A Learning Algorithm for Place Recognition

César Cadena and José Neira

Abstract— **We present a place recognition algorithm for SLAM systems using stereo cameras that considers both appearance and geometric information. Both near and far scene points provide information for the recognition process. Hypotheses about loop closings are generated using a fast appearance technique based on the bag-of-words (BoW) method. Loop closing candidates are evaluated in the context of recent images in the sequence. In cases where similarity is not sufficiently clear, loop closing verification is carried out using a method based on Conditional Random Fields (CRFs). We compare our system with the state of the art using visual indoor and outdoor data from the RAWSEEDS project, and a multisession outdoor dataset obtained at the MIT campus. Our system achieves higher recall (less false negatives) for full precision (no false positives), as compared with the state of the art. It is also more robust to changes in appearance of places because of changes in illumination (different shadow configurations in different days or time of day). We discuss the promise of learning algorithms such as ours, where learning can be modified on-line to re-evaluate the knowledge that the system has about a changing environment.**

I. Introduction

In this paper, we consider the problem of recognising locations based on scene geometry and appearance. This problem is particularly relevant in the context of environment modeling and navigation in mobile robotics. Algorithms based on visual appearance are becoming popular to detect locations already visited, also known as loop closures, because cameras are inexpensive, lightweight and provide rich scene detail.

Many methods focus on place recognition, and mainly use the bag-of-words representation [1], supported by some probabilistic framework [2]. Techniques derived from BoW have been successfully applied to loop closing-related problems [2,3], but could exhibit false positives in cases when the same features are detected although in a different geometric configuration. An incorrect loop closure can result in a critical failure for SLAM algorithms. On the issue of recognition of revisited places we consider the state of the art is the FAB-MAP [3]. It has proved very high precision although with reduced recall. In other words, the low proportion of false positives is attained by sacrificing many true positives [4]. This system also presents problems in applications using front facing cameras [5].

To avoid mismatches in these appearance-based approaches, some geometrical constraint is generally added in a verification step. The epipolar geometry is the most common technique used to enforce consistency in matches [2,4]. Recently proposed by Paul and Newman [6], the FAB-MAP 3D uses the same FAB-MAP framework including 3D information provided by a laser scanner for the distances between features, but can only make such inferences about visual features in the laser range.

The algorithm for place recognition presented here, first proposed in [7] and improved in [8], uses two complementary techniques. Candidate loop closing locations are generated using an improved bag-of-words method (BoW) [1], which reduces images to sparse numerical vectors by quantising their local features. This enables quick comparisons among a set of images to find those which are similar. Hierarchical implementation improves efficiency [9]. Hypothesis verification is based on CRF-Matching. CRF-Matching is a probabilistic model able to jointly reason about the association of features. It was proposed initially for 2D laser scans [10] and monocular images [11]. The CRF infers over the scene's image and 3D geometry. Insufficiently clear loop closure candidates from the first stage are verified by matching the scenes with CRFs. This algorithm takes advantage of the sequential nature of the data in order to establish a suitable metric for comparison of candidates, resulting in a highly reliable detector for revisited places.

The basic idea of the algorithm is to exploit the efficiency of BoW for detecting revisited places in real-time and the higher robustness of CRF-Matching to ensure that revisiting matches are correct. In order to keep the real time execution, only insufficiently clear results of BoW are input to CRF-Matching.

In the next section we provide a description on our place recognition system. In section III we present experimental results on real data that demonstrate the improvement in robustness and reliability of our approach. Finally, in section IV we discuss the results and discuss the applicability of our system to operations over time.

II. The Place Recognition System

The place recognition system [7,8] is a loop closing candidate generation-verification scheme. In this section, we describe both components of the system.

A. Loop Candidates Detection

The first component is based on the bag-of-words method (BoW) of [1] which is implemented in a hierarchical way, thus improving efficiency [9]. In this implementation we use 64-SURF-features, see Fig. 1(a). λ_t is the BoW score computed between the current image and the previous one. The minimum confidence expected for a loop closure candidate is

César Cadena and José Neira are with the Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Zaragoza 50018, Spain. {ccadena, jneira}@unizar.es.

