Global Robot Ego-localization Combining Image Retrieval and HMM-based Filtering

Cédric Le Barz[1], Nicolas Thome[2], Matthieu Cord[2], Stéphane Herbin[3] and Martial Sanfourche[3]

***Abstract*— This paper addresses the problem of global visual ego-localization of a robot equipped with a monocular camera that has to navigate autonomously in an urban environment. The robot has access to a database of geo-referenced images of its environment and to the outputs of an odometric system (Inertial Measurement Unit or visual odometry). We suppose that no GPS information is available. The goal of the approach described and evaluated in this paper is to exploit a Hidden Markov Model (HMM) to combine the localization estimates provided by the odometric system and the visual similarities between acquired images and the geo-localized image database. It is shown that the use of spatial and temporal constraints reduces the mean localization error from 16 m to 4 m over a 11 km path evaluated on the Google Pittsburgh dataset when compared to an image based method alone.**

I. Introduction

The problem tackled in this paper is the visual autonomous navigation of a robot operating in an urban environment [1]. A typical target application could be the delivery of goods using unmanned ground or aerial vehicles where the robot trajectory has been defined before hand on a given map, and must be followed to reach its final destination (Fig. 1). Absolute localization system like GPS may be shadowed or completely unavailable in several areas of the trajectory and substitute localization means must be used.

Visual information is an appealing alternative because cameras and densely sampled geo-referenced images are now commonly available. Nevertheless, the localization of a robot exploiting only image content is challenging because two images of the same place acquired at different times and with different cameras may show huge appearance differences due to illumination and colorimetry variations (e.g. sunny or cloudy days), camera viewpoints changes, scene modifications (e.g. seasonal changes, building construction) and occlusions (e.g. by cars) (Fig. 2). Standard image retrieval (IR) methods such as k Nearest Neighbour (kNN) votes or Bag Of Visual Words (BoVW) [2] produce noisy results that necessitate filtering to be robustly exploited as primary global localization information source.

Odometric systems, IMU based or visually based, provide localization information at low cost: however this information is only relative to a given position and suffers from drift especially on complex trajectories. It can only be used reliably on small portions of the followed route and can't be the only source of measurement for absolute localization.

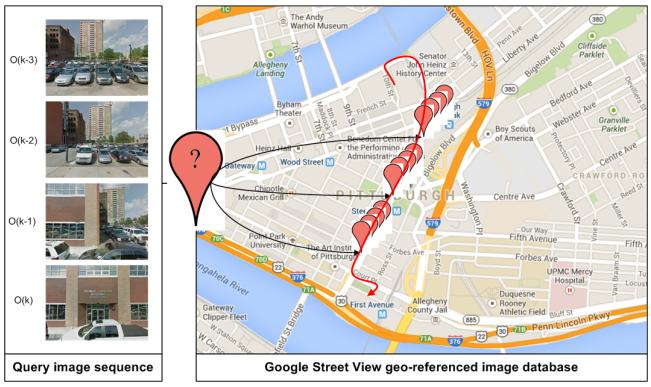

Fig. 1. Visual ego-localization system: Our system aims at matching a sequence of images with geo-referenced database images in order to determine accurate geo-localization from noisy odometric information.

The main contribution of this paper is to describe a general framework enabling to combine these two sources of noisy localization information: local odometry and visual similarity. More precisely, the solution we propose uses an IR algorithm applied to a database of geo-referenced images integrated into a Hidden Markov Model (HMM) accounting for odometry uncertainty. The role of the HMM is to exploit spatio-temporal constraints in order to filter out erroneous IR results.

The effectiveness of our approach has been evaluated over a 11 km path using two kinds of images: Google Streetview images [4] simulating images acquired online by the robot camera and Google Pittsburgh image dataset [3] as geo-referenced image database.

II. Related work

Visual place recognition problems have been addressed recently thanks to the availability of image databases. Most of them rely on the extraction of 2D and/or 3D features, that are compared to a geo-referenced feature database. Unlike [5] [6] that are loop closure algorithms developed for Simultaneous Localization and Mapping (SLAM) systems, we focus on position tracking.

