A Fast and Incremental Method for Loop-Closure Detection Using Bags of Visual Words

Adrien Angeli, David Filliat, Stéphane Doncieux, Jean-Arcady Meyer

A Fast and Incremental Method for Loop-Closure Detection Using Bags of Visual Words

Adrien Angeli, David Filliat, Stéphane Doncieux, and Jean-Arcady Meyer

***Abstract*—In robotic applications of visual simultaneous localization and mapping techniques, loop-closure detection and global localization are two issues that require the capacity to recognize a previously visited place from current camera measurements. We present an online method that makes it possible to detect when an image comes from an already perceived scene using local shape and color information. Our approach extends the bag-of-words method used in image classification to incremental conditions and relies on Bayesian filtering to estimate loop-closure probability. We demonstrate the efficiency of our solution by real-time loop-closure detection under strong perceptual aliasing conditions in both indoor and outdoor image sequences taken with a handheld camera.**

***Index Terms*—Loop-closure detection, localization, SLAM.**

I. Introduction

Over the last decade, the increase in computing power has helped to supplement traditional approaches to simultaneous localization and mapping (SLAM [1], [2], [3], [4]) with the qualitative information provided by vision. As a consequence, in robotics research, commonly used range and bearing sensors such as laser scanners, radars and sonars tend to be associated with, or replaced by, single cameras or stereo-camera rigs. For example, in previous work [5], we performed vision-based 2D SLAM for Unmanned Aerial Vehicles (UAV). Likewise, in [6], the authors performed 3D SLAM in real-time at 30Hz using a monocular handheld camera, while the authors of [7] present visual SLAM solutions based on both monocular and stereo vision.

However, there are still difficulties to overcome in robotic vision in general, and in SLAM applications in particular. Among them, the loop-closure detection issue concerns the difficulty of recognizing already mapped areas, while the global localization issue concerns the difficulty of retrieving the robot's location in an existing map. Those problems can be addressed by detecting when the robot is navigating through a previously visited place from local measurements. The overall goal of the research effort reported in this article is thus to design a vision-based framework tackling these issues so as to make it possible for a robot to reinitialize a visual 3D-SLAM algorithm like one of those presented in [6] or [7] in such situations. This comes down to an online image retrieval task that consists in determining if current image has been taken from a known location. Such task bears strong similarities with image classification methods like those described in [8] and [9], but an important difference is our commitment to online processing.

In this paper, we present a real-time vision-based method to detect loop-closures in a Bayesian filtering scheme: at each new image acquisition, we compute the probability that the current image comes from an already perceived scene. To this end, we designed a scene recognition framework that relies on an incremental version [10] of the bag-of-words method [9]. Loop-closure hypotheses whose probability is above some threshold are confirmed when a coherent structure between the corresponding images is found - i.e. when the epipolar geometry constraint is satisfied. This ultimate validation step is accomplished using a multiple-view geometry algorithm similar to the one proposed in [11]. We provide experimental results demonstrating the quality of our approach by performing loop-closure detection in incremental and real-time conditions in both indoor and outdoor image sequences using a single monocular camera.

In section 2, we present a review of related work on visual loop-closure and global localization. Section 3 briefly introduces our implementation of the bag-of-words paradigm. The filtering scheme is detailed in section 4 and experimental results are given in section 5. The last two sections are devoted to discussion and conclusion.

II. Related Work

The Monte Carlo Localization (MCL) method was originally designed [12] to make global localization capitalizing on range and bearing sensors possible. Although successfully adapted to vision [13], this method does not match our requirements since it relies on the existence of a map obtained beforehand. From the same principle, the Rao-Blackwellised particle filter (RBpf) enables loop-closure capabilities in SLAM algorithms (e.g the FastSLAM [14] framework). It has also been adapted to vision [15], but it suffers degeneration when closing a loop due to inaccurate resampling policies [3]. In addition, RBpf are not loop-closure detection methods per se, but rather SLAM methods robust to loop-closure events.

Loop-closure detection has also been performed using an Extended Kalman Filter (EKF) application to visual SLAM ([16], [17]). The overall idea is to detect loop-closures from advanced data association techniques that try to match visual features found in current images with those stored in the map. This approach limits the information used to detect loop-closure to the information used for mapping (which is designed

Adrien Angeli (corresponding author), Stéphane Doncieux and Jean-Arcady Meyer are with the Université Pierre et Marie Curie - Paris 6, FRE 2507, ISIR, 4 place Jussieu, F-75005 Paris, France. Email: firstname.lastname@isir.fr

David Filliat is with Ecole Nationale Supérieure des Techniques Avancées, 32, bvd Victor, F-75015 Paris, France. Email: david.filliat@ensta.fr

Regular Paper

for SLAM, and not optimized for loop-closure detection). It is also linked to a particular SLAM algorithm, whereas our approach may be adapted to any SLAM method (even not vision-based).

In this work, we wish to design a simple visual system able to perform loop-closure detection and global localization, within the framework of an online image retrieval task. Following a similar approach, but in a non-incremental perspective, voting methods presented in [18] and [19] call upon maximum likelihood estimation to match the current image with a database of images acquired beforehand. The likelihood depends upon the number of feature correspondences between the images, and leads to a vote assessing the amount of similarity. In [18], the authors also use multiple-view geometry to validate each matching hypothesis, while in [19] the accuracy of the likelihood is qualitatively evaluated in order to reject outliers. Even though they are easy to implement, the aforementioned voting methods rely on an offline construction of the image database and need expensive one-to-one image comparisons when searching for the most likely hypotheses. Moreover, the maximum likelihood framework is not suitable for managing multiple hypotheses over time, as it does not ensure the time coherency of the estimation (i.e. information from past estimates is not integrated over time so as to be fused with actual ones). As a consequence, this framework is prone to transient detection errors, especially under strong perceptual aliasing conditions.

In [20] and [21], bag-of-words methods are used to perform global localization and loop-closure detection in an image classification scheme (see also [22] for an extended version of [21], with multi-robot map-joining addressed as a loop-closure problem). Bag-of-words methods ([8], [9]) rely on a representation of images as a set of unordered elementary features (the visual words) taken from a dictionary. The dictionary is built by clustering similar visual descriptors extracted from the images into visual words. Using a given dictionary, image classification is based on the occurrence of the words in an image to infer its class. In [20] and [21], images are represented as vectors of visual words' statistics with size equal to the number of words in the dictionary. The dictionary is built beforehand in an offline process, clustering the visual features extracted from a training database of images into representative words of the environment. Matching between current and past images is defined as a Nearest Neighbor (NN) search among the cosine distances separating the corresponding vectors. In [20], a simple voting scheme selects the n best candidates from the NN search and multiple-view geometry is used to discard outliers. In [21], the NN search results are used to fill a *similarity matrix* whose off-diagonal elements represent loop-closure events, thus providing a powerful way to manage multiple hypotheses. In both approaches, the use of a dictionary enhances the robustness of matches, enabling a good tolerance to image noise, but the NN search involved, relying on exhaustive one-to-one vector comparisons, is very expensive.

More recently, the authors of [23] have proposed a vision-based probabilistic framework that makes it possible to estimate the probability that two observations originate from the same location. This approach, based on the bag-of-words scheme, is very robust to perceptual aliasing: a generative model of appearance is learned in an offline process, approximating the probabilities of co-occurrences of the words contained in the offline-built dictionary. Using this model, loop-closure detection can be performed with a complexity linear in the number of locations. The main asset of this model is its ability to evaluate the distinctiveness of each word, thus accounting for perceptual aliasing at the word level, while its principal drawback lies in the offline process needed for model learning and dictionary computation.