This research has been funded by the Dirección General de Investigación of Spain under projects DPI2009-13710, DPI2009-07130.

α^-; the confidence for a loop closure to be accepted without further verification is α^+. The images from one session are added to the database at one frame per second. This implementation enables quick comparisons of one image at time t with a database of images in order to find those that are similar according to the score s. There are 3 possibilities:

1) if $s \geq \alpha^+\lambda_t$ the match is considered highly reliable and accepted;
2) if $\alpha^-\lambda_t < s < \alpha^+\lambda_t$ the match is checked by CRF-Matching in the next step of verification.
3) otherwise. the match is ignored.

B. Loop Closure Verification

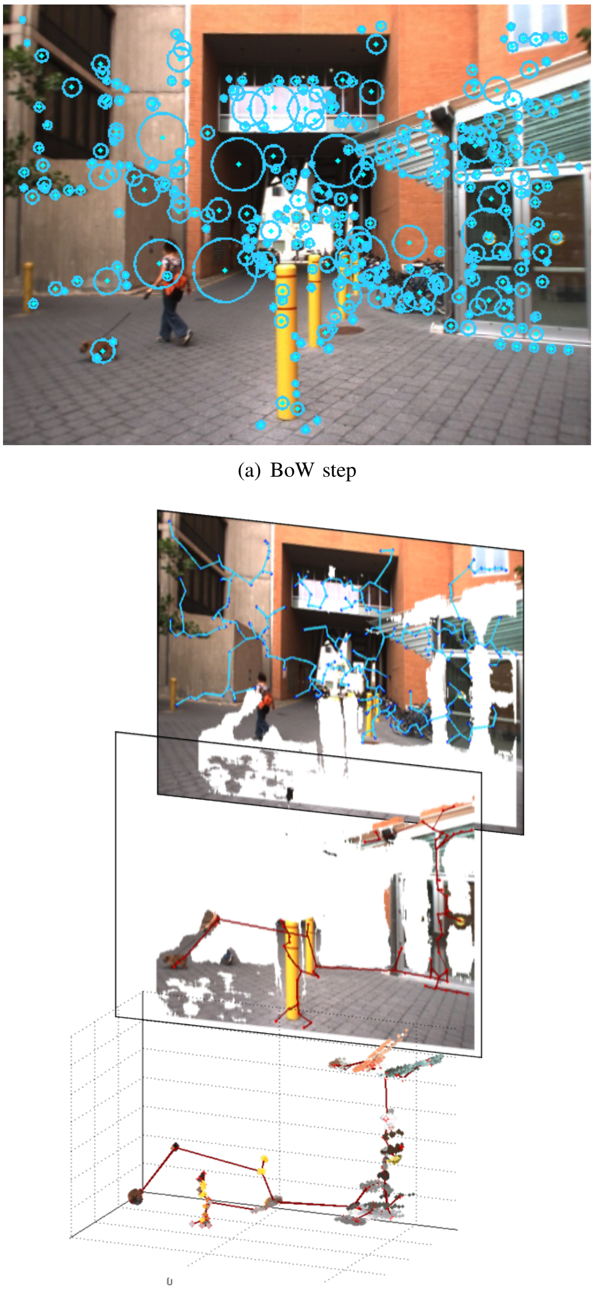
When further verification is required, loop closing candidates are verified for consistency in 3D and in image space with CRF-Matching, an algorithm based on Conditional Random Fields (CRF) [12]. CRF-Matching is a probabilistic graphical model for reasoning about joint association between features of different scenes. We model the scene with two graphs, the first one for SURF-features with 3D information (near), and the second one over the remaining SURF-features (far), see Fig. 1(b). The graph structure is given for the minimum spanning tree over the euclidean distances, either the 3D metric coordinates ($\mathcal{G}_{3D}$) or 2D pixel coordinates ($\mathcal{G}_{Im}$). We use the CRF-Matching stage over the loop closing candidates provided by the BoW stage. Then, we compute the negative log-likelihood (Λ) from the MAP associations between the scene in time t, against the loop closing candidate in time t', $\Lambda_{t,t'}$, and the scene in $t-1$, $\Lambda_{t,t-1}$.

