brower of Harris Corporation for making available the CorvusEye [5] WAMI datasets used in this research.

REFERENCES

[1] A. Elliethy and G. Sharma, "Vector road map registration to oblique wide area motion imagery by exploiting vehicles movements," in *IS&T Electronic Imaging: Video Surveillance and Transportation Imaging Applications*, San Francisco, California, 2016, pp. VSTIA–520.1–8. [Online]. Available: http://ist.publisher.ingentaconnect.com/contentone/ist/ei/2016/00002016/00000003/art00008

[2] K. Palaniappan, R. M. Rao, and G. Seetharaman, "Wide-area persistent airborne video: Architecture and challenges," in *Distributed Video Sensor Networks*. Springer, 2011, pp. 349–371.

[3] E. Blasch, G. Seetharaman, S. Suddarth, K. Palaniappan, G. Chen, H. Ling, and A. Basharat, "Summary of methods in wide-area motion imagery (WAMI)," in *Proc. SPIE*, vol. 9089, 2014, pp. 90 890C–90 890C–10.

[4] R. Porter, A. Fraser, and D. Hush, "Wide-area motion imagery," *IEEE Sig. Proc. Mag.*, vol. 27, no. 5, pp. 56–65, Sept 2010.

[5] "CorvusEye™1500 Data Sheet," http://www.exelisinc.com/solutions/corvuseye1500/Documents/CorvusEye500DataSheetAUG14.pdf.

[6] A. Rosenfeld and J. L. Pfaltz, "Sequential operations in digital picture processing," *Journal of the ACM*, vol. 13, no. 4, pp. 471–494, 1966.

[7] P. Felzenszwalb and D. Huttenlocher, "Distance transforms of sampled functions," Cornell University, Tech. Rep., 2004.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] W. Song, J. Keller, T. Haithcoat, and C. Davis, "Automated geospatial conflation of vector road maps to high resolution imagery," *IEEE Trans. Image Proc.*, vol. 18, no. 2, pp. 388–400, Feb 2009.

[10] C.-C. Chen, C. A. Knoblock, and C. Shahabi, "Automatically conflating road vector data with orthoimagery," *GeoInformatica*, vol. 10, no. 4, pp. 495–530, 2006.

[11] C.-C. Chen, C. A. Knoblock, C. Shahabi, Y.-Y. Chiang, and S. Thakkar, "Automatically and accurately conflating orthoimagery and street maps," in *Proc. ACM Int. Workshop on Geographic Information Systems*. ACM, 2004, pp. 47–56.

[12] A. Elliethy and G. Sharma, "A joint approach to vector road map registration and vehicle tracking for wide area motion imagery," in *IEEE Intl. Conf. Acoust., Speech, and Signal Proc.*, 2016, pp. 1100–1104, Shanghai, China, 20-25 March 2016.

[13] J. Xiao, H. Cheng, H. Sawhney, and F. Han, "Vehicle detection and tracking in wide field-of-view aerial video," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, June 2010, pp. 679–684.

[14] J. Xiao, H. Cheng, F. Han, and H. Sawhney, "Geo-spatial aerial video processing for scene understanding and object tracking," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, June 2008, pp. 1–8.

[15] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comp. Vis. and Image Understanding.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[16] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[17] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge University Press, 2003.

[18] C. M. Grinstead and J. L. Snell, *Introduction to probability*. American Mathematical Soc., 2012.

[19] P. H. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Comp. Vis. and Image Understanding.*, vol. 78, no. 1, pp. 138–156, 2000.

[20] R. Horaud, F. Forbes, M. Yguel, G. Dewaele, and J. Zhang, "Rigid and articulated point registration with expectation conditional maximization," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 33, no. 3, pp. 587–602, 2011.

[21] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. and Remote Sensing*, vol. 53, no. 12, pp. 6469–6481, 2015.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[23] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[24] 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.

[25] M. Teutsch and W. Kruger, "Robust and fast detection of moving vehicles in aerial videos using sliding windows," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog. Workshops*, June 2015, pp. 26–34.

[26] H. Grabner, T. T. Nguyen, B. Gruber, and H. Bischof, "On-line boosting-based car detection from aerial images," *ISPRS*, vol. 63, no. 3, pp. 382 – 396, 2008.

[27] K. Palaniappan, F. Bunyak, P. Kumar, I. Ersoy, S. Jaeger, K. Ganguli, A. Haridas, J. Fraser, R. Rao, and G. Seetharaman, "Efficient feature extraction and likelihood fusion for vehicle tracking in low frame rate airborne video," in *Intl. Conf. on Info. Fusion*, July 2010, pp. 1–8.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2016.2601265