Zamir et al. propose in [7] a hierarchical method to localize a group of images. SIFT descriptors from database images are indexed using a tree. A nearest neighbour tree search is then done for each SIFT query image feature. Weak votes are removed and each reliable feature votes for

[1]Cédric Le Barz is with Theresis department, THALES company, 91767 Palaiseau, France cedric.lebarz@thalesgroup.fr

[2]Nicolas Thome and Matthieu Cord are Sorbonne University, UPMC University, Paris 06, UMR 7606, LIP6, 75005 Paris, France nicolas.thome@lip6.fr, matthieu.cord@lip6.fr

[3]Stéphane Herbin and Martial Sanfourche are with the French Aerospace Lab, ONERA, 91123 Palaiseau, France stephane.herbin@onera.fr, martial.sanfourche@onera.fr

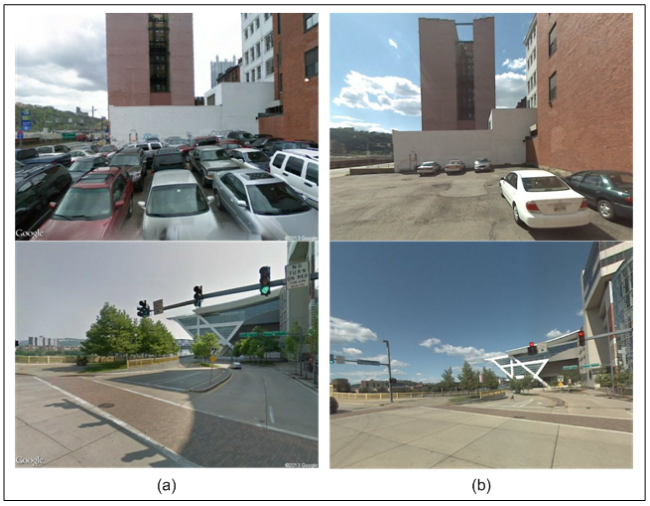

Fig. 2. Google Streetview images (a) and Robot images (b). Note the impact of different focal lenses, weather conditions, viewpoint changes and the presence/absence of cars in the scene.

a location. All accumulated spatial votes are then filtered by a Gaussian kernel. The geo-referenced image with the highest number of votes determines the location. In [8], the method described in [7] is improved by interpreting the 2D map votes as a likelihood. This likelihood is then used in a Bayesian tracking filter to estimate the temporal evolution based on the previous state. Both solutions are dedicated to web video annotation, and localization is not realized on the fly which makes it useless for navigation.

In [9], the vehicle localization algorithm uses simple visual features and 3D features. The solution requires in a preliminary phase to build a compact map described as a graph. Nodes include vehicle position at fixed distance interval and visual and 3D features. At runtime, a Bayesian filter is used to estimate the probability of the vehicle position by matching features extracted from sensors with database features. Their solution uses two lateral cameras and two lateral LIDARs. Same sensors are used during the map building step and the localization step. In contrast, our solution is monocular and uses different cameras for acquisition and reference database.

The solution proposed in [10] is based on the match of visual odometric measurements with a 2D road-map. The map is represented by a directed graph and a probabilistic approach is defined in order to navigate within this graph. They are able to localize themselves after a few seconds of driving with an accuracy of 3 m on a 18 km^2 map containing 2150 km of roads. Our navigation solution does not use any 2D road-map. It uses only visual features from images along the specified trajectory combined with odometric information.

In [11], the localization is achieved by recognizing temporal coherent sequences of local best matches. These local best matches are based on a Sum of Absolute Difference (SAD) on resolution-reduced and patch-normalized images between last acquire image and M previous images. They make the assumption that the robot velocity is constant between all image sub-sequences. The proposed solution is robust to extreme perceptual changes, but sensitive to point of view.

In [12], authors work on visual similarity for UAV ego-localization. They propose to generate artificial views of the scene in order to overcome the large view-point differences. Nevertheless, spatio-temporal constraint is not taken into account.

Another type of approach is to cast the problem as a classification task, as in [13]. A classifier for each image in the database is trained using per-exemplar SVM approach. The main contribution of the mentioned paper is the calibration of all SVM classifiers using mainly negative examples in order to be able to compare all classifiers scores.

As in [7] [8] [9] [11], our solution uses spatio-temporal coherency. Along with this, our solution uses a HMM enabling to take into account in a more flexible way robot dynamics. No assumption is done concerning the constant velocity of the robot, but as in [14] we consider coarse position estimates provided by an odometric sensor and their uncertainties. Furthermore, in contrast to [7] [8] [9], the latest part of the trajectory is re-estimated for each new acquisition.

