Robust Wearable Camera Localization as a Target Tracking Problem on SE(3)

Guillaume Bourmaud
guillaume.bourmaud@ims-bordeaux.fr
Audrey Giremus
audrey.giremus@ims-bordeaux.fr

IMS Laboratory CNRS UMR 5218
University of Bordeaux
France

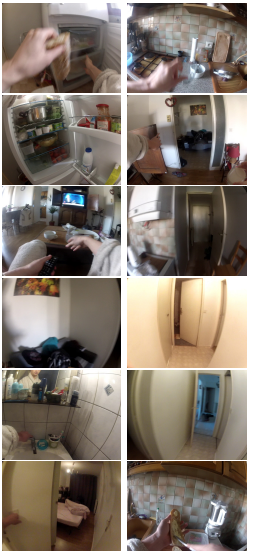
Abstract

This paper deals with the trajectory estimation of a wearable camera evolving in an indoor environment where a database of images has been previously annotated with map coordinates. The difficulty of this problem resides in the fact that: i) hand-held objects are frequently interposed between the camera and the environment during daily living activities; ii) strong motion blur and differences in illumination occur; iii) the environment changes between the images of the database and the video frames to localize, and the viewpoints can be significantly different. The contribution of this paper is threefold: 1) We formulate the localization problem as a **target tracking problem** on the Lie group of camera motions $SE(3)$, where the measurements are map coordinates obtained by applying a Content Based Image Retrieval algorithm to the video frames. 2) In order to solve this problem, we derive a novel **Rao-Blackwellized particle smoother on Lie groups**, which builds upon the recently proposed Extended Kalman Filter on Lie groups and the Rauch-Tung-Striebel Smoother on Lie groups that we also derive in this paper. 3) To take into account the topology of the environment, we propose to introduce **virtual measurements** that guide the particles and prevent them from crossing walls. We demonstrate, on several challenging video sequences, where the person wearing the camera performs daily living activities, that the proposed method is able to efficiently deal with outlier measurements and achieves a sub-meter level accuracy while the state of the art algorithms lack in robustness.

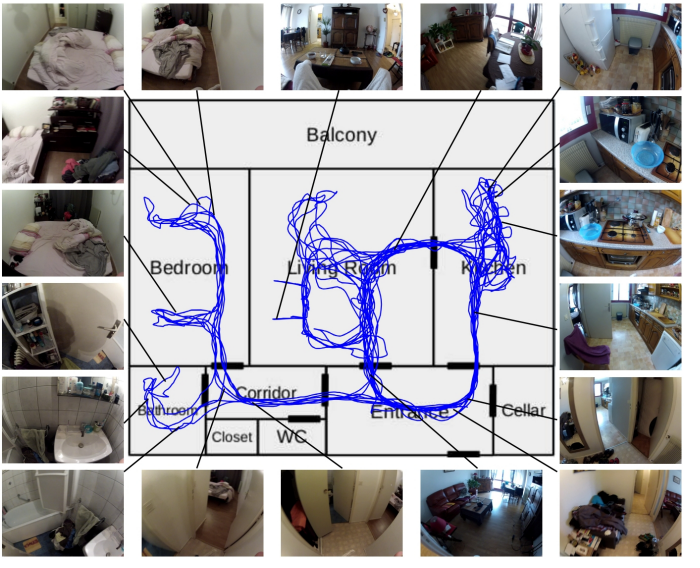
1 Introduction

Localizing a monocular camera evolving in an indoor environment where a database of images has been previously annotated with map coordinates, also known as Visual Indoor Localization (VIL), is a challenging problem. As a matter of fact, a wide range of applications, such as location based services, autonomous mobile robots or recognition of instrumental activities of daily living [12], require a meter level accuracy as well as the knowledge of the orientation [16]. Compared with other technologies such as Radio-frequency identification or WiFi radios, a monocular wearable camera is an informative low-cost passive sensor that does not require any modification of the environment. Thus being able to robustly and accurately estimate *both the position and the orientation* of a camera is essential.

Pages 39.1-39.12
DOI: https://dx.doi.org/10.5244/C.29.39



(a) Video frame examples of a person performing daily living activities. Hand-held objects are frequently interposed between the camera and the environment.



(b) Illustration of a database of images of an apartment annotated with map coordinates. This database of 6000 images was generated automatically from a training video sequence as explained in section 3.1. The blue line corresponds to the 2D positions of all the images of the database. Note that each image is actually annotated with a 6-dof pose (3D position and 3D orientation) but only the 3D position projected onto the plan of the apartment is presented in this figure.

Figure 1: (Best seen in colors) Left: Examples of video frames to localize. Right: Illustration of the database

1.1 Context and Objectives

In this paper, we are interested in VIL for challenging video sequences coming from a single monocular camera where the person wearing the camera performs daily living activities (see Fig.1(a)). The difficulty of this problem resides in the fact that: i) hand-held objects are frequently interposed between the camera and the environment; ii) strong motion blur and differences in illumination occur; iii) the environment changes between the images of the database and the video frames to localize, and the viewpoints can be significantly different.