[28] X. Shi, H. Ling, E. Blasch, and W. Hu, "Context-driven moving vehicle detection in wide area motion imagery," in *IEEE Intl. Conf. on Pattern Recog.*, Nov 2012, pp. 2512–2515.
[29] V. Reilly, H. Idrees, and M. Shah, "Detection and tracking of large number of targets in wide area surveillance," in *Proc. European Conf. Computer Vision*, 2010, vol. 6313, pp. 186–199.
[30] J. Prokaj, M. Duchaineau, and G. Medioni, "Inferring tracklets for multi-object tracking," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog. Workshops*, June, pp. 37–44.
[31] A. M. Tekalp, *Digital Video Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1995.
[32] J. P. Snyder, *Map projections–A working manual*. US Government Printing Office, 1987, vol. 1395.
[33] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York: Springer, 2006.
[34] "AFRL WPAFB 2009 data set," https://www.sdms.afrl.af.mil.
[35] National imagery transmission format (NITF) standard, version 2.1. [Online]. Available: http://www.gwg.nga.mil/ntb/baseline/index.html
[36] "OpenStreetMap," http://www.openstreetmap.org.
[37] M. F. Goodchild, "Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0," *Intl. J. of Spatial Data Infrastructures Research*, vol. 2, pp. 24–32, 2007.

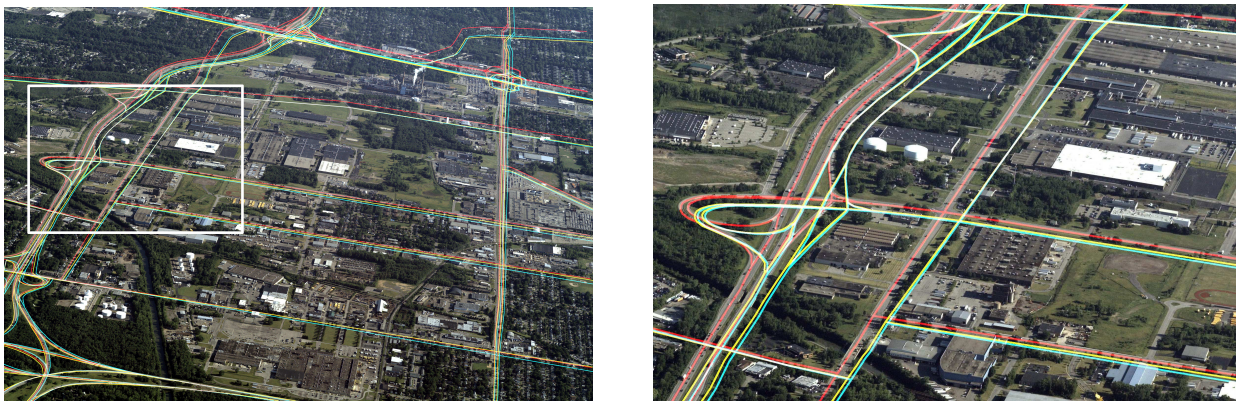
Ahmed Elliethy (S'15) received the B.Sc. degree (excellent with honors) in computer engineering and the M.Sc. degree in electrical engineering from the Military Technical College, Cairo, Egypt, in 2003, and 2010, respectively. He is currently pursuing his Ph.D. degree at University of Rochester. His research interests are computer vision and security, specifically, multiple object tracking, optical flow, and media forensics.

Gaurav Sharma (S'88–M'96–SM'00–F'13) is a professor at the University of Rochester in the Department of Electrical and Computer Engineering, in the Department of Computer Science and in the Department of Biostatistics and Computational Biology. From 2008–2010, he served as the Director for the Center for Emerging and Innovative Sciences (CEIS), a New York state funded center for promoting joint university-industry research and technology development, which is housed at the University of Rochester. He received the BE degree in Electronics and Communication Engineering from Indian Institute of Technology Roorkee (formerly Univ. of Roorkee), India in 1990; the ME degree in Electrical Communication Engineering from the Indian Institute of Science, Bangalore, India in 1992; and the MS degree in Applied Mathematics and PhD degree in Electrical and Computer Engineering from North Carolina State University, Raleigh in 1995 and 1996, respectively. From Aug. 1996 through Aug. 2003, he was with Xerox Research and Technology, in Webster, NY, initially as a Member of Research Staff and subsequently at the position of Principal Scientist.