III. Proposed solution

Preliminary experiments made clear that IR approaches are not selective enough for urban areas because the same features tend to be shared by several neighbour images and produce erroneous matches (Fig. 4). That is why we propose to exploit the spatio-temporal coherency in order to filter out the wrong matches provided by standard IR algorithms. This is achieved by combining the similarities supplied by an IR algorithm with a HMM where hidden states represent places. The idea is to find the trajectory that best explains the M past observations and therefore the current position. The definition of a HMM for each new image acquired by the robot will enable to re-estimate the latest part of the trajectory so that past errors are corrected on a long term basis. Furthermore, taking into account odometric information reduces online the number of database images used in the IR task.

A. General principle

At each time t the robot acquires an image O_k and receives an estimate of its current position $\tilde{S}_k$ from the odometric system. The goal of the global localization algorithm is to produce a better estimate $\hat{S}_k$ of the current robot position from the past observations and odometric estimate (Fig. 3). The estimator is a function of the M past observations $\mathbf{O}_k = \{O_{k-M+1}, \dots O_k\}$ (i.e. the current location estimate exploits a set of observations in a sliding window based approach of length M) and the estimated position $\tilde{S}_k$.

Estimation is realized in a classical random variable setting where the robot location at time t is considered as a random variable q_t taking values in a discrete set of possible location $S_j \ j \in \{1 \dots N\}$. The main modeling hypothesis is that its random behaviour is represented by a HMM.

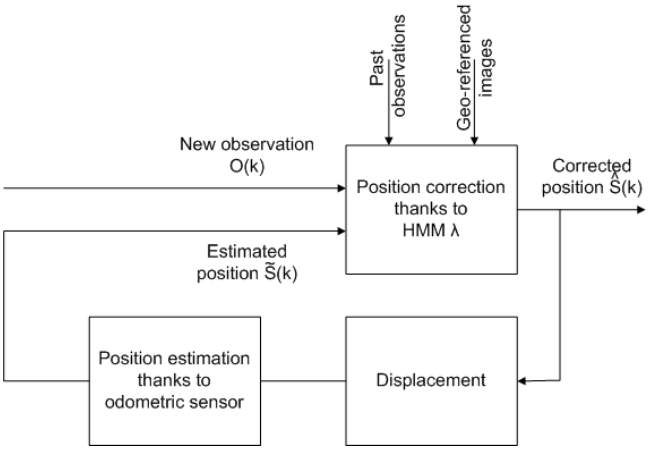

Fig. 3. System overview: for each new observation O_k, an odometric sensor provides a rough position estimate $\tilde{S}_k$ that is corrected thanks to a new HMM λ combining visual information and spatio-temporal constraints. This corrected position is noted $\hat{S}_k$.

Using the classical notations of [18], the use of a HMM requires the definition of the adequate model $\lambda = \{N, M, \Pi, A, B\}$ where N is the number of states, M is the number of observations, Π is the prior on the initial state, A is the transition probability matrix between the states and B is the observation probability matrix given some states.

The HMM approach provides a standard way to estimate the most likely state sequence $\hat{\mathbf{S}}_k$, i.e the M successive places, explaining the sequence of observations $\mathbf{O}_k = \{O_{k-M+1}, ..., O_{k-1}, O_k\}$ (Viterbi algorithm):

$$\hat{\mathbf{S}}_k = \arg\max_{\mathbf{S}} P(\mathbf{S}|\mathbf{O}_k, \lambda) \quad (1)$$

The question is now to design the HMM adapted to the global estimate of the robot location. This will be detailed in two steps: construction of the state transition matrix A and initial state vector, and computation of the conditional observation matrix B.

B. State transition matrix and initial state vector

The state transition matrix A and initial state vector are built from knowledge of the odometric system behaviour, robot kinematics and quality of the available database of geo-referenced images.

From the robot kinematics, images are approximately acquired every D meters with an odometric uncertainty of Δ meters. The image database consists of overlapping images acquired every $D^{'}$ meters with $D^{'} \leq D$. In this setting, the database is therefore assumed to have a bigger sampling rate than the online image acquisition rate.

Each possible state location S_j is uniquely defined by a geo-referenced database image I_j.

The filtering capacity of the HMM depends on the number M of past observations. This control parameter is free and its influence will be studied in the experiments.

One critical parameter is the localization uncertainty U which defines the area where the robot is supposed to be.