In the majority of the methods presented above, SIFT (Scale Invariant Feature Transform [24]) features are the preferred input information because of their robustness to reasonable 2D affine transformations, scale and viewpoint changes. However, other visual features could be used for loop-closure detection and global localization (see [25] for a comparison of visual local descriptors). For example, as stated in [19], color histograms are powerful features providing a compact geometry-less image representation that exhibits some attractive invariance properties to viewpoint changes. Hence, it may be suitable to merge several complementary visual information, like shape and color for example, in order to obtain a reliable solution in different contexts.

III. Visual Dictionary

The implementation of the bag-of-words method used here is detailed in [10]: the dictionary construction is performed online along with the image acquisition, in an incremental fashion. The words are stored using a tree structure (see [26] for more details), enabling logarithmic-time complexity when searching for a word and thereby entailing real-time processing. In the work reported here, we used two different feature spaces to describe the images:

- SIFT features [24]: interest points are detected as maxima over scale and space in differences of Gaussians convolutions. The features are memorized as histograms of gradient orientations around the detected point at the detected scale. The corresponding descriptors are of dimension 128 and are compared using L2 distance.
- Local color histograms: the image is decomposed in a set of regularly spaced windows of several sizes to improve scale invariance. The normalized H histograms in the HSV color space for each window are used as features. The windows used here are of size 20x20 (respectively 40x40) taken every 10 (respectively 20) pixels. The descriptors are of dimension 16 and are compared using diffusion distance [27].

A dictionary is built for each feature space.

IV. Bayesian Loop-Closure Detection

In this paper, we address the problem of loop-closure detection as an image retrieval task: we are seeking for the past image, if it exists, that looks similar enough to the current one to consider that they come from close viewpoints. The overall processing, illustrated in the diagram of figure 1, is achieved in a Bayesian filtering framework estimating the probability that current and past images pertain to the same scene: we thus look for the past image that maximizes the probability of loop-closure with the current image. When such an image is found (i.e. when probability is high for a particular loop-closure hypothesis), the consistency of the structure underlying those two images is checked by a multiple-view geometry algorithm [11]. When perceptual aliasing is present in the environment (i.e. when different places look similar), epipolar geometry provides a powerful way to reject outliers (i.e. past images that look like the current image but do not come from the same scene). In order to take advantage of different types of information, several feature spaces (i.e. SIFT features and H histograms) are used here for representing the images. Compared to maximum likelihood methods, the Bayesian filtering scheme proposed here takes temporal coherency of image acquisition into account in order to bring robustness to transient detection errors.

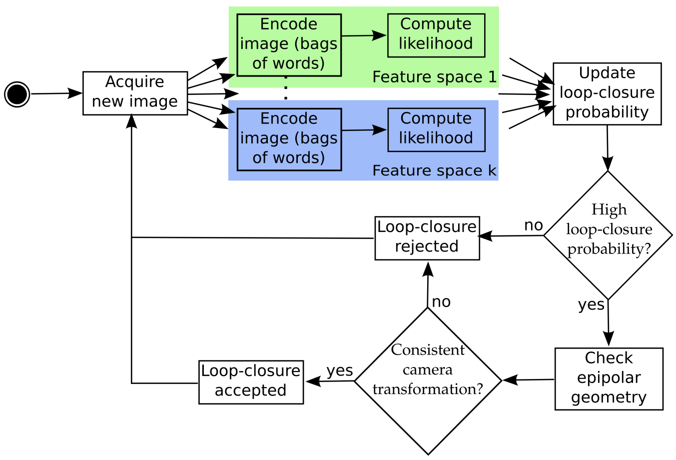

Fig. 1. Overall processing diagram (see text for details).

In this section, we first give the mathematical derivation of the filtering scheme used for the estimation of loop-closure probability. Then, we focus on issues regarding temporal coherency, likelihood computation and hypotheses management.

A. Discrete Bayes Filter

Let S_t be the random variable representing loop-closure hypotheses at time t. The event $S_t = i$ is the event that current image I_t "closes the loop" with past image I_i. This implies that the corresponding viewpoints x_t and x_i are close, and that I_t and I_i are similar. The event $S_t = -1$ is the event that no loop-closure occurred at time t. In a probabilistic Bayesian framework, the loop-closure detection problem can hence be formulated as searching for the past image I_j whose index satisfies:

$$j = \text{argmax}_{i=-1,\ldots,t-p} p(S_t = i | I^t) \qquad (1)$$

where $I^t = I_0, \ldots, I_t$, with $j = -1$ if no loop-closure has been detected. This search is not performed over the last p images because I_t always looks similar to its neighbors in time (since they come from close locations), and doing so would result in loop-closure detections between I_t and recently seen images (i.e. $I_{t-1}, I_{t-2}, \ldots, I_{t-(p+1)}$). This parameter, set to 10 in our experiments, is adjusted depending on the frame rate and on the velocity of camera motion.

We therefore need to estimate the *full posterior*, $p(S_t|I^t)$ for all $i = -1, \ldots, t-p$, in order to find, if a loop-closure occurred, the corresponding past image.

Following Bayes' rule and under the Markov assumption the posterior can be decomposed into:

$$p\big(S_t|I^t\big) = \eta p\big(I_t|S_t\big) p\big(S_t|I^{t-1}\big) \qquad (2)$$

where η is the normalization term. Let $(Z_k)_i$ be the state of the dictionary associated with the feature space k (SIFT features or H histograms in this paper) at time index i. The time subscript i is inherent to the incremental aspect of the dictionary construction: $(Z_k)_0 \subseteq (Z_k)_1 \subseteq \ldots \subseteq (Z_k)_{i-1} \subseteq (Z_k)_i$, with $(Z_k)_0 = \emptyset$ (features from the feature space k extracted in I_i are used to build $(Z_k)_{i+1}$). Also, let the subset $(z_k)_i$ of words taken from $(Z_k)_i$ and found in image I_i denote one representation of this image: $I_i \Leftrightarrow (z_k)_i$, with $(z_k)_i \subseteq (Z_k)_i$. Since several feature spaces are involved here, several image representations exist (one per feature space). Thus, let $(z^n)_i$ be the overall representation of image I_i, all feature spaces $k = 0, \ldots, n$ combined. The sequence of images I^t acquired up to time t can therefore be represented by the sequence $(z^n)^t = (z^n)_0, \ldots, (z^n)_t$.

So, the full posterior, now rewritten $p\big(S_t|(z^n)^t\big)$, can be expressed as follows:

$$p\big(S_t|(z^n)^t\big) = \eta p\big((z^n)_t|S_t\big) p\big(S_t|(z^n)^{t-1}\big) \qquad (3)$$

Assuming independence between the feature spaces, we can derive a more tractable mathematical formulation for equation 3 so as to make computation of the full posterior easier. However, capturing the correlations existing between the different dictionaries could provide additional information about the occurrence of the words. Under the independence assumption, the full posterior's expression can be written:

$$p\big(S_t|(z^n)^t\big) = \eta \left[\prod_{k=0}^{n} p\big((z_k)_t|S_t\big)\right] p\big(S_t|(z^n)^{t-1}\big) \qquad (4)$$

where the conditional probability $p\big((z_k)_t|S_t\big)$ is considered as a likelihood function $\mathcal{L}\left(S_t|(z_k)_t\right)$ of its second argument (i.e. S_t), with its first argument (i.e. $(z_k)_t$) held fixed: we evaluate, for each entry $S_t = i$ of the model, the likelihood of the currently observed words $(z_k)_t$ (see section IV-C).