The negative log-likelihood Λ^{3D} of the MAP association for $\mathcal{G}_{3D}$ provides a measure of how similar two scenes are in terms of close range, and Λ^{Im} for $\mathcal{G}_{Im}$ in terms of far range. Thus, we compare how similar the current scene is with the scene in t' with respect to $t-1$ with $\Lambda_{t,t'} \leq \beta\Lambda_{t,t-1}$ for each graph. With the β parameters we can control the level we demand of similarity to $(t, t-1)$, a low β means a high demand. By choosing different parameters for near and far information we can make a balance between the weight of each in our acceptance. Our place recognition system can be summarized in the algorithm 1.

III. Experiments

We have evaluated our system with the public datasets from the RAWSEEDS Project [13]. The data were collected by a robotic platform in static and dynamic indoor, outdoor and mixed environments. We have used the data corresponding to the Stereo Vision System with an 18cm baseline. Images are (640x480 px) taken at 15 fps.

We used 200 images uniformly distributed in time, from a static mixed dataset taken on 01-Sep-2008, for training the vocabulary for BoW and for learning the weights for CRF-Matching. In order to learn the weights for the CRF-Matching, we obtained the SURF features from the right image in the stereo system and computed their 3D coordinates. Then, we ran a RANSAC algorithm over the rigid-body transformation between the scene at time t and the scene at time $t-\delta_t$. The results from RANSAC were our labels. Since the stereo system has high noise in the dense 3D information, we selected $\delta_t = 1/15s$. The same procedure is done over the SURF features with no 3D information, where we obtain


(a) BoW step

(b) CRF step

Fig. 1. Outdoor scene from the MIT campus. We get the SURF-features for the BoW stage over one image of the stereo pair 1(a). For the CRF stage we compute the two minimum spanning trees (MST), one for features with 3D information (near features), and the second for the remaining ones, with image information (far features). In 1(b), we show the two resulting graphs: in blue the graph for far features ($\mathcal{G}_{Im}$), in dark red the graph for near features ($\mathcal{G}_{3D}$). We apply CRF-Matching over both graphs. The minimum spanning tree of $\mathcal{G}_{3D}$ is computed according to the metric coordinates, projected over the middle image only for visualisation. In the bottom, we show $\mathcal{G}_{3D}$ in metric coordinates with the 3D point cloud (textured) of each vertex in the tree. The MST gives us an idea of the dependencies between features in a scene, and allows for robust consistency checks of feature associations between scenes.

TABLE I

PARAMETERS FOR THE EXPERIMENTS

	FAB-MAP 2.0					Our System			
	RAWSEEDS			MIT		RAWSEEDS			MIT
	Indoor	Outdoor	Mixed	Campus		Indoor	Outdoor	Mixed	Campus
p	50%	96%	62%	33%	α^{+}	60%	60%	60%	60%
$P(\text{obs}\|\text{exist})$	0.31	0.39	0.37	0.39	α^{-}	15%	15%	15%	15%
$P(\text{obs}\|!\text{exist})$	0.05	0.05	0.05	0.05	β_{3D}	1	1.5	1.5	1.5
Motion Model	0.8	0.8	0.6	0.6	β_{Im}	1.3	1.7	1.7	1.7