We wish to develop a method that:

- relies only on the images coming from the wearable camera, i.e no other sensor such as Inertial Measurement Units should be used
- estimates the camera position with a sub-meter level accuracy as well as its orientation
- is consistent with the topology of the environment, i.e the camera trajectory should not cross walls
- is able to detect when the data is not sufficient to disambiguate the situation, i.e when the posterior distribution of the camera trajectory is multi-modal and/or too dispersed.

1.2 Contributions and Outline of the paper

In this context, we propose a novel VIL framework which is able to satisfy the previous technical specifications. The contribution of this paper is threefold:

1. We formulate the localization problem as a **target tracking problem** on the Lie group of camera motions $SE(3)$, where the measurements are map coordinates obtained by applying a Content Based Image Retrieval (CBIR) algorithm to the video frames.

2. In order to solve this problem, we derive a novel **Rao-Blackwellized particle smoother on Lie groups**, which builds upon the recently proposed Extended Kalman Filter on Lie groups [8] and the Rauch-Tung-Striebel Smoother on Lie groups that we also derive in this paper.
3. To take into account the topology of the environment, we propose to introduce **virtual measurements** that guide the particles and prevent them from crossing walls.

The rest of the paper is organized as follows: section 2 deals with the related work. Our VIL framework is presented in section 3. The formulation of the target tracking problem as well as the proposed Rao-Blackwellized particle smoother on Lie groups are described in section 4. In section 5, the limitations of the proposed approach are discussed, while in section 6, our VIL framework is evaluated experimentally. Finally, conclusions and future research directions are provided in section 7.

2 Related Work

There exist several ways to tackle the VIL problem.

First of all, it can be seen as a monocular Visual Simultaneous Localization and Mapping (VSLAM) problem with the additional knowledge of a database of images that are annotated with map coordinates. Thus, recent advances in monocular VSLAM such as [11] or [10] could be applied.

Secondly, from the database of images, a 3D point cloud of the environment can be reconstructed using a Structure from Motion (SfM) method such as [17] or [12]. Then, using this point cloud, it is possible to localize the video frames as in [25], [24], [20] or [18].

Finally, the approaches developed in the context of CBIR [7, 21] and appearance-only SLAM [8] are able to efficiently retrieve the nearest neighbor of an image in the database of annotated images. Consequently, the map coordinates of a retrieved image can be interpreted as the location of the query video frame.

All the previously cited approaches, are dedicated to "clean" data and fail when they are applied to the difficult problem of localizing a wearable camera during daily living activities (see section 1.1). Indeed, monocular VSLAM approaches are still fragile and require specific conditions, such as large camera translations, brightness constancy and a static environment, that are not met in our context. 3D-based localization algorithms assume that the environment did not change between the images of the database and the test videos. This assumption is usually verified for outdoor environments, but it is far from being true in an occupied apartment (see Fig.1(a) and 2). Image retrieval algorithms frequently output wrong results because of viewpoint changes and hand-held objects that are interposed between the camera and the environment (see Fig.2).

In this paper, we propose a method that overcomes the lack of robustness in the state of the art approaches. As a consequence, our method is able to deal with the VIL problem in the context of daily living activities. It consists in first employing a CBIR algorithm and then using the map coordinates of the retrieved images as measurements in a *target tracking problem* on the Lie group of camera motions $SE(3)$. The erroneous measurements are filtered by leveraging the fact that the camera:

- should have a smooth trajectory
- evolves in a constrained indoor environment and thus cannot cross walls.

The approaches described in [15], [22] and [27] are closely related to our own. For instance, [15] employs a Kalman filter while [22] and [27] use particle filters to estimate a smooth camera trajectory from the result of a CBIR algorithm. However, all these methods

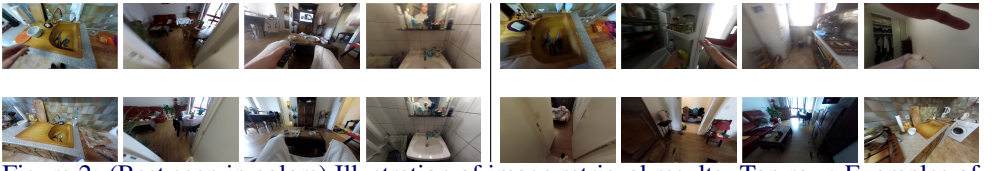


Figure 2: (Best seen in colors) Illustration of image retrieval results. Top row: Examples of video frames to localize. Bottom row: Nearest neighbor retrieved with a CBIR algorithm similar to [7]. Left: Examples of successfully retrieved images. Right: examples of images where the CBIR algorithm failed.

are assisted with other sensors, perform the trajectory estimation in a 2D plan and do not exploit the fact that the environment is constrained.

Contrary to these approaches, our method relies only on the images. In order to robustly estimate the 6-dof camera trajectory, we propose a novel *Rao-Blackwellized particle smoother on Lie groups* combined with *virtual measurements* that guide the particles and prevent them from crossing walls.