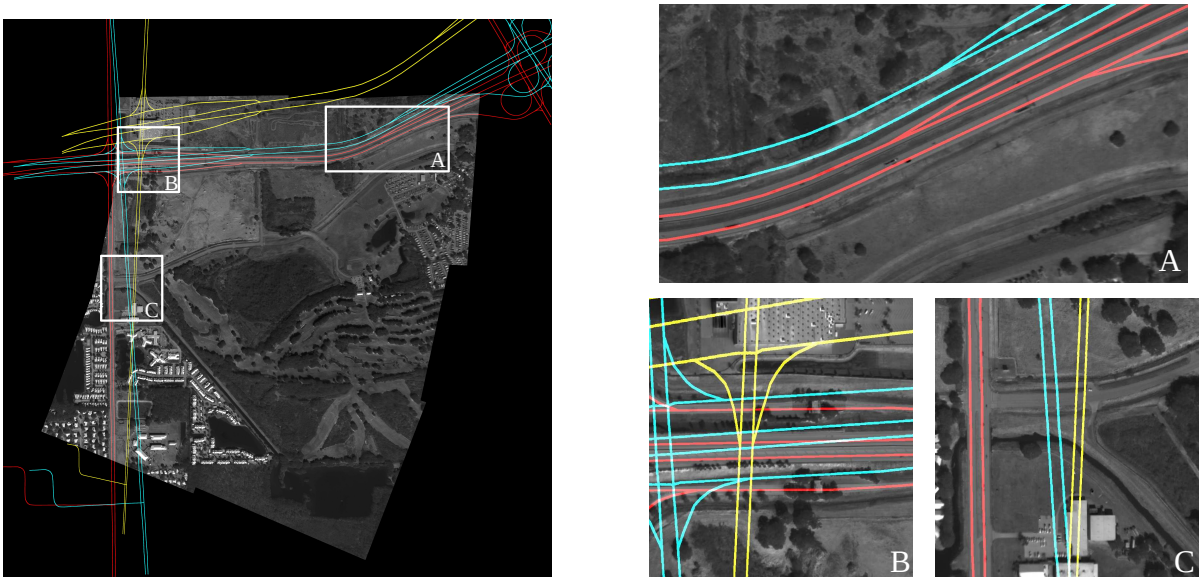
Dr. Sharma's research interests include image processing and computer vision, bioinformatics, color science and imaging, media security, and distributed signal processing. He is the editor of the "Color Imaging Handbook", published by CRC press in 2003. He is a fellow of the IEEE, of SPIE, and of the Society of Imaging Science and Technology (IS&T) and a member of Sigma Xi. He is a Technical Program Co-Chair for the 2016 IEEE International Conference on Image Processing (ICIP) and has served as a Technical Program Chair for ICIP 2012, as the Symposium Chair for the 2013 SPIE/IS&T Electronic Imaging symposium, as the 2010-2011 Chair IEEE Signal Processing Society's Image Video and Multi-dimensional Signal Processing (IVMSP) technical committee, the 2007 chair for the Rochester section of the IEEE and the 2003 chair for the Rochester chapter of the IEEE Signal Processing Society. From 2011 through 2015, he served as the Editor-in-Chief for the Journal of Electronic Imaging and in the past has served as an associate editor for the Journal of Electronic Imaging, IEEE Transactions on Image Processing, and IEEE Transactions on Information Forensics and Security.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

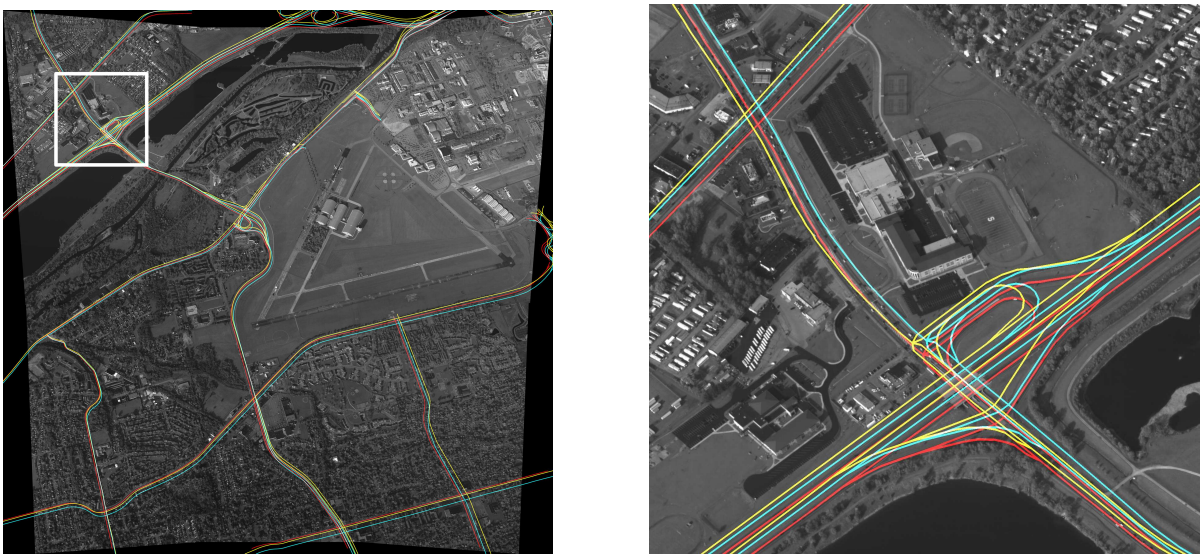
This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2016.2601265



(a)



(b)



(c)

Fig. 7: Road network alignment results using different methods for: (a) CORVUS(V) Area 1, (b) CORVUS(IR) Area 6, (c) WPAFB Area 9. The initial road network obtained from the WAMI frame meta-data is shown in cyan, while the result of the SBA method, and our proposed method appear in yellow and red colors, respectively. Left column is the full WAMI frame, while the right column shows a smaller cropped region that is marked on the corresponding full frame by a white rectangle.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.