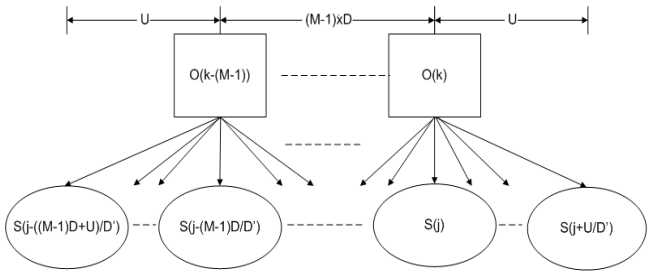

Fig. 5. Relation between states to consider and localization uncertainty.

This localization uncertainty can be for example the initial position uncertainty when the robot starts its mission.

The number of states N, i.e. the number of potentially matching images in the database, the initial state probability Π and the state transition probability matrix A depend on U, D, Δ and M. They are defined the following way:

- N : Given the putative position of the robot $\tilde{S}_k = S_j$, the localization uncertainty U, the approximative displacement D, and the observation number M, the potential states, i.e. the set of database images considered for matching is defined according to the schema on Fig. 5.
- Π: $\Pi = \{\pi_j\}_{j=1}^{j=N}$ where $\pi_j = P[q_1 = S_j]$. It depends on initial position estimate (i.e. estimated position by previous HMM) and localization uncertainty U. We use uniform uncertainty on interval of size $F = 1 + 2 \cdot \lceil U/D^{'} \rceil$.
- A: $A = \{a_{ij}\}$ where $a_{ij} = P[q_{t+1} = S_j | q_t = S_i]$, $1 \leq i, j \leq N$: To take into account odometric uncertainty for a displacement D, we defined A as $a_{ij} = \frac{D^{'}}{\Delta} rect_{\Delta/D'}(j - i - (D/D^{'}))$. [1]

C. Observation matrix

The observation matrix B is computed from visual similarity between the M observations and the set of potentially matching database images as shown in Fig 5.

Visual similarity measurement is based on a state of the art IR solution. During the navigation phase, SIFT descriptors for all interest points detected by a SIFT detector [15] are extracted in a similar way as during the off-line phase. A kNN voting algorithm is then performed: 1) For each descriptor of a query image the k nearest neighbours are found from a subset of database descriptors, i.e. those that are near to the putative robot position. This subset is determined thanks to the estimated robot position $\tilde{S}_k$, D, U, Δ and M. 2) As noisy interest points are usually detected in an image, a filtering process based on the ratio of the distance between the query descriptor and the first and second nearest neighbours is used [15]. 3) Query descriptors that match with multiple database descriptors are removed, and finally 4) Outliers are rejected through a geometric verification, i.e. a

[1]The function $rect_\alpha(x)$ is the rectangular function defined by $rect_\alpha(x) = 1$ if $|x| \leq \alpha$ and else $rect_\alpha(x) = 0$.