Also, in the field of place recognition, approaches such as [9] proposed to apply machine learning algorithms to learn places. These approaches are robust, however they cannot provide the camera position with a sub-meter level accuracy nor its orientation.

3 Proposed Visual Indoor Localization Framework

The proposed VIL framework consists in 2 modules: a CBIR algorithm followed by our novel Rao-Blackwellized Particle Smoother on Lie groups. The image retrieval algorithm that we employ is briefly presented in this section as well as the way we automatically build a database of images of an indoor environment, while the proposed Rao-Blackwellized Particle Smoother on Lie groups is described in the next section.

3.1 Automatic database creation and annotation

In order to apply a VIL algorithm, a database of images representing the indoor environment has to be created and annotated with map coordinates. While this step is usually tedious, we propose an (almost) automatic way to do it. First, a "clean" training video of the environment, as complete as possible, is taken. Then a monocular VSLAM approach similar to [11] is applied. Finally, the estimated camera trajectory is manually aligned with the 2D plan of the building. In this way, each frame of the training sequence has been automatically annotated with a 3D position and a 3D orientation. An example of database automatically created from a video of 20 minutes is presented Fig.1(b).

3.2 Content Based Image Retrieval Algorithm

The CBIR algorithm that we employ shares similarities with [20] and [7], yet different. Let us now describe this algorithm for one video frame. First of all, a 32×32 miniature of the frame is created [26] and compared to the miniatures of the database images using a sum of absolute distances. At the end of this stage, only the 100 closest database images are kept. Secondly, SURF points of interest [3] are detected and matched to the SURF descriptors of the database images using a kd-tree. At the end of this stage, only the 5 database images having the highest number of matches are kept. Finally, only the database images having at least 15 matches are kept. In order to reduce the computational cost of this step, the CBIR algorithm is only applied to one video frame per second. Examples of results produced by this algorithm are presented in Fig.2 (only the "nearest neighbor", i.e the database image

with the highest number of matches, is presented). This CBIR algorithm allows to efficiently find the nearest neighbors of a video frame in the annotated database of images.

4 A target tracking problem on $SE(3)$

After having found the nearest neighbors of the video frames in the annotated database of images, we wish to estimate the camera trajectory. First, we formulate this estimation problem as a target tracking problem on the Lie group of camera motions $SE(3)$ where the measurements are the map coordinates (3D orientation and 3D position) of the retrieved images. In order to take into account the topology of the environment, we propose to introduce virtual measurements that prevent the estimated trajectory from crossing walls. Then, in order to solve this problem, we derive a novel Rao-Blackwellized particle smoother on Lie groups.

4.1 Formulation of the problem

The map coordinates, obtained by applying a CBIR algorithm to the video frames, can be interpreted as target measurements coming from a target detector. Since the map coordinates take their values on the Lie group of camera motions $SE(3)$ (3D position and 3D orientation), we propose to formulate this estimation problem as a target tracking problem on $SE(3)$.

4.1.1 Preliminaries on the geometry of camera poses: the matrix Lie group $SE(3)$

A camera pose $C_{ig} = \begin{bmatrix} R_{ig} & T_{ig} \\ 0_{1\times 3} & 1 \end{bmatrix} \subset \mathbb{R}^{4\times 4}$ is a transformation matrix where R_{ig} is a 3D rotation matrix and T_{ig} is a 3D vector. Applying C_{ig} to a 3D point $x^g \in \mathbb{R}^3$ defined in a reference frame (RF) g allows to transform x^g from RF g to RF i, i.e $\begin{bmatrix} x^i \\ 1 \end{bmatrix} = C_{ig} \begin{bmatrix} x^g \\ 1 \end{bmatrix}$. Two poses C_{ji} and C_{ig} can be composed using matrix multiplication to obtain another pose $C_{jg} = C_{ji}C_{ig}$. Inverting a pose matrix C_{ig} produces the inverse transformation, i.e $C_{ig}^{-1} = C_{gi}$. Consequently multiplying a transformation with its inverse produces the identity matrix: $C_{ig}C_{gi} = Id$. From a mathematical point of view, the set of camera poses form the 6-dimensional matrix Lie group $SE(3)$ [6]. The matrix exponential expm and matrix logarithm logm establish a local diffeomorphism between an open neighborhood of Id in $SE(3)$ and an open neighborhood of $0_{4\times 4}$ in the tangent space at the identity, called the *Lie Algebra* $\mathfrak{se}(3)$. $\mathfrak{se}(3)$ is a 6-dimensional vector space. We denote the linear isomorphism between $\mathfrak{se}(3)$ and $\mathbb{R}^6$ as follows: $[\cdot]^\vee_{SE(3)} : \mathfrak{se}(3) \to \mathbb{R}^6$ and $[\cdot]^\wedge_{SE(3)} : \mathbb{R}^6 \to \mathfrak{se}(3)$. We also introduce the following notations: $\exp^\wedge_{SE(3)}(\cdot) = \text{expm}\left([\cdot]^\wedge_{SE(3)}\right)$ and $\log^\vee_{SE(3)}(\cdot) = [\text{logm}(\cdot)]^\vee_{SE(3)}$. Finally, the distribution of a random variable $C_{ig} \in SE(3)$ is called a (right) concentrated Gaussian distribution on $SE(3)$ [1] of "mean" μ_{ig} and "covariance" P denoted $\mathcal{N}_{SE(3)}(C_{ig}; \mu_{ig}, P)$ if:

$$C_{ig} = \exp^\wedge_{SE(3)}\left(\varepsilon^i_{ig}\right)\mu_{ig} \text{ where } \varepsilon^i_{ig} \sim \mathcal{N}_{\mathbb{R}^6}\left(\varepsilon^i_{ig}; 0_{6\times 1}, P\right) \tag{1}$$

Such a distribution provides a meaningful covariance representation and allows us to quantify the uncertainty of the camera poses.