Recursive estimation of the full posterior is made possible by decomposing the right hand side of equation 4 as follows:

$$p\big(S_t|(z^n)^t\big) = \eta \left[\prod_{k=0}^{n} p\big((z_k)_t|S_t\big)\right] \underbrace{\sum_{j=-1}^{t-p} p\big(S_t|S_{t-1}=j\big) p\big(S_{t-1}=j|(z^n)^{t-1}\big)}_{belief} \qquad (5)$$

where $p\big(S_t|S_{t-1}\big)$ is the time evolution model (see section IV-B) of the probability density function (p.d.f.). From equation 5, we can see that the estimation of the full posterior at time t is done by first applying the time evolution model to the previous estimation of the full posterior, leading to what we can call the *belief* at time t, which is in turn multiplied successively by the likelihoods obtained from the different feature spaces in order to get the actual estimation for the posterior.

Note that in our framework, the sequence of words $(z^n)^t$ evolve in time with the acquisition of new images, diverging from the classical Bayesian framework where such sequences would be fixed. Moreover, in spite of the incremental evolution of the dictionary, the representation of each past image is fixed and does not need to be updated.

B. Transition from $t-1$ to t

Between $t-1$ and t, the full posterior is updated according to the time evolution model of the p.d.f., $p\big(S_t|S_{t-1}=j\big)$, which gives the probability of transition from one state j at time $t-1$ to every possible state at time t. It therefore plays a key role in reducing transient detection errors by ensuring the temporal coherency of the detection. Depending on the respective values of S_t and S_{t-1}, this probability takes one of the following values:

- $p\big(S_t=-1|S_{t-1}=-1\big)=0.9$, the probability that no loop-closure event will occur at time t is high given that none occurred at time $t-1$.
- $p\big(S_t=i|S_{t-1}=-1\big)=\frac{0.1}{(t-p)+1}$ with $i\in[0;t-p]$, the probability of a loop-closure event at time t is low given that none occurred at time $t-1$.
- $p\big(S_t=-1|S_{t-1}=j\big)=0.1$ with $j\in[0;t-p]$, the probability of the event "no loop-closure at time t" is low given that a loop-closure occurred at time $t-1$.
- $p\big(S_t=i|S_{t-1}=j\big)$, with $i,j\in[0;t-p]$, is a Gaussian on the distance in time between i and j whose sigma value is chosen so that it is non zero for exactly 4 neighbors (i.e. $i=j-2\ldots j+2$). The size of this neighborhood is adjusted depending on the frame rate and on the velocity of camera motion. This corresponds to a diffusion of the posterior in order to account for the similarities between neighboring images.

Note that in order to have $p\big(S_t>=-1|S_{t-1}=j\big)=1$ when $j\in[0;t-p]$, the coefficients of the Gaussian used in the last case have to sum to 0.9.

C. Likelihood in a Voting Scheme

In section IV-A, we saw how using multiple feature spaces gave the opportunity to represent an image in different ways. From a perceptual point a view, each representation brings its own piece of information about the state of the world, independently from other feature spaces. This entails computing a likelihood measure for the loop-closure hypotheses S_t for each of the feature spaces considered. From the computational point of view, all the representations rely on the bag-of-words paradigm, providing a generic interface to compute and manage image representations. Therefore, the details given here about the estimation of the likelihood associated to a specific feature space k apply identically to each other feature space.

During the computation of the likelihood associated to the feature space k, we wish to avoid an exhaustive image-to-image comparison of the visual features, as implemented in most of the voting and bag-of-words methods cited in section II. In order to efficiently find the most likely past image I_i that closes the loop with the current one, we take advantage of the *inverted index* associated with the dictionary. The inverted index lists the images from which each word has been seen in the past. Then, during the quantization of the current image I_t with the words $(z_k)_t$ it contains, each time a word is found, we retrieve from the inverted index the list of the past images in which it has been previously seen. This list is used to update the score (originally set to 0) that is assigned to every loop-closure hypothesis $S_t=i$ in a simple voting scheme: when we find a word that has been seen in image I_i, statistics about the word are added to the score (see figure 2). The chosen statistics are inspired from the *term frequency–inverted document frequency (tf–idf)* weighting [28]:

$$\text{tf–idf}=\frac{n_{wi}}{n_i}\log\frac{N}{n_w} \tag{6}$$

where n_{wi} is the number of occurrences of word w in I_i, n_i is the total number of words in I_i, n_w is the number of images containing word w, and N is the total number of images seen so far. From equation 6, we can see that the tf–idf coefficient is the product of the term frequency (i.e. the frequency of a word in an image), by the inverted document frequency (i.e. the inverse frequency of the images containing this word). It is calculated each time a likelihood score is computed, giving increased emphasis to words seen frequently in a small number of images, and penalizing common words (i.e. words that are seen everywhere), according to the most recent statistics.

To summarize, when a word is found in the current image, the images where this word has been previously seen have their scores updated with the tf–idf coefficient associated with the pair $\{$*word–image*$\}$. The score associated with each loop-closure hypothesis $S_t=i$ will be used to compute the corresponding likelihood, as we shall see later on. But before, we must give some details about the computation of the score associated to the event "no loop-closure occurred at time t". Indeed, it is evaluated here as the event "a loop-closure is found with I_{-1}". I_{-1} is a virtual image built at each likelihood computation step with the m most frequently seen words of $(Z_k)_t$ (m being the average number of words found per image): it is the "most likely" image.

The idea is that the score associated with I_{-1} will change depending on the location of the current image, so as to behave as the score of the "no loop-closure" event. When no loop-closure occurs, I_t will be statistically more similar to I_{-1} than to any other I_i, because I_t will have more words in common with I_{-1} than with any other I_i. On the other hand, in a real unambiguous loop-closure situation, the score of I_{-1} will be low compared to the score of the loop-closing image I_i: as the words responsible for this detection are only present in two images (i.e. I_t and I_i), they are not frequently seen words and they are in consequence unlikely to be found in I_{-1}. The design of the virtual image proposed here is also relevant in case of perceptual aliasing (i.e. when I_t comes from a location that is similar to several previously visited places). In such situation, as multiple past images have equivalent likelihoods, it is important to ensure that I_{-1} receives a score that is in the same order of magnitude as the score of these images, so as to prevent an erroneous loop-closure detection. Here, as part of the most common words composing I_{-1} will originate from the images that are responsible for perceptual aliasing, it is guaranteed that I_{-1} will be granted with an important score (but not necessary the highest one).

The construction of a virtual image with existing words is similar to the addition of new locations from words sampling used in [23]. In our filtering scheme, the existence of the virtual image can be simulated simply by adding a I_{-1} entry to the inverted index for each of the most frequently seen words. Therefore, if one of them is found in I_t, it will vote for I_{-1} as shown in figure 2 and the corresponding score will be computed as for the "true" images.