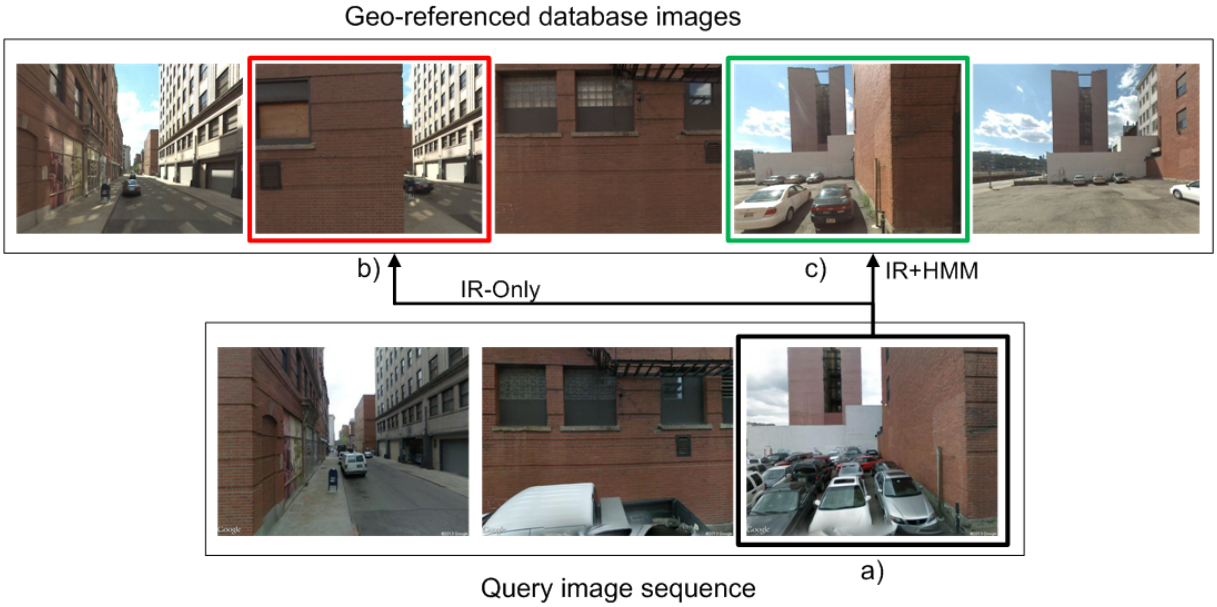

Fig. 4. Top images: Sequential database images considered for IR task - Bottom images: Sequential query images acquired by the robot - (a) Latest query image, (b) image returned by an IR algorithm only and (c) by an IR algorithm combined with a HMM.

RANSAC [16] 4-point algorithm [17] (homography). Hence, we get the number of descriptor correspondences between the descriptors of the query image O_k acquired by the robot and the descriptors of the database image I_j. This is our similarity measure, noted $f(O_k, I_j)$.

The observation matrix $B = \{b_j(k)\}$, where $b_j(k) = P[O_k \ at \ t | q_t = S_j]$, $1 \leq j \leq N$ and $1 \leq k \leq M$ is the probability of observing O_k when location is S_j. We propose to compute this probability from the similarity measure using the following formula:

$$b_j(k) = P[O_k \ at \ t | q_t = S_j] = \frac{\alpha}{1 + exp(a \cdot (f(O_k, I_j) + b))} \quad (2)$$

where a and b are two constants, $f(O_k, I_j)$ is the visual similarity measure and α is a normalization constant to impose $\sum_{j=1}^{j=N} b_j(k) = 1$.

A summary of the general estimation scheme is presented in algorithm 1.

Given $\lambda = \{N, M, \Pi, A, B\}$, (3) can be solved.

$$\hat{\mathbf{S}}_k = \arg\max_{\mathbf{S}} P(\mathbf{O}_k | \mathbf{S}, \lambda) \cdot P(\mathbf{S}, \lambda)$$
$$= \arg\max_{\mathbf{S}} \left(\prod_{k=1}^{k=M} P(O_k | \mathbf{S}, \lambda) \right) \cdot \left(\pi_1 \cdot \prod_{k=2}^{k=M} a_{k-1,k} \right) \quad (3)$$

The first term of (3) refers to visual similarities between observations and the image database, whereas the second term refers to the dynamics of the robot and models spatio-temporal constraints. We study in section IV the achieved performances by mixing these two complementary aspects.

Algorithm 1: Vision based global localization from odometric estimates

Input: Estimated robot position $\tilde{S}_k$, Localization uncertainty U, Estimated displacement D with odometric uncertainty Δ, M last past observations, Geo-referenced image features database.

Output: Corrected robot position $\hat{S}_k$.

1 HMM initialization (A and Π) from D, U, Δ and M as explained in section III-B;
2 Select relevant database images from estimated position $\tilde{S}_k$, D, U, Δ, and M (Fig. 5);
3 Compute similarities between the M past observations and relevant database images as explained in section III-C;
4 Compute B from similarities with (2);
5 Apply Viterbi algorithm to solve (3) to estimate the latest state $\hat{S}_k$;

IV. Experimental results

To evaluate our solution in a realistic situation, we conducted our experiments on a 11 km trajectory. The dataset used has been acquired at different times (more than one year between acquisitions) and with different camera fields of view resulting in visual changes for the same scenes (Fig. 2).

A. Image datasets and settings

We performed experiments on the Google Pittsburgh dataset as image database [3], and Google Streetview images as query images [4]. Pittsburgh dataset images have been resized to 640x480, so that their resolutions match the query image resolution. About 1160 SIFT descriptors are extracted and stored per image. From the original corpus, we keep one image every $D' = 5$ m and remove non-informative images (e.g. images acquired in tunnels) resulting in a corpus

of 2215 images. Query images are downloaded from the Internet via a HTTP request with the following settings: a resolution of 640x480, a field of view of 100^o and a camera tilt of 5^o. We requested one image every D = 10 m resulting in 1105 query images. For (2), a was set to 1 and b was set to -4.