4.1.2 Propagation model

First, we wish to take advantage of the fact that the camera should have a smooth trajectory. We propose to employ a white noise acceleration model of the form:

$$C_{t+1} = \exp^\wedge_{SE(3)}(v_t \Delta t)\, C_t \quad \text{and} \quad v_{t+1} = v_t + n_t \tag{2}$$

where $C_t \in SE(3)$ is the camera pose at time t. More precisely, C_t is defined as the transformation matrix $C_{i_t g}$ where g is the RF of the map and i_t is the RF of the camera at time t. $v_t \in \mathbb{R}^6$ corresponds to its speed, $n_t \sim \mathcal{N}_{\mathbb{R}^6}(n_t; 0_{6\times 1}, R_t)$ is a white Gaussian noise and Δt is the time interval between two consecutive frames. Consequently, the state $X_t \in SE(3) \times \mathbb{R}^6$ that we wish to estimate is the concatenation of the camera pose C_t and its speed v_t.

4.1.3 Likelihood, Virtual Measurements and Latent variables

We wish to define a likelihood involving the map coordinates of the images retrieved by the CBIR algorithm. However, the CBIR algorithm might have failed and retrieved wrong images (see Fig.2). Thus, we need to define a robust likelihood that allows to discard the retrieved map coordinates when they are erroneous. Also, when the CBIR algorithm does not manage to retrieve any image in the database, nothing prevents the camera pose neither from "leaving" the map, nor from crossing walls.

In order to answer these two problems, we first define the measurement variable y_t that concatenates:
- all the N map coordinates of the database images that we call Virtual Measurements
- the N_{CBIR} map coordinates (maximum $N_{CBIR} = 5$, see section 3.2) of the retrieved images at time t .

$$y_t = [\underbrace{y_t(1), y_t(2), ..., y_t(N)}_{\text{Virtual Measurements}}, \underbrace{y_t(N+1), ..., y_t(N+N_{CBIR})}_{\text{CBIR output}}] \tag{3}$$

Secondly, we introduce a latent discrete variable s_t that acts as a selector among the components of y_t such that $y_t(s_t = i)$ selects the ith component in y_t. Let us assume that, at time t, the CBIR algorithm retrieved N_{CBIR} images in the database. Then, $s_t \in \{1, ..., N+N_{CBIR}\}$ and $y_t(s_t)$ corresponds to the s_tth component of y_t. The likelihood we consider is defined as:

$$p(y_t|C_t, s_t = i) = \prod_{j=1}^{N+N_{CBIR}} p(y_t(j)|C_t, s_t = i) \propto \mathcal{N}^R_{SE(3)}(y_t(i); C_t, Q_{t,i}) \tag{4}$$

where $Q_{t,i} = Q_{VM}$ if $i \leq N$ and $Q_{t,i} = Q_{CBIR}$ otherwise. Q_{VM} and Q_{CBIR} are covariance matrices to be defined.

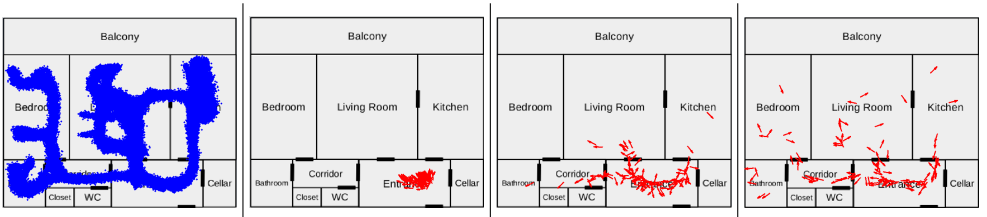
By introducing both virtual measurements and latent variables, the system is now able to select a virtual measurement instead of a true measurement when the CBIR algorithm retrieves wrong images. Moreover, when the CBIR algorithm does not retrieve any image, the system still selects a virtual measurement, which prevents the estimate camera pose from crossing walls (see Fig.3).

The probability transition of s_t is defined as $p(s_t|s_{t-1}) = p(s_t)$. In practice, $p(s_t = i) \ll p(s_t = j)$ for $i \leq N$ and $j > N$ in order to encourage the system to "use" the map coordinates retrieved by the CBIR algorithm as much as possible.

