Vehicle Geolocalization based on video synchronization

Ferran Diego, Daniel Ponsa, Joan Serrat and Antonio M. López

Abstract— **This paper proposes a novel method for estimating the geospatial localization of a vehicle. I uses as input a georeferenced video sequence recorded by a forward–facing camera attached to the windscreen. The core of the proposed method is an on–line video synchronization which finds out the corresponding frame in the georeferenced video sequence to the one recorded at each time by the camera on a second drive through the same track. Once found the corresponding frame in the georeferenced video sequence, we transfer its geospatial information of this frame. The key advantages of this method are: 1) the increase of the update rate and the geospatial accuracy with regard to a standard low–cost GPS and 2) the ability to localize a vehicle even when a GPS is not available or is not reliable enough, like in certain urban areas. Experimental results for an urban environments are presented, showing an average of relative accuracy of 1.5 meters.**

I. INTRODUCTION

In the recent decade the most employed sensor for consumer vehicle navigation and localization is the GPS receiver. At present, the standalone information of GPS has an approximate accuracy of 5–10 meters [1]. However, it can degrade specially on urban environments due to satellites occlusion and multi–path reception provoked by tall buildings, tunnels, etc. In this paper, we focus on precisely localizing a moving vehicle based just on a visual input where the GPS is not available or is not reliable enough. To achieve it, we assume the path of the vehicle is planned and known previously.

In the literature, a variety of methods have been proposed for computing the geospatial location of a vehicle or a robot. They can be decomposed on dead reckoning and map–matching algorithms, specifically using visual inputs. The former ones use an on–board inertial measurement unit to measure the vehicle travel distance and a gyro and compass to provide the moving direction in order to refine the geospatial location obtained from a GPS receiver. The latter corrects the vehicle position based on recovering its pose with respect to an environment model and differ on the way they construct the environment model. An approach for simultaneous localization and mapping (SLAM) is proposed in [2], which is based on the extended Kalman filter, and assumes that a robot moves in a stationary world of landmarks in order to estimate the environment map on–line. Some works [3], [4], [5], [6], [7] built a topological world representation estimated on–line by adding images to a database and maintaining a link graph. The global location is done by efficient image matching scheme w.r.t. all the image in the topological map. In particular, Schleicher *et al.* [7] combines the visual–information provided by a stereo camera and a GPS receiver in order to localize a vehicle and estimate a large–scale environment map. On the other hand, some works [8], [9] employ a $3D$ reconstruction of the environment built during an off–line learning phase and recover the geospatial location by matching the current view with projectives of the learned environment map. The work done by Hakeen *et al.* [10] consists in localizing the geospatial information and estimating the trajectory of a moving camera without a $3D$ reconstruction of the environment using a set of reference images with known GPS which were previously acquired. However, the algorithms based on mapping only estimate the geospatial location according to an efficient image matching scheme without considering the temporal coherence, that is, a vehicle follows a planned continuous and smooth planar trajectory.

In this paper, we describe an new approach in that we model a planned route using visual data. The visual data is acquired using a forward–facing camera attached at the windscreen. The key idea is to first record a video sequence along some road track and at each frame record its geospatial information with a better accuracy than that provided by consumer GPS receivers, by using either a differential GPS (DGPS) or an RTK–GPS. On a second round, when the vehicle drives later along this track, we record a second video, which we call the 'observed' sequence, without any GPS receiver. Note that, of course, the vehicle speed varies and this variation is independent in the two videos so that at one same location the speed is different, in general. For each current frame of the observed sequence, we will be able to find out the corresponding frame in the reference sequence, that is, the one recorded at the same or the closest camera location in the first ride. In other words, we synchronize the two video sequences on–line, whilst recording the second one. Once found the corresponding frame, we transfer the geospatial information of the corresponding frame in the first ride to the current frame.

II. SYSTEM OVERVIEW

The aim of vehicle geolocalization is to compute the position of a vehicle using a GPS receiver. However, we focus on computing the geospatial localization of a vehicle replacing the GPS receiver with a forward facing camera and

This work was supported by the Spanish Ministry of Education and Science (MEC) under project TRA2007-62526/AUT, the research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and the FPU MEC grant AP2007-01558. We gratefully acknowledge to Dr. Joan Nunes from the Laboratorio de Información Geográfica y de Teledetección (LIGIT) of UAB and Ernest Bosch from the Institut Cartogràfic de Catalunya (ICC) for the helpful discussions and providing access to DGPS and software.