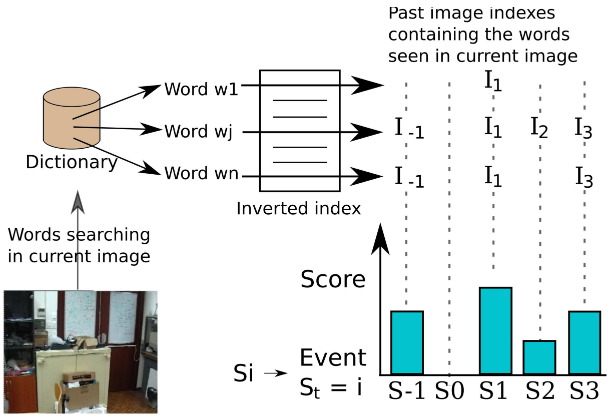

Fig. 2. The voting scheme: the list of the past images in which current words $(z_k)_t$ have been seen is obtained from the inverted index and is used to update the hypotheses' scores.

Once all the words found in the current image have been processed and the computation of the scores is complete, we select the subset $(H_k)_t \subseteq I^{t-p}$ of images for which the *particular* coefficient of variation (c.o.v.) (i.e. particular deviation from the mean of the scores normalized by the mean) is higher than the *standard* c.o.v. (i.e. standard deviation normalized by the mean). $(H_k)_t \subseteq I^{t-p}$ is the subset of the most likely images according to the feature space k. Then, if I_i appears in $(H_k)_t$, the belief at time t (see equation 5) is multiplied by the difference between the particular c.o.v. of I_i and the standard c.o.v., plus 1 (which can be simplified into the difference between the score s_i of the hypothesis and the standard deviation σ, normalized by the mean μ):

$$\mathcal{L}\left(S_t = i | (z_k)_t\right) = \begin{cases} \frac{s_i - \mu}{\mu} - \frac{\sigma}{\mu} + 1 = \frac{s_i - \sigma}{\mu} & \text{if } s_i \geq \mu + \sigma \\ 1 & \text{otherwise} \end{cases} \quad (7)$$

The update of the belief for the restricted set of the most likely hypotheses is illustrated in figure 3. The selection done on the hypotheses at this stage makes it possible to simplify the update of the posterior (as only a restricted set of hypotheses is updated), considering that non-selected hypotheses have a likelihood of 1 and therefore multiply the posterior by 1. When all the images of $(H_k)_t$ have been processed for all the feature spaces, the full posterior is normalized.

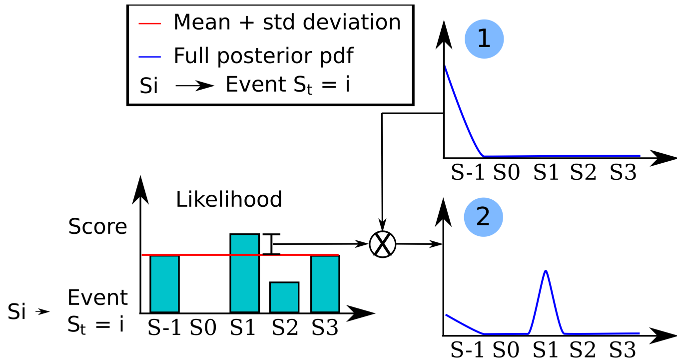

Fig. 3. The belief at time t (frame "1", see equation 5, section IV-A), is updated according to the likelihood model (frame "2"): when the score of a hypothesis is above the mean + standard deviation threshold, the corresponding probability is updated.

D. A Posteriori Hypotheses Management

When the full posterior has been updated and normalized, we search for the hypothesis $S_t = i$ whose a posteriori probability is above some threshold (0.8 in our experiments). However, the posterior does not necessarily exhibit a strong single peak for a unique hypothesis even if a loop-closure occurred. It may rather be diffused over a set of neighboring hypotheses (except for $S_t = -1$). This is mainly imputable to the similarities among neighboring images in time: some of the words commonly found in I_t and I_i are also probably in I_{i-1} or I_{i+1} for example. Thus, instead of searching for single peaks among the full posterior, we look for a hypothesis for which the sum of the probabilities over neighboring hypotheses is above the threshold (the neighborhood chosen here is the same as the neighborhood selected for the diffusion in section IV-B).

When a hypothesis fulfills the above condition, a multiple-view geometry algorithm [11] helps discarding outliers by verifying that the two images of the loop-closure (i.e. I_t and I_i) satisfy the epipolar geometry constraint, which would imply that they share some common structure and that they could hence come from the same 3D scene. To this end, a RANSAC procedure entails rapidly computing several camera transformations by matching SIFT features between the two frames, discarding inconsistent ones using a threshold on the average reprojection error. If successful, the algorithm returns the 3D transformation between x_t and x_i (i.e. the viewpoints associated with I_t and I_i) and the hypothesis is accepted. Otherwise, the hypothesis is discarded. However, even if a hypothesis has been discarded by the multiple-view geometry algorithm, its a posteriori probability will not fall to 0 immediately: it will diffuse over neighboring images during the propagation of the full posterior from t to $t+1$. Thus, correct hypotheses erroneously discarded by epipolar geometry will be reinforced by the likelihoods of further time instants until a valid 3D transformation is found. Note that since SIFT features are extracted from the images and stored during the online dictionary construction, we do not need to process the images again when applying the multiple-view geometry algorithm.

V. Experimental Results

We obtained results[1] from several indoor and outdoor image sequences grabbed with a single monocular handheld camera (i.e. a simple camcorder with a 60° field of view and automatic exposure). In this paper, we present the results obtained from two experiments: an indoor image sequence with strong perceptual aliasing and a long outdoor image sequence. In both experiments, illumination conditions remained constant: the indoor sequence has been captured under artificial lighting conditions, while the length of the outdoor one (i.e. nearly 20 minutes) was too short to experience changes in lighting conditions.

A. Indoor experiment

The overall camera trajectory followed during this experiment is shown in figure 4 using three different styles. When the posterior is below the threshold, the trajectory is shown with a blue (dotted) line. When it is above the threshold and the epipolar constraint is satisfied, a loop-closure is detected and the trajectory is shown with a green (dashed) line. But, when the posterior is above the threshold and the epipolar constraint is not satisfied, the loop-closure hypothesis is rejected and the trajectory is shown with a red (circled) line.

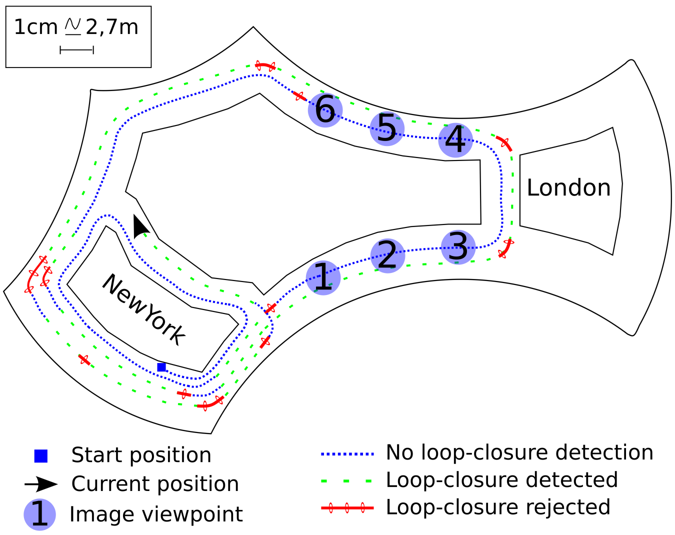

Fig. 4. Overall camera trajectory for the indoor image sequence. A first short loop is done around the "New York" elevators on the left before going to the "London" elevators on the right. The short loop is travelled again when the camera is back from the "London" elevators following the top-most corridor on the plan. Then, the camera repeats the long loop (i.e. to the "London" elevators and back) before ending in front of the "New York" elevators. The numbers in the circles indicate the positions from which the images shown in figure 5 were taken. See text for details about the trajectory.