4.2 Rao-Blackwellized particle smoother on Lie groups

In order to satisfy the technical specifications described in section 1.1, we are interested in approximating the posterior distribution $p(X_{0:T}, s_{1:T}|y_{1:T})$[1]. To do so, we propose to derive a Rao-Blackwellized particle smoother on Lie groups (LG-RBPS) that samples the discrete latent variables s_t but takes advantage of the properties of (2) and (4) to marginalize out X_t.

[1]The notation $p(X_{0:T}, s_{1:T}|y_{1:T})$ stands for $p(X_0, X_1, ..., X_T, s_1, s_2, ..., s_T|y_1, y_2, ..., y_T)$



(a) 20 samples of each virtual measurement are plotted. It can be interpreted as all the places where the camera is "allowed" to go.

(b) t=0.3: In this example, all the particles are initialized at the same position and orientation with null speed

(c) t=3 sec: The particles begin to spread inside rooms, through doors and along the corridors.

(d) t=6 sec: The particles have now reached all the rooms.

Figure 3: Illustration of the diffusion of the particles using the virtual measurements. Only the projection of the 3D position (base of the arrow) and 3D orientation (direction of the arrow) onto the plan of the apartment is presented in this figure.

4.2.1 Related work

In the field of SLAM, several works proposed Rao-Blackwellized Particle Smoothers (RBPS) [4]. Nevertheless these approaches are "classical" RBPS and cannot be considered as RBPS on Lie groups because the Lie group part of the state is sampled and not marginalized out. Moreover, they do not deal with measurements evolving on a Lie group. In [19], a Rao-Blackwellized Particle Filter (RBPF) able to deal with measurements on a Lie group is proposed. However, once again, the Lie group part of the state is sampled.

In [2], an invariant Rao-Blackwellized Particle Filter (RBPF) was recently proposed. It is dedicated to systems possessing invariance properties once some of the state variables are known.

Contrary to these approaches, we propose an LG-RBPS algorithm able to deal with measurements evolving on a Lie group and to marginalize out the Lie group part of the state completely. It builds upon the recently proposed Extended Kalman Filter on Lie groups (LG-EKF) [5] and the Rauch-Tung-Striebel Smoother on Lie groups (LG-RTS) that we also derive in this paper. Our approach can be seen as a generalization of [13] to Lie groups.

4.2.2 Algorithm

Mathematical Derivations Due to the lack of space, the mathematical derivations of both the proposed Rao-Blackwellized particle smoother on Lie groups and the Rauch-Tung-Striebel Smoother on Lie groups are provided as supplementary material. However, the pseudo-code of the LG-RBPS is presented in Alg 1 and a detailed explanation is provided below.

Explanation The LG-RBPS presented in Alg 1, which allows to estimate the camera trajectory, consists in two main steps: filtering and smoothing.

In the filtering part, N_p weighted trajectories (or particles) are computed. A trajectory is parametrized by its mean μ and its covariance P as well as a weight w at each time instant. We now explain how these trajectories are obtained. The mean of the trajectory is initialized by choosing randomly a camera pose among the map coordinates with a null speed. The covariance of the trajectory is initialized with a predefined covariance matrix P_0 and the initial value of w is $\frac{1}{N_p}$. Then for each time instant t, from $t = 1$ (first video frame) to $t = T$ (last video frame), the algorithm operates as follows:

1. LG-EKF propagation: the mean and the covariance of the trajectory are propagated using the motion model (2)

2. Optimal Sampling: the likelihood of each measurement available at time instant t given the propagated trajectory is computed and used to draw one of the measurements
3. LG-EKF update: the mean and the covariance of the propagated trajectory are corrected (using the measurement model (4)) by taking into account the information coming from the selected measurement
4. the weight of the trajectory is updated by summing all the likelihoods from step 2

After having applied these four steps to each trajectory, the weights are normalized and unlikely trajectories are replaced with more likely ones during the resampling stage.

The trajectories produced by the filtering stage might contain "jumps" because of the resampling stage. More importantly, at each time instant, only the past measurements (summarized by the covariance of the trajectory) are used to select the next measurement, while the future measurements are not. However, being able to take into account the future measurements usually significantly improves the performances of an algorithm especially in cases where past measurements are not informative. This is the purpose of the smoothing stage.

In the smoothing part, the algorithm runs backward, that is from $t = T$ to $t = 1$, and draws N_p samples from the full posterior distribution $p(X_{0:T}, s_{1:T}|y_{1:T})$. For each sample, the algorithm starts from one of the (weighted) trajectories produced by the filtering stage and recursively (involving the LG-RTS smoother) tries to find a likely path among the possible trajectories produced by the filtering stage.

Finally, for each time instant t, the centroid and the covariance of the values of the samples (at time instant t) are computed.

The centroids are taken as estimate for the camera trajectory, while the covariances can be employed to detect when the posterior distribution of the camera trajectory is multi-modal and/or dispersed (see video provided as supplementary material).