The authors are with Computer Vision Center & Computer Science Dept., Edifici O, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallés, Spain. ferran.diego@cvc.uab.es

a video sequence with known GPS. The video sequence is recorded from moving vehicles whose frames are annotated with geospatial information. This video sequence are called reference sequence. Note that we assume that the vehicles follow similar trajectories on the same track. This assumption is plausible due to transportation vehicles normally follow planned routes.

The overall vehicle geolocalization method consists of two stages, which are shown in Fig. 1. In the first stage, an image descriptor $\mathbf{a}_t$ and $\mathbf{a}_*$ are computed for each frame t in the observed video sequence and all video frames in the georeferenced sequence, respectively (see Sect. II-A). The image descriptor $\mathbf{a_t}$ is a simply representation of the image acquired at time t used to compare it among all images descriptors of the reference sequence, $\mathbf{a}_*$. The image descriptor is robust to different illumination conditions which allow us to handle the comparison among frames acquired at different times. In the second stage, a video synchronization is proposed to estimate the corresponding frame in the georeferenced sequence to the most recently acquired frame in the observed video, in order to transfer its geospatial information to the current frame(see Sect. II-B). The temporal coherence of the video synchronization algorithm consists in relating the frames of the observed sequence with regard to the frames of georeferenced sequences which maximizes jointly a 'similar' content. These stages are repeated while the observed sequence is being acquired in order to estimate the GPS locations of the vehicle. The following sections detail these two steps.

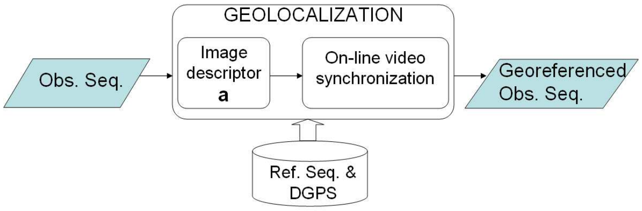


Fig. 1. Illustration of the overall framework of our vehicle geolocalization decomposed into two stages: computation of the image descriptor $\mathbf{a}$ and on–line video synchronization, and the database.

A. Image descriptor

Let $\mathbf{F}^o_t$ and $\mathbf{F}^r_{x_t}$ denote the t^{th} frame and the x^{th}_t frame of the observed and a reference sequence, respectively. The image descriptors $\mathbf{a}_t$ and $\mathbf{a}_{x_t}$ describe the images $\mathbf{F}^o_t$ and $\mathbf{F}^r_{x_t}$, respectively. These image descriptors are compared in the video synchronization stage in order to measure the degree of similarity between the two frames they represent, in order to later estimate the likelihood that t and x_t are corresponding frames. The image descriptor $\mathbf{a}_*$ is computed as follows. First, the image is smoothed using a Gaussian kernel and downsampled at $1/16^{th}$ of the original resolution. Then, partial derivatives $(\frac{\partial \cdot}{\partial x}, \frac{\partial \cdot}{\partial y})$ are computed and the value at each pixel is set to zero if the gradient magnitude is less than 5% of the maximum. Finally, the partial derivatives are stacked all into a column vector $\mathbf{a}_*$ which is normalized to unit norm. This image descriptor is adopted because it is simple to compute and compare in order to evaluate instantaneously all the similarities among a subset of image descriptors in the reference sequence. At the same time, this image descriptor deals with contrast or lighting changes and, of course, when they show different foreground objects like vehicles. Fig. 2 summarizes the computation of the image descriptors $\mathbf{a}_t$ and $\mathbf{a}_{x_t}$ with regard to the frames $\mathbf{F}^o_t$ and $\mathbf{F}^r_{x_t}$.

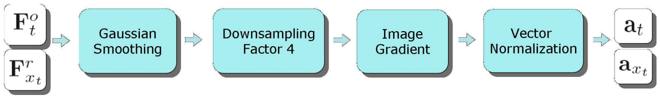


Fig. 2. Illustration of the computation flow of an image descriptor decomposed as follows: (1) smooth the input image, (2) downsample at the $(1/2^4)^{th}$ of the original resolution, (3) compute partial derivatives and finally, (4) stack the partial derivatives into a column vector normalized to unit norm.