B. Results

First, we compared our method (noted IR-HMM) to a state of the art IR algorithm based on visual similarities only (noted IR-Only). The meaningful metrics used are mean localization error and recall rate[2]. The recall rate increases from 36% to 84%, and the mean error localization decreases from to 16 m to 4 m (Tab. I). This considerable improvement confirms that exploiting the spatio-temporal constraint is essential. Our solution corrects ambiguous image matches (Fig. 4) thanks to spatio-temporal constraints imposed via the A matrix.

TABLE I

MEAN ERROR DISTANCE AND RECALL RATE FOR AN IR ALGORITHM BASED ON VISUAL SIMILARITIES ONLY (IR-ONLY), FOR AN IR ALGORITHM FOLLOWED BY A SPATIAL-TEMPORAL FILTER (IR-ST) AND FOR THE SAME IR ALGORITHM COMBINED WITH A HMM (IR-HMM) ON PITTSBURGH DATASET FOR M = 15 AND U = 50 M.

	IR-Only	IR-ST	IR-HMM
Mean error distance	15.8m	7.7m	3.9m
Recall	36.1%	71.2%	84.0%

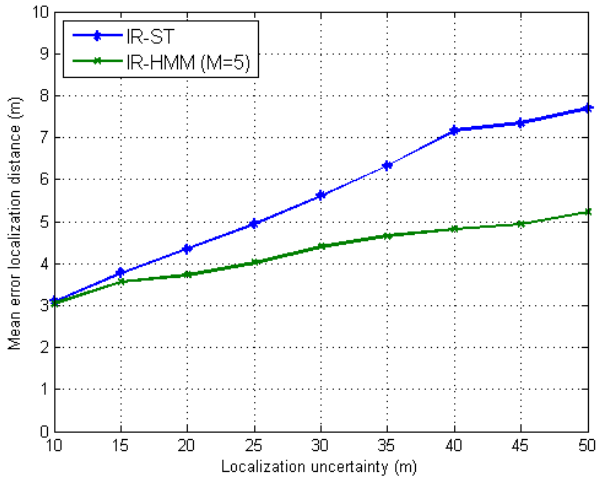

Fig. 6. Mean localization error distance vs. Localization uncertainty U. Spatio-temporal constraints reduce significantly false match that may appear with an IR algorithm, improving the mean error localization distance.

Then, we compared our solution with a method similar to the one described in [7] and reminded in section II: a Gaussian spatial filter is applied on putative positions obtained by query descriptor votes. Like ours, this method (noted IR-ST) takes into account spatio-temporal information. We

[2]True positive images are defined as geo-referenced images whose distance with ground truth image is less than 5 m.

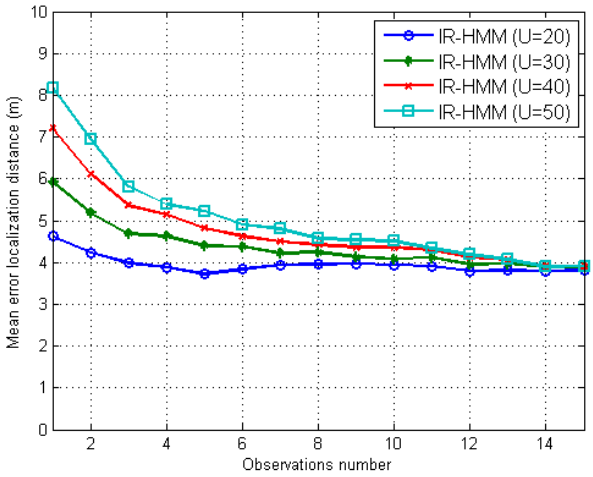

Fig. 7. Mean localization error distance vs. observation number M. When localization uncertainty increases, M must be also increased to guarantee a given mean localization error.

noticed that for a localization uncertainty of 50 m, the use of the HMM enables to decrease the mean localization error from 8 m to 4 m (Fig. 6). Furthermore, when localization uncertainty increases, performance differences between both solutions increase. The trajectory estimate with a HMM is more precise than with a spatio-temporal filter that tends to smooth the trajectory. HMM removes impossible matches, whereas in a spatio-temporal filter false matches are used for position estimates.