Algorithm 1 Pseudo-algorithm of our place recognition system

```
Input: Scene at time $t$, Database $\langle 1, \ldots, t-1 \rangle$
Output: Time $t'$ of the revisited place, or null
  $Output = Null$
  Find the best score $s_{t,t'}$ from the query in the database of the bag-of words
  if $s_{t,t'} \geq \alpha^{+} s_{t,t-1}$ then
    $Output = t'$
  else
    if $s_{t,t'} \geq \alpha^{-} s_{t,t-1}$ then
      Build the $\mathcal{G}_{3D}$ and $\mathcal{G}_{Im}$
      Infer with CRFs and compute the neg-log-likelihoods $\Lambda$
      if $\Lambda^{3D}_{t,t'} \leq \beta_{3D}\Lambda^{3D}_{t,t-1} \wedge \Lambda^{Im}_{t,t'} \leq \beta_{Im}\Lambda^{Im}_{t,t-1}$ then
        $Output = t'$
      end if
    end if
  end if
  Add current scene to the Database
```

the labels by calculating with RANSAC the fundamental matrix between the images. Thus, we obtained a reliable enough labelling for the training. Although this automatic labelling can return some outliers, the learning algorithm has demonstrated being robust in their presence. Afterwards, we tested the whole system in three other datasets: static indoor, static outdoor and dynamic mixed. The four datasets were collected on different dates and in two different campuses. Refer to the RAWSEEDS Project [13] for more details. In the fig. 2 we show the ground truth trajectories and results.

For the first bag-of-words stage, we have to set the minimum confidence expected for a loop closure candidate, α^{-}, and the minimum confidence for a trusted loop closure, α^{+}. We selected the working values $\alpha^{-} = 15\%$ and $\alpha^{+} = 60\%$ in all experiments. Since these datasets are fairly heterogeneous, we think these values can work well in many situations. As It might depend on the datasets and the vocabulary size, though. Then, for the CRF-Matching stage, we set the β parameters in order to obtain 100% precision. This allows comparisons with alternative systems in terms of reliability. All the parameters used are shown in Table I.

We have compared the results from our system against the state-of-the-art technique FAB-MAP 2.0 [4]. The FAB-MAP software[1] provides some predefined vocabularies. We have used the FAB-MAP indoor vocabulary for the RAWSEEDS indoor dataset and the FAB-MAP outdoor vocabulary for the others datasets. This technique has a set of parameters to tune in order to obtain the best performance in each experiment. The parameters that we have modified are the following ones (for further description please see [3] and [4]):

- p: Probability threshold. The minimum matching probability required to accept that two images were generated from the same place.
- $P(\text{obs}|\text{exist})$: True positive rate of the sensor. Prior probability for detecting a feature given that it exists in the location.
- $P(\text{obs}|!\text{exist})$: False positive rate of the sensor. Prior probability for detecting a feature given that it does not exist in the location.
- *Motion Model*: Model Motion Prior. This biases the matching probabilities according to the expected motion of the robot. A value of 1.0 means that all the probability mass goes forward, and 0.5, means that probability goes equally forward and backward.

In both systems, our and FAB-MAP, we disallow the matches with frames in the previous 20 seconds. The final values used by us are shown in Table I. We have chosen the parameter set in order to obtain the maximum possible recall at one hundred percent precision. All the place recognition experiments are carried out at 1 fps.