B. video synchronization

The aim of video synchronization is the alignment along the time dimension of two video sequences recorded at different times. Video synchronization estimates a discrete mapping $c(t) = x_t$ at time $t = 1, \ldots, n_o$ of the observed sequence such that $\mathbf{F}^r_{x_t}$ maximizes some measure if similarity with $\mathbf{F}^o_t$, among a subset of frames of $\mathbf{F}^r$, being n_o the number of frames of the observed sequence and $\mathbf{F}^r_{x_t}$ and $\mathbf{F}^o_t$ the x^{th}_t and t^{th} frame in the reference and observed sequence respectively. Once the discrete mapping is found, the geospatial information of the corresponding frame $\mathbf{F}^r_{x_t}$ at time t is transferred to the current input frame $\mathbf{F}^o_t$. These video sequences are recorded by a pair of independently moving cameras, although their motion is not completely free because we impose the vehicles to follow *approximately* coincident trajectories. As consequence, the video frames have a large overlapping in the field of view of the two cameras. The video synchronization is a challenging task because it must face (1) varying and independent speed of the cameras in the two sequences which implies a non–linear time correspondences and (2) slight rotations and translations of the camera location due to dissimilar trajectories. Although several video synchronization techniques have been proposed [11], [12], [13], only our previous work [14] on video alignment addresses these two specific requirements. Now, we need to add a third important requirement: the temporal correspondence between the observed and the reference sequence must be computed on–line, because we need to obtain the geospatial location after each frame has been acquired. Therefore, we propose a on–line video synchronization algorithm by extending [14]. That is because the video synchronization jointly compares the similarity content among consecutive frames in order to exploit that the vehicles follow similar trajectories whereas image retrieval techniques retrieves a frame with the highest similarity being a challenging task to distinguish the corresponding frame among a huge amount of frames which show similar content.

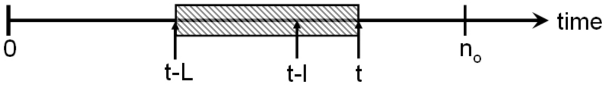


Fig. 3. Temporal meaning of a fixed lag–smoothing of a hidden Markov model where the label x_{t-l} is estimated at time t using L images in the observed sequence, which are from $(t-L)^{th}$ to t^{th} frame in the observed sequence.

We state the problem of estimating the corresponding frame as one of probabilistic inference. A label $x_t \in 1, \ldots, n_r$ is the number of corresponding frame in the reference sequence to the t^{th} frame in the observed sequence, being n_r the number of frames in the reference sequence. The estimation of the label x_t is posed as a maximum a posteriori inference problem of a fixed–lag smoothing dynamic Bayesian network (DBN) which is defined as

$$x_{t-l}^{MAP} = \operatorname*{argmax}_{x_{t-l} \in \Omega_t} p(x_{t-l} | \mathbf{y}_{t-L:t}) \ , \tag{1}$$

where Ω_t is the set of labellings allowed to infer the label x_{t-l}, $l \geq 0$ is a lag or delay, $L \geq l$ is the total set of observed frames used to infer the label x_{t-l} and $\mathbf{y}_{t-L:t}$ are the observations (image descriptors described in Sect. II-A) from the $(t-L)^{th}$ to the t^{th} in the observed sequence. The range of the set of labellings Ω_t is set as $[x_{t-L-1}, x_{t-l-1} + \Delta]$, being Δ the maximum label difference between consecutive frames. Note that x_{t-L-1} and x_{t-l-1} have been estimated before the set of labellings Ω_t is defined at time t. The estimation of x_{t-l}^{MAP} requires $L+1$ frames of the observed sequence and the fixed–lag smoothing infers the label $t-l$ at time t with a delay of l frames. Fig. 3 illustrates the meaning of a fixed–lag smoothing. In order to estimate the label x_{t-l}, max–product inference algorithm is applied in Eq. (1) as

$$x_{t-l}^{MAP} = \operatorname*{argmax}_{x_{t-l} \in \Omega_t} \max_{\mathbf{x}_{t-L:t} \setminus x_{t-l}} p(\mathbf{x}_{t-L:t} | \mathbf{y}_{t-L:t}) \ , \tag{2}$$

where $\mathbf{x}_{t-L:t} = [x_{t-L}, \ldots, x_t]$ is a list of labels which corresponds the temporal correspondence among the reference sequence and the observations $\mathbf{y}_{t-L:t}$. The maximization of the posterior probability density $p(\mathbf{x}_{t-L:t} | \mathbf{y}_{t-L:t})$ is over all the temporal correspondence labels $\mathbf{x}_{t-L:t}$ expect for x_{t-l}. The posterior probability density of the temporal correspondence $\mathbf{x}_{t-L:t}$ is decomposed as