5 Limitations

The VIL framework proposed in this paper has several limitations. First of all, the database of the environment has to be as complete as possible. Actually, if the camera goes in a place that does not correspond to any image of the database, then the CBIR algorithm will not retrieve any image and the particles will diffuse in the environment. In practice, the proposed automatic database creation framework (see section 3.1) allows to create large and complete databases very efficiently. Secondly, the introduction of virtual measurements prevents the particles from crossing walls only "softly". In theory, a particle with a high velocity might cross a wall and reach another virtual measurement. In practice, we constrain the velocity below a given physically realistic threshold and use a small time interval Δt to prevent such "jumps".

6 Experiments

In this section, the proposed VIL framework is evaluated experimentally on several challenging video sequences where the person wearing the camera performs daily living activities.

For all these experiments, the parameters of our VIL framework have been optimized by hand on one video once and for all. We used: $Q_{VM} = \text{diag}\left(\pi_{1\times 3}, (5e-3)_{1\times 3}\right)$, $Q_{CBIR} = \text{diag}\left(0.1_{1\times 3}, (5e-2)_{1\times 3}\right)$, $R_t = \text{diag}\left((5e-3)_{1\times 6}\right)$, $P_0 = \text{diag}\left(0.5_{1\times 3}, (5e-2)_{1\times 3}, 1_{1\times 6}\right)$, $N_p = 100$ and $p(s_t = i) = 0.1$ for $i > N$. In practice, when updating the weights, if $s_t^{(i)}$ corresponds to selecting a virtual measurement, then $p\left(y_t|s_{0:t-1}^{(i)}, y_{1:t-1}\right)$ is replaced by a constant. This allows not to modify the weights of the particles when the CBIR algorithm does not retrieve any image. Also, each $15\Delta t$, we use $R_t = P_0$ to provide more variability to the particles.

Algorithm 1 LG-RBPS

Inputs: N_p (number of particles), P_0 (initial covariance), T (number of time steps), $\{y_t\}_{t=1,...,T}$ (measurements), Q_{VM} (virtual measurement covariance), Q_{CBIR} (true measurement covariance), Δt (time interval), R_t(propagation covariances),$\{p(s_t)\}_{t=1,...,T}$ (discrete probability distributions)

Outputs: $\{\mu_t\}_{t=1,...,T}$ (estimated trajectory), $\{P_t\}_{t=1,...,T}$ (covariance of the estimated trajectory)

Notations: $\mu^{(i)}_{t_a|t_b}$ and $P^{(i)}_{t_a|t_b}$ correspond to the mean and the covariance of the trajectory associated to particle i, at time t_a, having observed $\{y_t\}_{t=1,...,t_b}$

Filtering

- For $i = 1,2,...,N_p$
  - Initialize $\mu^{(i)}_{0|0}$ by choosing randomly a map coordinate in the database with a null speed
  - $P^{(i)}_{0|0} = P_0$ and $w^{(i)}_0 = \frac{1}{N_p}$
- EndFor
- For $t = 1,2,...,T$
  - For $i = 1,2,...,N_p$
    - **LG-EKF propagation**: Propagate $\mu^{(i)}_{t-1|t-1}$ and $P^{(i)}_{t-1|t-1}$ to get $\mu^{(i)}_{t|t-1}$ and $P^{(i)}_{t|t-1}$ using R_t and Δt
    - **Optimal Sampling**: Draw $s^{(i)}_t$ using $\mu^{(i)}_{t|t-1}$, $P^{(i)}_{t|t-1}$, y_t and $p(s_t)$ and evaluate $p\left(y_t|s^{(i)}_{0:t-1}, y_{1:t-1}\right)$
    - **LG-EKF update**: Update $\mu^{(i)}_{t|t-1}$ and $P^{(i)}_{t|t-1}$ to get $\mu^{(i)}_{t|t}$ and $P^{(i)}_{t|t}$ using $y_t\left(s^{(i)}_t\right)$ and Q_{CBIR} or Q_{VM}
    - Update weight: $w^{(i)}_t = w^{(i)}_{t-1} p\left(y_t|s^{(i)}_{0:t-1}, y_{1:t-1}\right)$
  - EndFor
  - Normalize weights and Resample particles
- EndFor

Smoothing

- For $i = 1,2,...,N_p$
  - Set $\tilde{s}_T = s^{(j)}_T$ with probability $w^{(j)}_T$
  - Set $\tilde{\mu}_{T|T} = \mu^{(j)}_{T|T}$, $\tilde{P}_{T|T} = P^{(j)}_{T|T}$
  - Draw $x^{(i)}_T \sim \mathcal{N}_{SE(3)\times\mathbb{R}^6}\left(x_T; \tilde{\mu}_{T|T}, \tilde{P}_{T|T}\right)$
  - For $t = T-1, T-2, ..., 1$
    - For $k = 1,2,...,N_p$
      - Set r with the value of $\mathcal{N}_{SE(3)\times\mathbb{R}^6}\left(X_{t+1}; \mu^{(k)}_{t+1|t}, P^{(k)}_{t+1|t}\right)$ evaluated in $X_{t+1} = x^{(i)}_{t+1}$
      - $w^{(k)}_{t|t+1} \propto w^{(k)}_t p(\tilde{s}_{t+1})\, r$
    - EndFor
    - Set $j = k$ with probability $w^{(k)}_{t|t+1}$
    - $\tilde{s}_t = s^{(j)}_t$, $\tilde{\mu}_{t|t} = \mu^{(j)}_{t|t}$ and $\tilde{P}_{t|t} = P^{(j)}_{t|t}$
    - **LG-RTS Smoother**: Smooth $\tilde{\mu}_{t|t}$ and $\tilde{P}_{t|t}$ to get $\tilde{\mu}_{t|T}$ and $\tilde{P}_{t|T}$ using $\tilde{\mu}_{t+1|T}$ and $\tilde{P}_{t+1|T}$
    - Draw $x^{(i)}_t \sim \mathcal{N}_{SE(3)\times\mathbb{R}^6}\left(x_t; \tilde{\mu}_{t|T}, \tilde{P}_{t|T}\right)$
  - EndFor
- EndFor