TABLE II

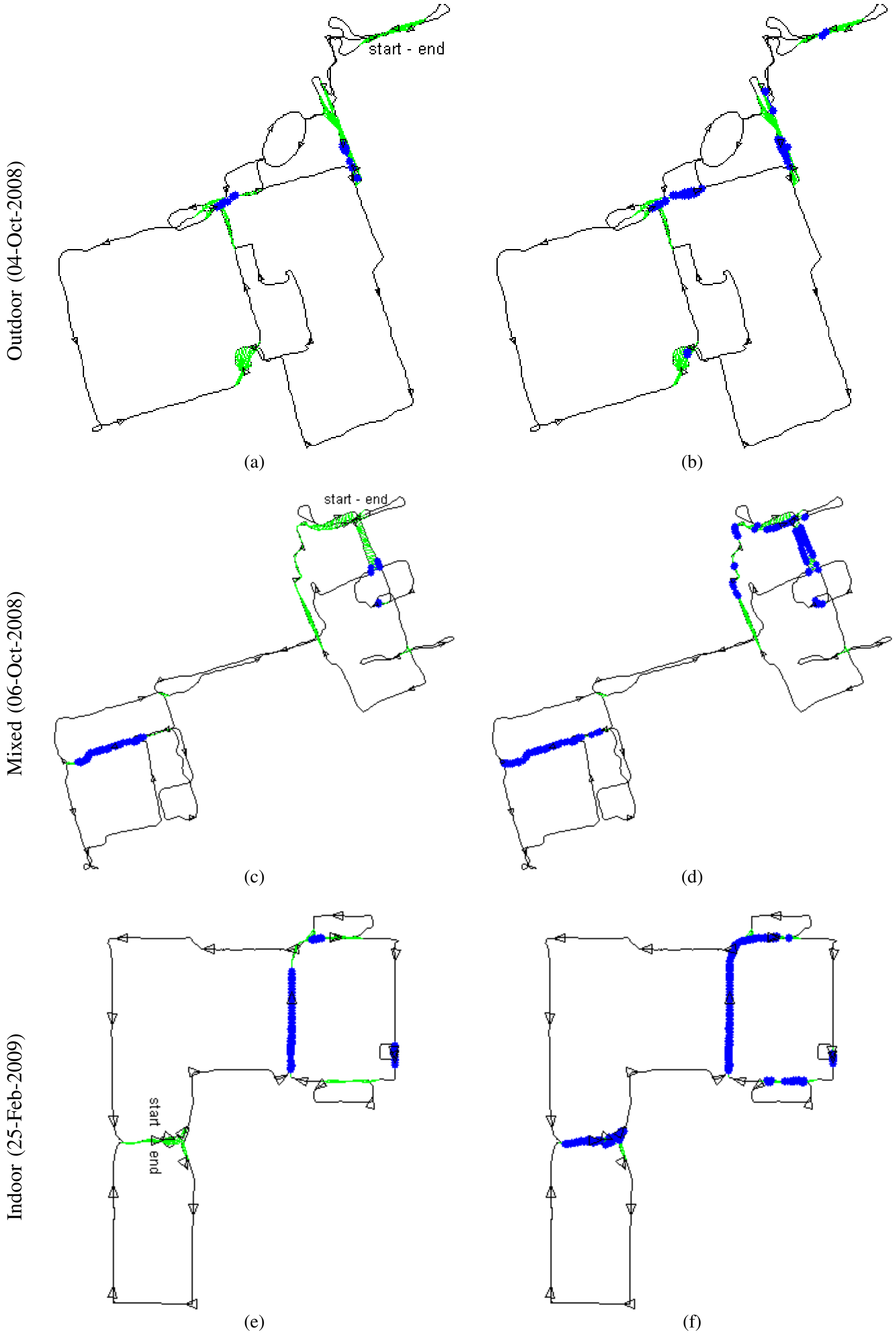
RESULTS FOR ALL DATASETS

	Precision	Recall
RAWSEEDS		
Outdoor (04-Oct-2008)		
FAB-MAP	100%	3.82%
BoW-CRF	100%	11.15%
Mixed (06-Oct-2008)		
FAB-MAP	100%	13.47%
BoW-CRF	100%	35.63%
Indoor (25-Feb-2009)		
FAB-MAP	100%	26.12%
BoW-CRF	100%	58.21%
Multisession MIT		
19-20 of July/2010		
FAB-MAP	100%	38.89%
BOW-CRF	100%	38.27%

The results of our system and of FAB-MAP over the RAWSEEDS datasets are shown in Fig. 2, and the statistics in Table II.