$$p(\mathbf{x}_{t-L:t} | \mathbf{y}_{t-L:t}) \propto p(\mathbf{y}_{t-L:t} | \mathbf{x}_{t-L:t}) p(\mathbf{x}_{t-L:t}) \ , \tag{3}$$

where $p(\mathbf{x}_{t-L:t})$ and $p(\mathbf{y}_{t-L:t} | \mathbf{x}_{t-L:t})$ are *a prior* and an *observation likelihood* respectively. The estimation of label x_{t-l} is the argument that maximizes the temporal coherence between two video sequences summarized as

$$x_{t-l}^{MAP} = \operatorname*{argmax}_{x_{t-l} \in \Omega_t} \max_{\mathbf{x}_{t-L:t} \setminus x_{t-l}} p(\mathbf{y}_{t-L:t} | \mathbf{x}_{t-L:t}) p(\mathbf{x}_{t-L:t}) \ . \tag{4}$$

The prior $p(\mathbf{x}_{t-L:t})$ constraints the temporal correspondence between two video sequences depending on the assumption adopted between these two sequences. For simplicity, the prior probability is assumed to be independent given the label values. Hence, it is written as

$$p(\mathbf{x}_{t-L:t}) \ = \ P(x_{t-L}) \prod_{k=t-L}^{t-1} p(x_{k+1} \mid x_k) \ , \tag{5}$$

where $P(x_{t-L})$ is the probability for the first label of the current estimation of the temporal correspondence $\mathbf{x}_{t-L:t}$ that gives the same probability to all labels inside Ω_t. The intended meaning of $p(x_{k+1} \mid x_k)$ is the following: we assume that vehicles do not go backward, that they move always forward or at most stop for some time. Therefore, the labels x_t must increase monotonically. Hence, $p(x_{k+1} \mid x_k)$ is defined as

$$p(x_{k+1} \mid x_k) \ = \ \begin{cases} \beta & \text{if } x_{k+1} \geq x_k \\ 0 & \text{otherwise,} \end{cases} \ , \tag{6}$$

where β is a constant that gives equal probability to any label greater than x_k.

On the other hand, the observation likelihood $p(\mathbf{y}_{t-L:t} | \mathbf{x}_{t-L:t})$ describes the similarity of two video sequences given a temporal correspondence $\mathbf{x}_{t-L:t}$. For simplicity, we also assume that the likelihood of observations $\mathbf{y}_{t-L:t}$ is independent given their corresponding label values and hence $p(\mathbf{y}_{t-L:t} | \mathbf{x}_{t-L:t})$ factorizes as

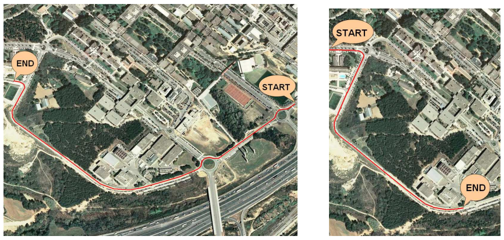
$$p(\mathbf{y}_{t-L:t} | \mathbf{x}_{t-L:t}) \ = \ \prod_{k=t-L}^{t} p(y_k | x_k) \ , \tag{7}$$

where $p(y_k | x_k)$ describes the similarity between two frames, one frame from the reference sequence and another from the observed sequence. We want the similarity to be maximum or at least high, if two frames are corresponding. The observations $\mathbf{y}_{t-L:t}$ corresponds to the image descriptors $[\mathbf{a}_{t-L}, \ldots, \mathbf{a}_t]$, which are described in Sect. II-A. In order to measure the similarity among the image descriptor of the current frame $\mathbf{a}_t$ and all image descriptors of the frames inside Ω_t, we consider the inner product of two image descriptor because it measures the coincidence of the gradient orientation in the subsampled image. Hence, our observation probability is defined as

$$P(y_k | x_k) = \max_{\substack{-\Delta_x < i < \Delta_x \\ -\Delta_y < j < \Delta_y}} \Phi(< \mathbf{a}_{x_k}^{i,j}, \mathbf{a}_k >; 1, \sigma_s^2) \ , \tag{8}$$