As we can see in figure 4, the trajectory is shown with a blue (dotted) line every time the camera is discovering unexplored areas, in spite of the strong perceptual aliasing present in the corridors to and from the "London" elevators (see figure 5 for examples of the images composing the sequence). During the run, no *false positive* detections were made (i.e. when a loop-closure is detected whereas none occurred), thus demonstrating the robustness of our solution to perceptual aliasing.

Fig. 5. Top-most corridor (top row) and bottom-most corridor (bottom row) image examples, showing the high level of perceptual aliasing in the environment. The numbers in the circles help associating the images with the positions labelled in figure 4.

From figure 4, we can also see that the trajectory is shown with a green (dashed) line most of the time spent in previously visited places, indicating that *true positive* detections were made (i.e. when a loop-closure occurs and it is correctly detected). Figure 6 gives an example of a true positive detection.

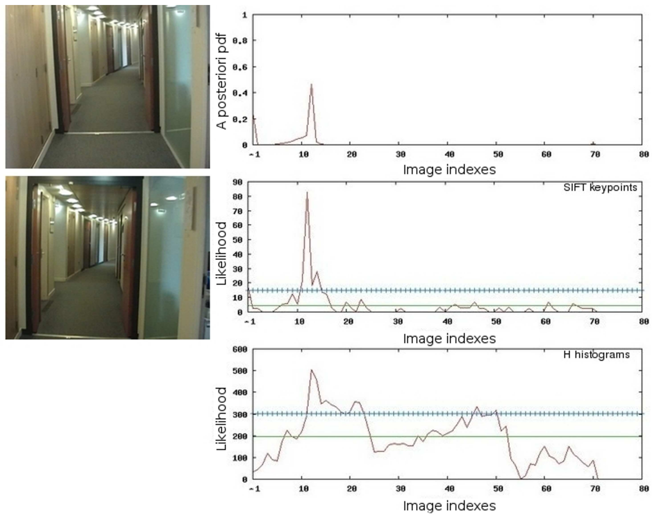

Fig. 6. First loop-closure detection for the indoor image sequence. The full posterior and the likelihood computed from the SIFT and H histograms feature spaces are shown, along with the current image I_t (top left) and the loop-closing image I_i (bottom left). Likelihoods are obtained from the scores (tf–idf) of the different hypotheses. Also shown with the likelihoods are the score mean (solid green) and the score mean + standard deviation threshold (blue crosses). As it can be seen, the likelihood is very strong around images corresponding to hypotheses 10 to 13, causing the sum of the corresponding probabilities in the posterior to reach the 0.8 threshold. Also, it clearly appears here that I_t and I_i come from very close viewpoints.

During passings in already explored places, it may be noticed that the line representing the trajectory switches from green (dashed) to red (circled) each time the camera was

[1] Videos available at http://animatlab.lip6.fr/AngeliVideosEn, but also at http://ieeexplore.ieee.org as supplemental material to this paper.

turning around corners. In these particular cases, loop-closure detection fails only because the epipolar constraint is not satisfied: the a posteriori probability of loop-closure is above the threshold but, due to the large and fast rotations made by the camera, precise keypoints associations are difficult. Indeed, in this narrow indoor environment, when the camera is turning around corners, the viewpoint variation between current and loop-closing images may be large, resulting in small overlap between these images and preventing SIFT features from matching correctly. This corresponds to *false negative* detections (i.e. when a loop-closure occurs but it is not detected).

When considering the trajectory of the camera with more attention, it may be observed that the first loop-closure detection that should be done (i.e. when the camera reaches again its starting position for the first time, during its first travel behind the "New York" elevators) is missed and the trajectory remains shown with a blue (dotted) line. This is imputable to the low responsiveness of the probabilistic framework: the likelihood associated with a particular hypothesis has to be very high relative to the other likelihoods to trigger a fast loop-closure detection. Usually, the likelihood associated with a hypothesis must have a good support during 2 or 3 consecutive images in order to trigger a loop-closure detection. The responsiveness of our system is governed by the transition model of the probabilistic framework: we assume that the probability of remaining in a "no loop-closure" event is high (i.e. 0.9, see section IV-B). Decreasing this probability to lower values makes it possible to detect loop-closures faster (i.e. with fewer images required), but this also produces false positive detections, which is not acceptable. The delay involved here therefore enhances the robustness to transient detection errors, considering only hypotheses with repeated support over time as possible candidates for loop-closure.

During the run, there was only one case where the probability was above the threshold but the selected hypothesis was wrong and it has been conveniently rejected by the multiple-view geometry algorithm. This event, that can be considered as a *false alarm*, can be identified in figure 4 as the red (circled) portion of the trajectory that occurs when the camera is coming back for the first time from the "London" elevators (just near the 6^{th} circle on the figure). This false alarm can be explained by the strong perceptual aliasing that makes the corridors to and from the "London" elevators look the same (see figure 7): since our bag-of-words algorithm relies on the occurrence of the words rather than on their position, the current image may look like a past image but the structures of the scenes may not be consistent, thus preventing the epipolar constraint from being satisfied.

In order to test the robustness of the detection to camera viewpoint changes, we rotated the camera around its optical axis when passing behind the "New York" elevators for the second and third times. As shown by the green (dashed) line representing the trajectory during these periods, the loop-closure detection results were not affected. The figure 8 gives an example of loop-closure detection with different camera orientations between current and loop-closing images. The loop-closure detection shown in this figure corresponds to the third passing of the camera behind the "New York" elevators. This is why we observe two distinct peaks on the likelihoods: two hypotheses are valid in this case, because I_t closes the loop with images from the first and the second visits. But due to the temporal coherency of the p.d.f., the hypotheses that have high a posteriori probabilities are those from the second passing.

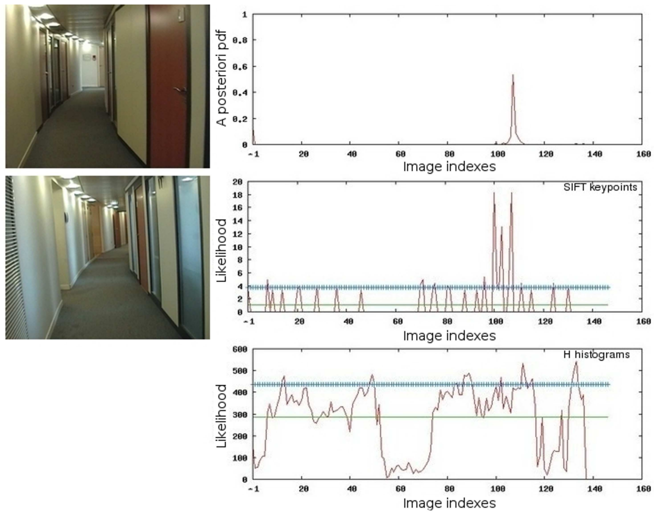

Fig. 7. The only false alarm due to perceptual aliasing: as we can see, the likelihoods are confused (we can note two similar high peaks on the SIFT's likelihood, while the H histograms' likelihood does not give helpful information) and the images look very similar. This hypothesis has been rejected by the multiple-view geometry algorithm.