Finally, for $t = 1,2,...,T$ compute the centroid μ_t and the covariance P_t of the samples $\left\{x^{(i)}_t\right\}_{i=1,...,N_p}$

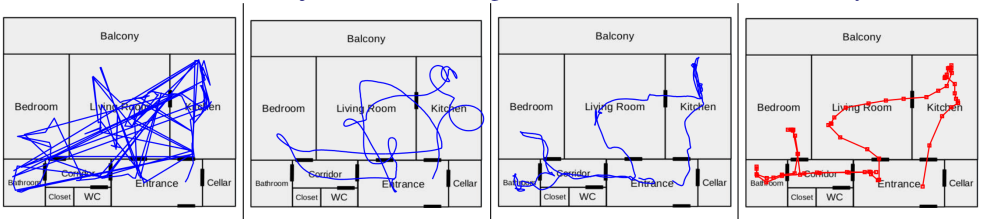
Technical details about **LG-EKF propagation**, **Optimal Sampling**, **LG-EKF update** and **LG-RTS Smoother** are provided as supplementary material.

We manually annotated the trajectories of 6 different videos where the person wearing the camera performs daily living activities. As explained in section 2, to the best of our knowledge, there are not concurrent approach to address the VIL problem in this difficult context. Consequently, to evaluate the performances of our approach, we estimated the trajectories of the 6 video sequences with a CBIR algorithm only (see section 3.2) as it is proposed in [7], the proposed LG-RBPS without virtual measurement[2] and the LG-RBPS with virtual measurements. The RMSE (m) of each approach is presented in Table 1. For each video sequence, as expected, the proposed LG-RBPS, which employs the output of the CBIR algorithm, produces a significantly lower RMSE than the CBIR algorithm alone.

[2] In this case, the algorithm chooses at each time instant whether it is better to select a measurement among the available measurements or to discard all the available measurements (see [23]).

	GO80	GO81	GO82	GO83	GO84	GO85
CBIR only (similar to [7])	1.7	1.3	2.4	2.2	1.3	2.2
CBIR + LG-RBPS No Virt. Meas.	**0.5**	0.7	1.7	0.9	**<0.5**	1.3
CBIR + LG-RBPS With Virt. Meas.	**0.5**	**<0.5**	**0.8**	**0.7**	**<0.5**	**0.9**

Table 1: Results on 6 challenging video sequences (GO80,..., GO85). Examples of video frames from these videos are presented Fig.1(a) and Fig.2. The figures represent the RMSE in meter of the estimated trajectories w.r.t the ground truth which has an accuracy of 0.5m.



(a) Result of the CBIR algorithm (nearest neighbor) similar to [7]

(b) CBIR+LG-RBPS without virtual measurements

(c) CBIR+LG-RBPS with virtual measurements

(d) Ground truth

Figure 4: Illustration of estimated trajectories on the video sequence GO82. Only the projection of the 3D position onto the apartment plan is presented.

Adding virtual measurements also significantly improves the performances of the LG-RBPS when the CBIR algorithm produces poor results. In Fig.4, the camera trajectory estimated with the 3 different approaches on the video sequence GO82 are presented. For this video sequence, the proposed LG-RBPS with virtual measurements is the only approach where the estimated camera trajectory does not cross walls. A video illustrating how the particles can be employed to detect when the posterior distribution of the camera trajectory is multi-modal and/or dispersed is provided as supplementary material.

7 Conclusion

In this paper, we presented a Visual Indoor Localization system for challenging video sequences coming from a single monocular camera where the person wearing the camera performs daily living activities. We derived a novel Rao-Blackwellized particle smoother on Lie groups that allows to significantly improve the output of a Content Based Image Retrieval (CBIR) algorithm and thus overcomes the lack of robustness in the state of the art approaches. Moreover, we proposed to introduce virtual measurements in order to guide the particles and prevent them from crossing walls especially when the CBIR algorithm produces poor results. To the best of our knowledge, it is the first time that a VIL system, relying on the video frames only, is able to estimate the camera trajectory with a sub-meter accuracy when: i) hand-held objects are frequently interposed between the camera and the environment; ii) strong motion blur and differences in illumination occur; iii) the environment changes between the images of the database and the video frames to localize.