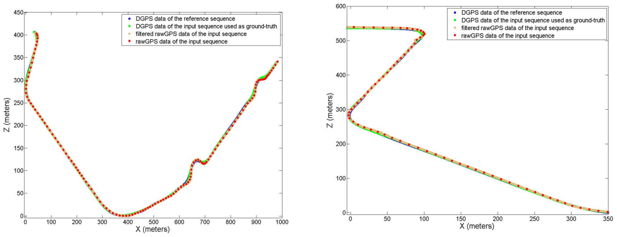
where $\Phi(\beta; \mu, \sigma^2)$ denotes the evaluation of the Gaussian pdf $\mathcal{N}(\mu, \sigma^2)$ at β, and σ_s^2 controls the likelihood of the similarity measure between two image descriptors. We set σ_s^2

to 0.5 to give significant likelihood only those frames whose image descriptors form an angle less than approximately $5°$. The upper indexes of $\mathbf{a}_{x_t}^{i,j}$ mean that the image descriptor is computed from a smoothed low resolution image with a translation of i and j pixels over the x– and y–directions. The maximum translation over x– and y–direction are Δ_x and Δ_y respectively and, they are set to 2 pixels. These translations increase the robustness of the likelihood when slight camera rotations and translations appear due to trajectory dissimilarities.



(a) (b)

Fig. 4. Aerial view of the paths covered by the vehicle where the travel distance of (a) and (b) are 1.5 and 1 km, respectively.



(a) (b)

Fig. 5. Geospatial information of the vehicle paths shown on Fig. 4 using a DGPS in the first and the second ride, and the raw GPS and the filtered raw GPS of a standard low–cost GPS receiver which is synchronized with the DGPS.

III. RESULTS

In this section we present quantitative results of our vehicle geolocalization system. In order to test its behaviour two video sequences have been recorded with a SONY DCR-PC330E camcorder in our university campus. The camera is attached to the windscreen and forward–facing. It captures 720×560 pixel frames at a frame rate of 25 Hz. In order to evaluate the accuracy of our approach using the Euclidean distance between geospatial locations, the reference and the observed are annotated with geospatial information obtained from a DGPS Trimble-GeoXT, with an estimation accuracy of 0.5 meters after post–processing and an update rate of 1 Hz. The DGPS data of the reference sequence is used to transfer it to the observed sequence whereas the DGPS data of the observed sequence is used as a ground–truth. Furthermore, we will compare our method with regard to a the KEOMO 16 channel GPS receiver with Nemeric chipset with an update rate of 1 Hz, which is denoted as a *standard low–cost GPS*. These three sensors are synchronized to acquire their data at the same time. Note that the GPS data is available only in 4% of the frames of a video sequence. However, for the rest of frames there is still some knowledge that can be exploited, since a vehicle follows a regular trajectory as shown on Fig. 4. In order to estimate the geoespatial information of both GPS receivers for each video frame, we apply a Kalman smoother to process the available GPS data and interpolate the lacking information by means of the Rauch–Tung–Striebel Kalman smoother equations [15]. Hence, all frames of the reference sequence are georeferenced using the DGPS data. The observed sequence is georeferenced using a standard low–cost GPS and a DGPS data. The latter is used as a ground–truth. The GPS information of both GPS receivers is available only the 4% of frames in a video sequence is called *raw GPS* whereas the GPS data estimated by a Kalman smoother is called *filter raw GPS*.

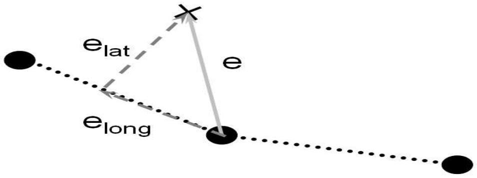
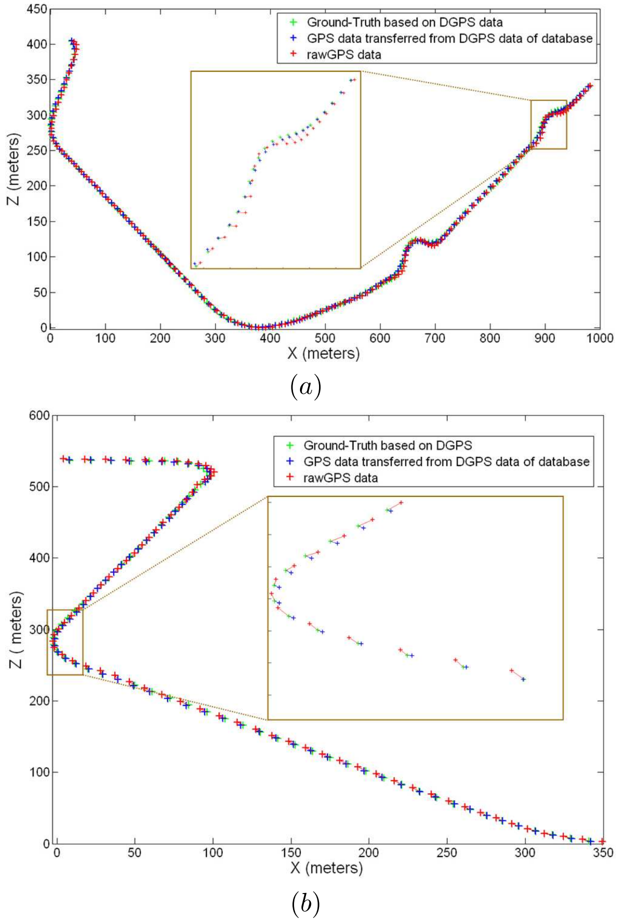