Finally, we studied the sensitivity of our solution to the number of past observations M used, according the localization uncertainty U (Fig. 7). The higher U, the more observations number have to be considered to keep the mean error localization under a threshold. As M approaches 0, only dynamics included in the A matrix is significant (3). In this case, our definition of A (possible transitions have equal probabilities) and Π (possible initial states have equal probabilities) implies equal probabilities for different states (3). The random selection performed among possible states explains the mean error localization increase.

Therefore, using dynamics only, or using visual similarity only, are insufficient in our context. Combining both improves significantly results.

V. CONCLUSION

We have proposed a general approach for global ego-localization able to combine noisy location estimates provided by an odometric system and visual place recognition. No GPS information is used. The solution exploits a Hidden Markov Model whose structure is adaptively defined from knowledge of the odometric system behaviour. Each new image acquisition by the robot allows a complete re-estimate of the M past observation locations ensuring odometric error correction on a long term basis.

The approach has been evaluated on the Pittsburgh Google dataset. We demonstrated the benefits of combining simple visual similarities and dynamics modelling: the proposed

solution improves significantly the mean error localization which decreases from 16 m to 4 m for a localization uncertainty of 50 m.

Improved image retrieval solutions can be easily integrated in the system without substantial structural modifications: this is the avenue of future work.

ACKNOWLEDGMENT

This work results from a collaboration between UPMC University, Onera and Thales Services SAS.

REFERENCES

[1] J. Ibañez-Guzmán, C. Laugier, J. Yoder, and S. Thrun, "Autonomous driving: Context and state-of-the-art," in *Handbook of Intelligent Vehicles*, ser. Springer Reference, A. Eskandarian, Ed. Springer, Mar 2012, pp. 1271–1310.

[2] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, vol. 2. IEEE, 2003, pp. 1470–1477.

[3] (2011) Pittsburgh dataset website (provided by google for research purposes). [Online]. Available: http://www.icmla-conference.org/icmla11/challenge.html

[4] (2014) Google street view API website. [Online]. Available: http://developers.google.com/maps/documentation/streetview

[5] M. Cummins and P. Newman, "Appearance-only slam at large scale with fab-map 2.0," *International Journal of Robotics Research*, vol. 30, pp. 1100–1123, August 2011.

[6] W. Maddern, M. Milford, and G. Wyeth, "Cat-slam: probabilistic localisation and mapping using continuous appearance-based trajectory," *International Journal of Robotics Research*, vol. 31, no. 4, pp. 429–451, April 2012.

[7] A. Zamir and M. Shah, "Accurate image localization based on google maps street view," in *Proceedings of the European Conference on Computer Vision*. IEEE, 2010, pp. 255–268.

[8] G. Vaca-Castano, A. Zamir, and M. Shah, "City scale geo-spatial trajectory estimation of a moving camera," in *Proceedings of the Computer Vision and Pattern Recognition conference*. IEEE, 2012, pp. 1186–1193.

[9] H. Badino, D. Huber, and T. Kanade, "Real-time topometric localization," in *Proceedings of the International Conference on Robotics and Automation*. IEEE, 2012, pp. 1635–1642.

[10] M. Brubacker, A. Geiger, and R. Urtasun, "Lost! leveraging the crowd for probabilistic visual self-localization," in *Proceedings of the Computer Vision and Pattern Recognition conference*. IEEE, 2013, pp. 3057–3064.

[11] M. Milford and G. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *International Conference on Robotics and Automation*. IEEE, 2012, pp. 1643–1649.

[12] A. Majdik, Y. Albers-Schoenberg, and D. Scaramuzza, "Mav urban localization from google street view data," in *Proceedings of the International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 3979–3986.

[13] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla, "Learning and calibrating per-location classifiers for visual place recognition," in *Proceedings of the Computer Vision and Pattern Recognition conference*. IEEE, 2013, pp. 907–914.

[14] J. Zhang, A. Hallquist, E. Liang, and A. Zakhor, "Location-based image retrieval for urban environments," in *Proceedings of the International Conference on Image Processing*. IEEE, 2011, pp. 3677–3680.

[15] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, November 2004.

[16] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381–395, 1981.

[17] A. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, 2nd ed. New York, NY, USA: Cambridge University Press, ISBN 978-0-521-54051-3, 2006.

[18] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, vol. 77, 1989, pp. 257–286.