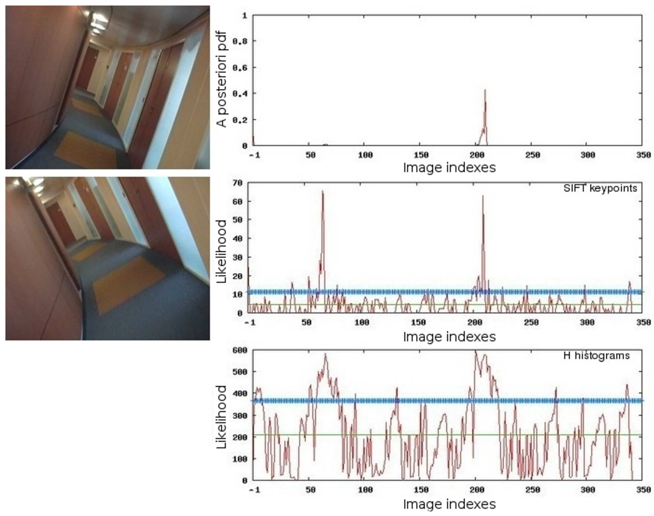

Fig. 8. Another loop-closure detection for the indoor image sequence. Although there is a significant camera viewpoint difference between current and past images, the loop-closure is correctly detected.

B. Outdoor experiment

During this second experiment, images were taken outdoor with a handheld camera while turning around the laboratory's building (figure 9 gives examples of images from this sequence).

Fig. 9. Examples of the images composing the outdoor sequence. The numbers in the circles help associating the images with the positions labelled in the figure 10.

The overall camera trajectory followed during this experiment is shown in figure 10 using the same style conventions as before. Here, we introduced red-green (circled-dashed) lines to denote fast alternations of true positive and false negative detections that occur when people or cars are passing in front of the camera, causing correct hypotheses to be rejected because not enough point correspondences can be found to satisfy the epipolar geometry constraint. These events (of which one example is given in figure 11) demonstrate the robustness of the probabilistic framework to transient detection errors: even though images are occluded by people or cars, correct loop-closure hypotheses are selected (i.e. they have a high a posteriori probability), but since the epipolar constraint cannot be satisfied, they cannot be fully validated to be accepted as true positive loop-closure detections.

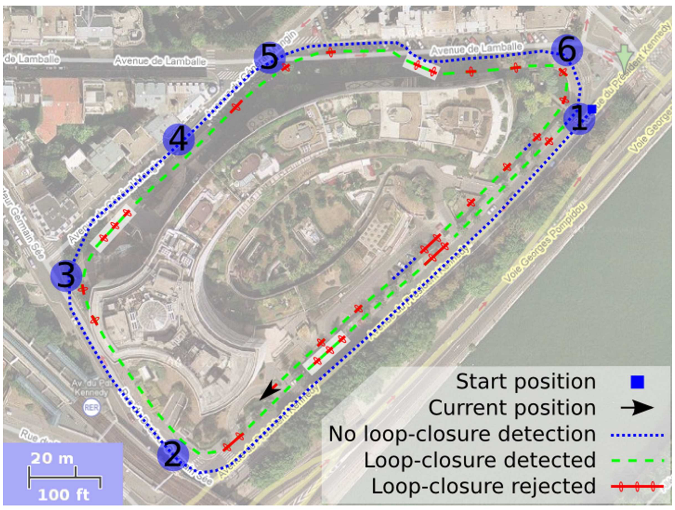

Fig. 10. Overall camera trajectory for the outdoor image sequence. Two loops are done around the "Lip6" laboratory, starting near the top-right end of the building on the image (indicated by the square) and ending at its bottom-left end. The path in front of the building (i.e. running parallel to the river) is thus travelled three times. The style conventions for the trajectory are the same as in figure 4, with the introduction here of red-green (circled-dashed) lines to denote fast alternations of true positive and false negative detections. Red-green (circled-dashed) lines are painted over white rectangles to distinguish them easily. See text for details about the trajectory.

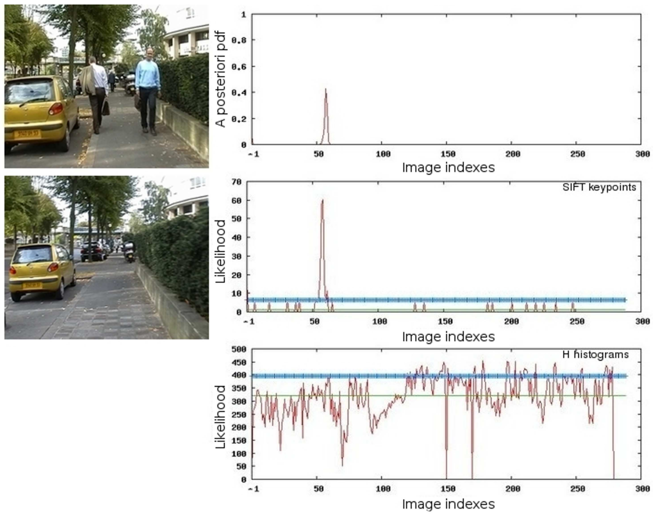

Fig. 11. Robustness of the probabilistic framework to transient detection errors: although current image is partially occluded by pedestrians, a correct loop-closure hypothesis is selected, but it is rejected by the multiple view geometry algorithm.

As in the indoor experiment, no false positive detections were made, whereas multiple true positives were found (see figure 12). Also, we can see from figure 10 that the first loop-closure detections occur tardily when the camera is coming back to its starting position, revealing again the low responsiveness of the probabilistic framework.

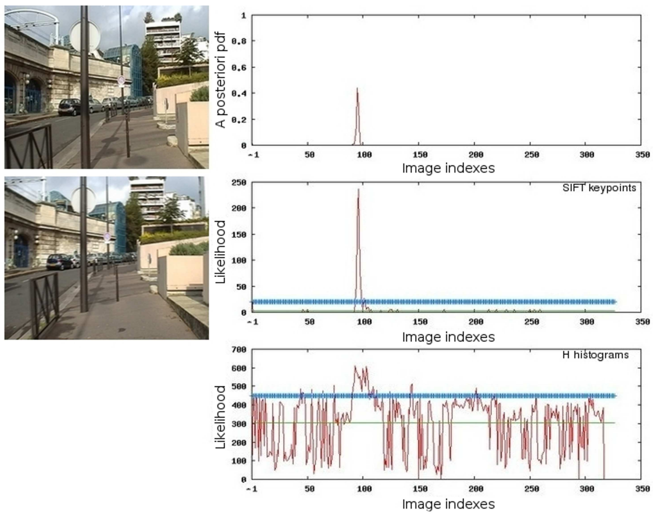

Fig. 12. Example of a true positive loop-closure detection for the outdoor image sequence. Again, we can observe that the likelihood from the SIFT feature space is very high and discriminative.

C. Influence of the visual dictionaries

In this section we will study the influence of the different visual dictionaries used here (i.e. SIFT features and H histograms) for loop-closure detection. To this end, we tried to perform loop-closure detection using only either SIFT features or H histograms. Although those tests have been done using both image sequences, the indoor one produces more valuable results since more loop-closures are done during the travel of the camera and because the indoor environment is much more diversified.

H histograms only carry colorimetric information, without any shape nor structure information. Therefore, the corresponding likelihood is always confused, and it will never be very peaked over one particular hypothesis unless the corresponding image contains specific colors that are seen nowhere else. However, H histograms can help distinguishing similarly structured environments that only differ in their colors (e.g. two corridors having the same dimensions but whose walls are painted with different colors). When used alone, H histograms cannot trigger a loop-closure detection. But when used in combination with SIFT features, they enhance loop-closure detection, improving notably the overall responsiveness of the probabilistic framework. Indeed, as shown in figure 13, we can see that the posterior obtained when using both SIFT features and H histograms is higher than when using SIFT features only. This is because H histograms' likelihood, although not discriminative enough to trigger a loop-closure detection, is a higher around the loop-closing hypothesis, and so it reinforces the votes from the SIFT feature space when updating the posterior.