Fig. 6. The error is the Euclidean distance($e = \sqrt{X^2 + Z^2}$) relative to the GPS ground–truth and their respective lateral and longitudinal component with regard to the trajectory of the vehicle using a linear interpolation of GPS data used as a ground–truth. The dot points are the geospatial location of the ground–truth whereas the cross point is the computed or captured geospatial location.

The paths covered by the vehicle are shown in Fig. 4 whereas the geospatial information on both video sequences is shown in Fig. 5. The average speed of the vehicle is approximately 50 kph on both video sequences and both paths. The length of each planned route shown in Fig. 5 is approximately 1.5 and 1 km respectively. The reference and the observed video sequence shown in Fig. 4a are 3200 and 3500 frames long respectively whereas, the reference and the observed video sequence shown in Fig. 4b are 1800 and 2100 frames long, respectively. The difference in the number of frames is due to varying and independent speed of the vehicle during each track. In addition, these sequences were recorded at noon with the presence of vehicles and some tall–building. Fig. 6 shows the error metric used to compare our approach with regard to a standard low–cost GPS. The error metrics are the Euclidean distance of the estimated geospatial location w.r.t. the geospatial location of the ground–truth and the projection of the Euclidian distance w.r.t. the trajectory of vehicle. This projection decomposes the error in longitudinal and lateral error component. The error is calculated at each frame of the observed sequence in order to illustrate its temporal evolution whereas its histogram summarizes it compactly. Furthermore, we compare our approach w.r.t. a standard low–cost GPS calculating the error and the histogram using the raw and filtered raw GPS data of the GPS receiver and the transferred geospatial



(a)

(b)

Fig. 7. Geospatial information of a vehicle using the transferred GPS and an off-the-shelf low–cost GPS relative, to the ground–truth. The links relates the corresponding geospatial informations among different geospatial information. The cross points are the geospatial location whereas the lines are their correspondence among geospatial locations.

information from the georeferenced reference sequence. Fig. 8 shows the errors and the histograms of the three geospatial information relative to the ground-truth, which are the transferred geospatial information, raw and filtered raw GPS data, in the two paths. Our approach obtains an average Euclidean error of 1.5 meters approximately whereas the standard low cost GPS, raw GPS and filtered raw GPS, obtains an error of 6 meters approximately in both GPS datas. Fig. 8 shows that the main error of our system is the longitudinal component w.r.t. the trajectory of the vehicle because to distinguish among consecutive frames is a challenging task due to they show similar image content. Although, the longitudinal error of our approach is approximately 4 meters lower than a standard low–cost GPS. Fig. 7 depicts the estimation of our vehicle geolocalization system w.r.t. the standard low–cost GPS and the ground–truth showing their correspondences. In spite of the large error on some short segments due to dissimilar vehicle trajectories and video synchronization errors, as shown in Fig. 8, we are still able to achieve an accuracy of less than 2 meters in approximately the 80% of the frames. Furthermore, the update rate of our approach is 25 Hz, which is 25 times more than a standard low–cost GPS, and our approach is able to locate the vehicle where GPS data is not available.