Future work will focus on robust visual odometry to help in the guidance of the particles when the CBIR algorithm fails because of changes in the environment.

Acknowledgments The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007- 2013) under grant agreement 288199 - Dem@Care. The authors would like to thank the reviewers and Cornelia Vacar for their valuable help.

References

[1] Timothy D. Barfoot and Paul T. Furgale. Associating uncertainty with three-dimensional poses for use in estimation problems. *IEEE Trans. Robot.*, 30(3):679–693, Jun 2014. ISSN 1941-0468. doi: 10.1109/tro.2014.2298059.

[2] Axel Barrau and Silvere Bonnabel. Invariant particle filtering with application to localization. In *IEEE Conference on Decision and Control*, 2014.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006*, pages 404–417. Springer, 2006.

[4] Kristopher R Beevers and Wesley H Huang. Fixed-lag sampling strategies for particle filtering slam. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 2433–2438. IEEE, 2007.

[5] Guillaume Bourmaud, Rémi Mégret, Audrey Giremus, and Yannick Berthoumieu. Discrete extended Kalman filter on Lie groups. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, 2013.

[6] Gregory S. Chirikjian. *Stochastic Models, Information Theory, and Lie Groups, Volume 2*. Springer-Verlag, 2012. ISBN 978-0-8176-4943-2. doi: 10.1007/978-0-8176-4944-9.

[7] Ciarán Ó Conaire, Michael Blighe, and Noel E O'connor. Sensecam image localisation using hierarchical SURF trees. In *Advances in Multimedia Modeling*, pages 15–26. Springer, 2009.

[8] Mark Joseph Cummins and Paul M. Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *I. J. Robotic Res.*, 27(6):647–665, 2008.

[9] Vladislavs Dovgalecs. *Indoor location estimation using a wearable camera with application to the monitoring of persons at home*. PhD thesis, 2011.

[10] Ethan Eade. *Monocular simultaneous localisation and mapping*. PhD thesis, 2008.

[11] Jakob Engel, Thomas Schops, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular SLAM. *ECCV, Lecture Notes in Computer Science*, pages 834–849, 2014. ISSN 1611-3349. doi: 10.1007/978-3-319-10605-2_54.

[12] Olof Enqvist, Fredrik Kahl, and Carl Olsson. Non-sequential structure from motion. In *ICCV Workshops*, pages 264–271, 2011.

[13] William Fong, Simon J Godsill, Arnaud Doucet, and Mike West. Monte carlo smoothing with application to audio signal enhancement. *Signal Processing, IEEE Transactions on*, 50(2):438–449, 2002.

[14] Ivan Gonzalez Diaz, Vincent Buso, and Jenny Benois-Pineau. Modeling instrumental activities of daily living in egocentric vision as sequences of active objects and context for alzheimer disease research. In *Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare*, pages 11–14. ACM, 2013.

[15] Jihun Ham, Yuanqing Lin, and Daniel D Lee. Learning nonlinear appearance manifolds for robot localization. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 2971–2976. IEEE, 2005. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1545149.

[16] Robert Huitl, Georg Schroth, Sebastian Hilsenbeck, Florian Schweiger, and E Steinbach. Tumindoor: An extensive image and point cloud dataset for visual indoor localization and mapping. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1773–1776. IEEE, 2012.

[17] Vadim Indelman, Richard Roberts, Chris Beall, and Frank Dellaert. Incremental light bundle adjustment. In *BMVC*, 2012.

[18] Till Kroeger and Luc Van Gool. Video registration to sfm models. In *Computer Vision–ECCV 2014*, pages 1–16. Springer, 2014.

[19] Junghyun Kwon and Kyoung Mu Lee. Monocular slam with locally planar landmarks via geometric rao-blackwellized particle filtering on lie groups. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1522–1529. IEEE, 2010.

[20] Jason Zhi Liang, Nicholas Corso, Eric Turner, and Avideh Zakhor. Image based localization in indoor environments. In *Computing for Geospatial Research and Application (COM. Geo), 2013 Fourth International Conference on*, pages 70–75. IEEE, 2013.

[21] Miguel Lourencco, Vitor Pedro, and João Pedro Barreto. Localization in indoor environments by querying omnidirectional visual maps using perspective images. In *ICRA*, pages 2189–2195, 2012.

[22] Morgan Quigley, David Stavens, Adam Coates, and Sebastian Thrun. Sub-meter indoor localization in unmodified environments with inexpensive sensors. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 2039–2046. IEEE, 2010.

[23] Simo Särkkä, Aki Vehtari, and Jouko Lampinen. Rao-blackwellized particle filter for multiple target tracking. *Information Fusion*, 8(1):2–15, 2007.

[24] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 667–674. IEEE, 2011. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6126302.

[25] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *BMVC*, 2012.

[26] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(11):1958–1970, 2008.

[27] Jürgen Wolf, Wolfram Burgard, and Hans Burkhardt. Robust vision-based localization by combining an image-retrieval system with monte carlo localization. *Robotics, IEEE Transactions on*, 21(2):208–216, 2005.