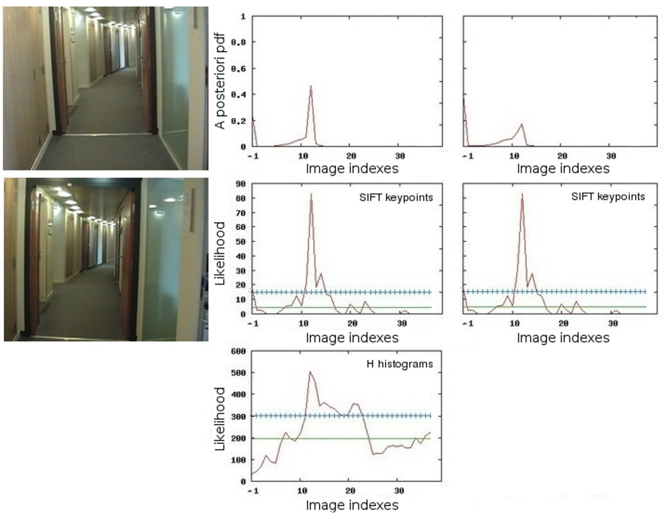

Fig. 13. Loop-closure detection enhancement using color and shape information in the indoor image sequence: when H histograms are combined to SIFT features (left part), the a posteriori probability is higher than when using SIFT features alone (right part).

Using SIFT features in conjunction with H histograms therefore enhances the responsiveness of the algorithm, making it able to detect loop-closures sooner, especially when the camera is back to its starting position for the first time: loop-closures are detected 2 or 3 images before when both feature spaces are involved. Table I gives additional clues for this improvement, with information about the loop-closure detection performances for the indoor and outdoor image sequences when using SIFT features alone or in conjunction with H histograms. Given are the number of images composing each sequence ("#img"), the corresponding number of loop-closures ("#LC", determined at hand from the camera trajectory), the rate of true positive detections ("%TP", the percentage of loop-closures correctly detected), and the number of false alarms ("#FA", erroneous hypotheses that receive a high probability but that are rejected by the multiple-view geometry algorithm).

TABLE I
COLOR INFORMATION IMPROVEMENTS

Sequence	#img	#LC	%TP	#FA
Indoor SIFT + H	388	217	80	1
Indoor SIFT	388	217	68	0
Outdoor SIFT + H	531	301	71	0
Outdoor SIFT	531	301	70	0

From table I, we can see that when adding color information, the true positive rate is improved: this is notably remarkable in the indoor sequence where the increase in recognition performances is 12%. On the outdoor sequence on the other hand, improvements are less significant. This is due to the impressive reliability of the SIFT features in this sequence. Indeed, as SIFT features are robust to scale variations in the images, the important depth of the outdoor scenes enables long term recognition of these features along the trajectory of the camera. Hence, adding color information in this case does not dramatically improve the number of correct loop-closure detections. We can also see in table I that adding color information has the unwanted effect of producing more false alarms: when using SIFT features only, no false alarms were raised for the indoor image sequence, whereas one was when combining them with H histograms (see section V-A).

D. *Performances*

During the experiments, the dictionaries were built online in an incremental fashion from images of size 240x192 pixels, enabling real-time performances with a Pentium Core2 Duo 2.33GHz laptop in both indoor and outdoor experiments.

Table II gives the length of the different sequences tested (with corresponding number of images), the CPU time needed to process them, and the sizes of the different dictionaries at the end of the run (expressed in number of words). For both sequences (i.e. indoor and outdoor), we give the performances obtained when SIFT features are used alone or in combination with H histograms.

TABLE II
PERFORMANCES

Sequence	Length	#img	CPU	#SIFT	#H hist.
Indoor SIFT + H	6m28s	388	2m52s	9201	7284
Indoor SIFT	6m28s	388	1m33s	9201	0
Outdoor SIFT + H	17m42s	531	10m16s	39175	18408
Outdoor SIFT	17m42s	531	6m48s	39175	0

For the indoor experiment, images were grabbed at 1Hz: the camera was moved along medium sized corridors, with curved shape and suddenly appearing corners, motivating the choice for a reasonable framerate in order for consecutive images to share some similarities. For the outdoor experiment however, images were grabbed with a lower framerate (i.e. 0.5Hz): outdoor images grabbed at distant time instants share some similarities because of the high depth of outdoor scenes.

From table II, we logically observe that when using SIFT features only, the CPU time needed to process a sequence is significantly lower than when H histograms are involved too: the overall processing is about 40% faster in the first case. However, with both feature spaces enabled, real-time processing is still achieved and, as mentioned before, the responsiveness of the probabilistic framework is enhanced, without causing false positive detections to appear. When processing an image, the most time consuming step is feature extraction and matching with the words of the corresponding dictionary. When trying to match a feature with the visual words of the dictionary, the search is done with logarithmic-time complexity in the number of words due to the tree structure of the dictionary [26]: real-time performances could not have been obtained with linear-time complexity in the number of words in view of the dictionary sizes involved here.

For the outdoor experiment, the overall camera trajectory was about 1.3km and a bit less than 40000 words were created (when considering the SIFT case only) from 531 images. In the results obtained by the authors of [23], the data collection for dictionary construction has been done over 30km, using 3000 images and generating approximately 35000 words. It is obvious that our model needs far more words than the solution proposed in [23], and the intuitive explanation of this is twofold. First, in our online dictionary construction, we cannot afford data rearranging, which would make it possible to obtain a more compact representation. Secondly, in order for the tf–idf weighting used here to perform efficiently, discriminative words are preferable in order to select unambiguous hypotheses. As shown in [10], the size of the cluster representing the words has a direct influence on the word's distinctiveness: a higher distinctiveness is obtained with a smaller cluster size, i.e. a larger dictionary size. The parameters used here are found experimentally to perform well on all the encountered environments.

VI. Discussion and Future Work

The solution proposed in this paper is a completely incremental and online vision-based method allowing loop-closure detection in real-time. The bag-of-words framework introduced in [10] and used here provides a simple way to manage multiple image representations, taking advantage of information gathered from distinct heterogeneous feature spaces. Moreover, building the dictionaries in an incremental fashion entails "learning" only that part of the environment in which the robot is operating, while bag-of-words methods applied to robotics usually use a static dictionary (e.g. [20], [21], [23]) learned beforehand from a training data set supposed to be a good representation of the environment. The consequence is that our system is able to work indoor and outdoor without hand-tuning the dictionary, and without prior information on the environment type.

The results presented here show the robustness of our solution to perceptual aliasing. However, the more complex probabilistic framework described in [23] handles it more properly, taking it into account at the word level (i.e., the input information level) while, in our case, it is managed at the detection level (i.e., the output level), when hypotheses are checked by the epipolar geometry algorithm. Still, the evaluation of the distinctiveness of every word proposed in [23] cannot be done incrementally because, to evaluate the co-occurrences of the words, representative images of the entire environment have to be processed beforehand. In our method, the distinctiveness of the words is taken into account using the online calculated tf–idf coefficient: the words seen multiple times in the same image will vote with a high score for this image (i.e. high tf), while the words seen in every images will have a small contribution (i.e. low idf).