IV. Conclusion

In this paper, we have proposed a new method for estimating the geospatial localization of a vehicle. We synchronized a captured sequence of images with regard to a video sequence with known GPS to transfer the geospatial location of a vehicle at each time a frame is acquired. The novelty of our approach is the geospatial information is computed without a GPS receiver but using an on–line video synchronization between two video sequences, where one of these video sequences has been georeferenced using DGPS or inertial GPS. In addition, we described a qualified method to estimate the geospatial information for all frames of a video sequence instead of the 4% of the frames. The advantages of our method are: (1) the increase of temporal resolution (update rate) which is 25 times faster than a standard GPS, (2) the increase of the relative geospatial accuracy with less than 2 meters of error more than 80% of the time instead of an average accuracy of 5 meters obtained by a standard GPS and finally, (3) this method is still able to estimate the geospatial location of a vehicle where the GPS is not available or is not reliable enough like in urban areas due to multipath reception and satellite occlusions. As future work, we plan to include a consumer GPS receiver as an additional observation in our on–line video synchronization in order to increase the accuracy.

References

[1] I. Skog and P. Händel, “In-car positioning and navigation technologies: a survey,” *Trans. Intell. Transport. Sys.*, vol. 10, no. 1, pp. 4–21, 2009.

[2] M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-whyte, and M. Csorba, “A solution to the simultaneous localization and map building (slam) problem,” *IEEE Transactions on Robotics and Automation*, vol. 17, pp. 229–241, 2001.

[3] K. Konolige, J. Bowman, J. D. Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua, “View-based maps,” in *International Journal of Robotics Research (IJJR)*, Accepted for publication.

[4] I. Mahon, S. Williams, O. Pizarro, and M. Johnson-Roberson, “Efficient view-based slam using visual loop closures,” *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1002–1014, 2008.

[5] J. Courbon, Y. Mezouar, and P. Martinet, “Autonomous navigation of vehicles from a visual memory using a generic camera model,” *Trans. Intell. Transport. Sys.*, vol. 10, no. 3, pp. 392–402, 2009.

[6] K. Konolige and M. Agrawal, “Frame-frame matching for realtime consistent visual mapping,” in *in Proc. International Conference on Robotics and Automation (ICRA*, 2007.

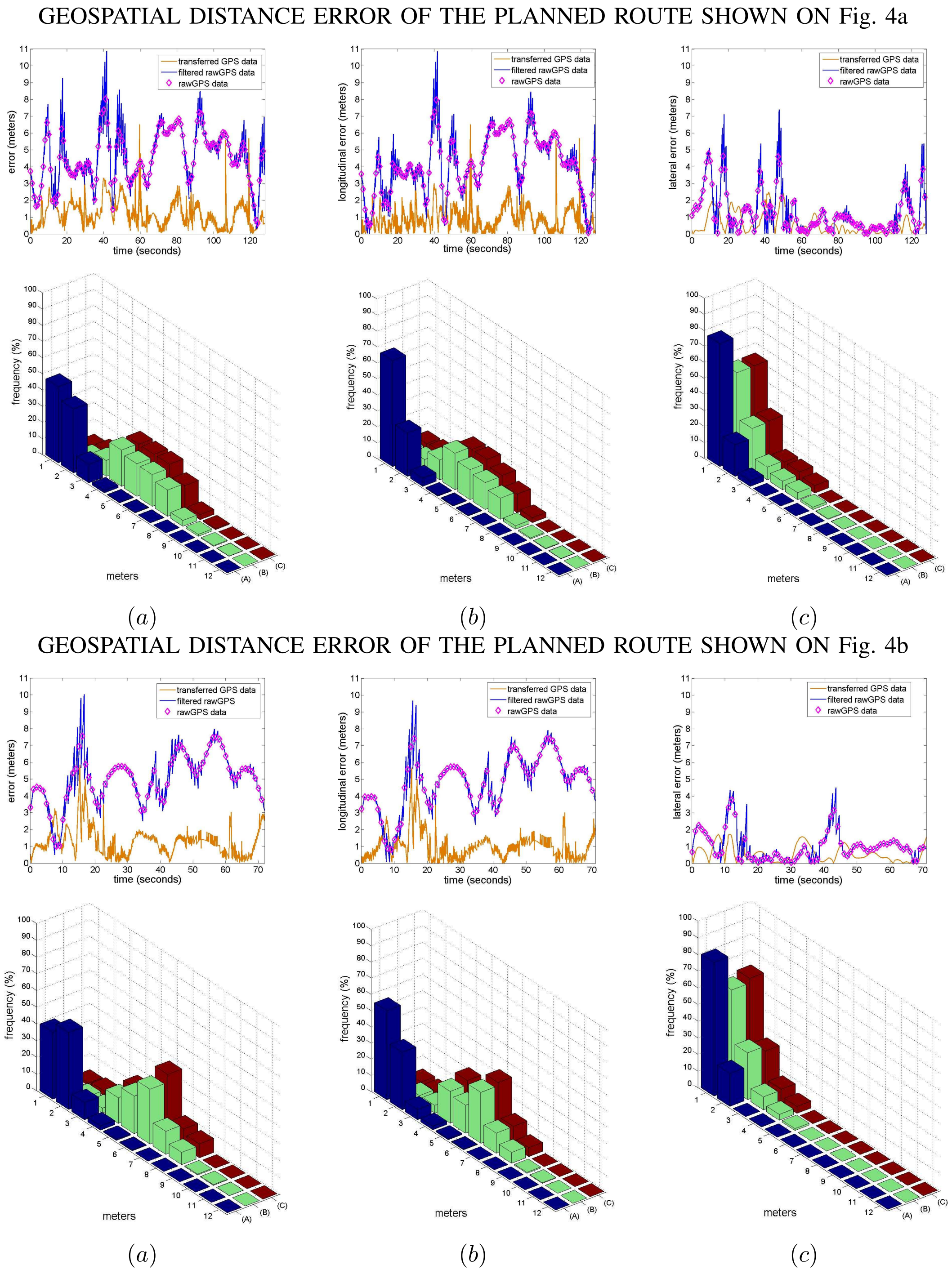
[7] D. Schleicher, L. M. Bergasa, M. Ocana, R. Barea, and M. E. Lopez, “Real-time hierarchical outdoor slam based on stereovision and gps fusion,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 10, no. 3, pp. 440–452, July 2009.

[8] A. Levin and R. Szeliski, “Visual odometry and map correlation,” *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 1, pp. 611–618, 2004.

[9] K. K. Jun, J. Miura, and Y. Shirai, “Autonomous visual navigation of a mobile robot using a human-guided experience,” *Robotics and Autonomous Systems*, vol. 40, pp. 2–3, 2000.

[10] A. Hakeem, R. Vezzani, M. Shah, and R. Cucchiara, “Estimating geospatial trajectory of a moving camera,” *Pattern Recognition, International Conference on*, vol. 2, pp. 82–87, 2006.

[11] L. Wolf and A. Zomet, “Wide baseline matching between unsynchronized video sequences,” *Int. Journal of Computer Vision*, vol. 68, no. 1, pp. 43–52, 2006.



GEOSPATIAL DISTANCE ERROR OF THE PLANNED ROUTE SHOWN ON Fig. 4a

(a) (b) (c)

GEOSPATIAL DISTANCE ERROR OF THE PLANNED ROUTE SHOWN ON Fig. 4b

(a) (b) (c)

Fig. 8. The error and its distribution of (a) the Euclidean distance between the estimated geospatial and the ground–truth whereas (b) and (c) are the error of the longitudinal and lateral component of the Euclidean distance relative the trajectory of the ground–truth. The error is calculated using three estimated geospatial locations: transferred location from a georeferenced sequence and the raw and filtered raw GPS of a low–cost GPS receiver. The first and third row are the temporal evolution of error computed at each frame in the observed sequence whereas the second and forth row are its distribution, being (A) the error of the transferred geospatial information, (B) the filtered raw GPS of a standard GPS and (C) only the raw GPS of a standard GPS. In addition, the numbers of x–direction of the distribution represents the rounded error expressed in meters.

[12] Y. Caspi and M. Irani, "Spatio–temporal alignment of sequences," *IEEE Trans. Pattern Analisys and Machine Intelligence*, vol. 24, no. 11, pp. 1409–1424, 2002.

[13] C. Lei and Y. Yang, "Trifocal tensor–based multiple video synchronization with subframe optimization," *IEEE Trans. Image Processing*, vol. 15, no. 9, pp. 2473–2480, 2006.

[14] F. Diego, D. Ponsa, J. Serrat, and A. López, "Video alignment for difference-spotting," in *ECCV 2008 Workshop on Multi Camera and Multi-modal Sensor Fusion Alg. and Apps.*, 2008. [Online]. Available: http://perception.inrialpes.fr/Publications/2008/ZCIBH08

[15] A. Gelb, Ed., *Applied Optimal Estimation*. Cambridge, MA: MIT Press, 1974.