The probabilistic framework presented here poorly handles the management of loop-closure hypotheses. Indeed, a new entry is added to the posterior each time a new image is acquired, while the evaluation of the corresponding hypotheses (i.e. checking if whether or not the newly acquired image closes the loop with one of the past images) is done afterwards: in other words, a new image is added to the model irrespectively of the loop-closure detection results. In future work, a topological map of the environment could be directly created by adding only images that do not close a loop with already memorized ones. These events would therefore represent positions in the environment, linked by their proximity in time and space, and not only images linked sequentially in time. This would avoid the presence of multiple high peaks due to the co-existence of multiple images taken from the same position (see figure 8).

In future work, we will adapt our approach to a purely vision-based SLAM system like [6] so as to reinitialize the SLAM algorithm when the camera position is lost or when there is a need to self-localize in a map acquired beforehand. The metrical information about the camera's pose coming from SLAM could help improving the definition of a location's neighborhood, using spatial transitions between adjacent locations instead of time indexes. As mentioned above, this would make it possible to fuse images taken from close metric locations to build a topological map of the environment.

Finally, other feature spaces could be explored, implementing for instance one of the visual descriptors tested in [25], whereas relative spatial positions between the visual words could be used to improve matching. Loop-closure detection at different moments of the day should also be experienced, so as to test the robustness of our solution to varying lighting conditions.

VII. Conclusion

In this paper, we have presented a fast and incremental bag-of-words method for performing loop-closure detection in real-time, with no false positive detections on the obtained experimental results even under strong perceptual aliasing conditions. We demonstrated the quality of our approach with results obtained in indoor and outdoor environments, reaching real-time performances even in long image sequences. Our approach calls upon a Bayesian filtering framework with likelihood computation in a simple voting scheme and should be extended to SLAM reinitialization in a near future.

VIII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the reviewers for their useful comments on reviewing the paper.

REFERENCES

[1] D. Filliat and J.-A. Meyer, "Map-based navigation in mobile robots - I. a review of localisation strategies," *Journal of Cognitive Systems Research*, vol. 4, no. 4, pp. 243–282, 2003.

[2] J.-A. Meyer and D. Filliat, "Map-based navigation in mobile robots - II. a review of map-learning and path-planing strategies," *Journal of Cognitive Systems Research*, vol. 4, no. 4, pp. 283–317, 2003.

[3] H. Durrant-Whyte and T. Bailey, "Simultaneous localisation and mapping (slam): Part i," *IEEE Robotics and Automation Magazine*, vol. 13, no. 1, pp. 99–110, 2006.

[4] T. Bailey and H. Durrant-Whyte, "Simultaneous localisation and mapping (slam): Part ii," *IEEE Robotics and Automation Magazine*, vol. 13, no. 3, pp. 108–117, 2006.

[5] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "2d simultaneous localization and mapping for micro aerial vehicles," in *European Micro Aerial Vehicles (EMAV)*, 2006.

[6] A. Davison, I. Reid, N. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, June 2007.

[7] T. Lemaire, C. Berger, I.-K. Jung, and S. Lacroix, "Vision-based slam: Stereo and monocular approaches," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 343–364, February 2007.

[8] M. E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *IEEE conference on Computer Vision and Pattern Recognition*, 2006.

[9] G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV04 workshop on Statistical Learning in Computer Vision*, 2004, pp. 59–74.

[10] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," in *IEEE International Conference on Robotics and Automation*, 2007.

[11] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 756–777, 2004.

[12] F. Dellaert, D. Fox, W. Burgard, and S. Thrun, "Monte carlo localization for mobile robots," in *IEEE International Conference on Robotics and Automation*, May 1999.

[13] J. Wolf, W. Burgard, and H. Burkhardt, "Robust vision-based localization by combining an image retrieval system with monte carlo localization," *IEEE Transactions on Robotics*, vol. 21, no. 2, pp. 208–216, 2005.

[14] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges," in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI)*. Acapulco, Mexico: IJCAI, 2003.

[15] M. Pupilli and A. Calway, "Real-time visual slam with resilience to erratic motion," in *IEEE Computer Vision and Pattern Recognition*, 2006.

[16] L. Clemente, A. Davison, I. Reid, J. Neira, and J. Tardòs, "Mapping large loops with a single hand-held camera," in *Proceedings of Robotics: Science and Systems*, 2007.

[17] B. Williams, P. Smith, and I. Reid, "Automatic relocalisation for a single-camera simultaneous localisation and mapping system," in *IEEE Iternational Conference on Robotics and Automation (ICRA)*, 2007.

[18] J. Kosecká, F. Li, and X. Yang, "Global localization and relative positioning based on scale-invariant keypoints," *Robotics and Autonomous Systems*, vol. 52, pp. 209–228, 2005.

[19] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *IEEE International Conference on Robotics and Automation*, 2000.

[20] J. Wang, H. Zha, and R. Cipolla, "Coarse-to-fine vision-based localization by indexing scale-invariant features," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 36, no. 2, pp. 413–422, April 2006.

[21] P. Newman, D. Cole, and K. Ho, "Outdoor slam using visual appearance and laser ranging," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2006.

[22] K. L. Ho and P. Newman, "Detecting loop closure with scene sequences," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 261–286, 2007.

[23] M. Cummins and P. Newman, "Probabilistic appearance based navigation and loop closing," in *Proc. IEEE International Conference on Robotics and Automation (ICRA'07)*, 2007.

[24] D. Lowe, "Distinctive image feature from scale-invariant keypoint," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in *International Conference on Computer Vision & Pattern Recognition*, vol. 2, June 2003, pp. 257–263.

[26] D. Filliat, "Interactive learning of visual topological navigation," in *To appear in the proceedings of the 2008 IEEE International Conference on Intelligent Robots and Systems (IROS 2008)*, 2008.

[27] H. Ling and K. Okada, "Diffusion distance for histogram comparison," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2006, pp. 246–253.

[28] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision (ICCV)*, 2003.

Adrien Angeli received in 2005 the Master's degree in computer engineering from the Ecole Centrale d'Electronique, Paris, France and the Master's degree in artificial intelligence from the Université Pierre et Marie Curie, Paris, France. He is currently a Ph.D. student with the Université Pierre et Marie Curie. His research interests include vision-based localization and SLAM applications to robotics.

David Filliat graduated from the Ecole Polytechnique in 1997 and obtained a PhD in robotics from Paris VI university in 2001. After 3 years of work for the robotic programs in the French armament procurement agency, he is now assistant professor at Ecole Nationale Supérieure de Techniques Avancées. His main research interest are perception, navigation and learning in the frame of the developmental approach to autonomous mobile robotics.

Stéphane Doncieux is Assistant Professor at Paris 6 University and is responsible for the SIMA (Integrated, Mobile and Autonomous Systems) research team of ISIR. He has been trained as an engineer and holds a PhD in computer science. He works on the autonomous design of control architectures thanks to evolutionary algorithms and on adding decisional autonomy to flying robots. He is also heading the ROBUR project of ISIR, which aims at building an autonomous flapping-wing robot.

Jean-Arcady Meyer is Emeritus Research Director at CNRS. He trained as an engineer, he graduated in Human and Animal Psychology, and received a PhD in Biology. He is the founder of the journal Adaptive Behavior, a former Director of the International Society for Adaptive Behavior and a current Director of The International Society for Artificial Life. His primary scientific interests are in adaptive behaviors in natural and artificial systems. He is the main coordinator of the Psikharpax project.