In the outdoor dataset, FAB-MAP does not detect all the loop closures zones, as shown in Fig. 2(a). The biggest loop is missed in the starting and final point of the experiment, in the top-right area of the map. One sample of this false negative area is shown in Fig. 3(a). The result of our system

[1] The software and vocabularies were downloaded from `http://www.robots.ox.ac.uk/~mobile/`



(a) (b)

(c) (d)

(e) (f)

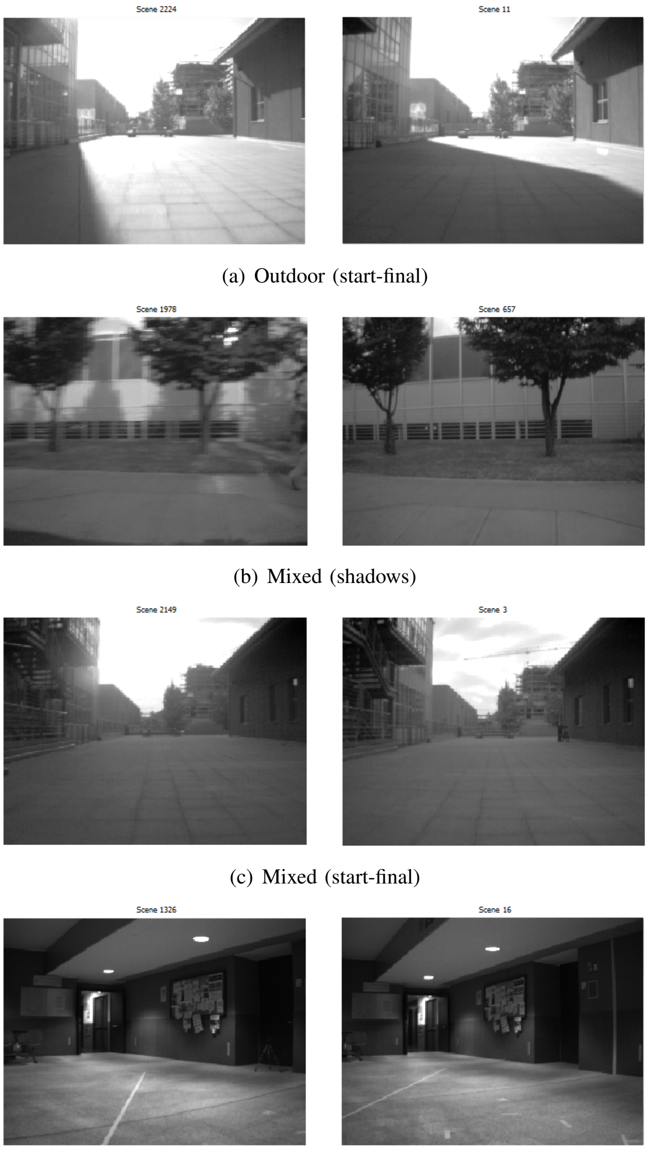
Fig. 2. Loops detected by each of the methods in the RAWSEEDS datasets. On the left results from FAB-MAP and on the right results from our system BoW + CRF-Matching. Black lines and triangles denote the trajectory of the robot; light green lines, actual loops, deep blue lines denote true loops detected.

is shown in Fig. 2(b). At 100% of precision we can detect all the loop closure areas.

For the experiment in the dynamic mixed environment we get 100% precision with both systems. Though the recall is lower in the FAB-MAP, see table II. Furthermore, all the loop closure zones are not detected, see Fig. 2(c), with false negatives as shown in Fig. 3(c), as compared with our results, see Fig. 2(d).

The indoor experiment is shown in Fig. 2. In Fig. 2(e), some loop closures are not detected by FAB-MAP, including the big area on the left hand side of the map (Fig. 3(d)), especially important in the experiment because if no loop is detected in that area, a SLAM algorithm can hardly build a correct map after having traversed such a long path (around 300 metres). The result from our system is shown in Fig. 2(f). At 100% precision we can detect all the loop closure areas.

The system also was evaluated using a dataset taken in the MIT campus in multiple sessions around of the Stata Center building, with indoor and outdoor routes taken on July of 2010. The stereo images were collected with a BumbleBee2, from PointGrey, with an 8cm baseline. We used 200 images (512x384 px) uniformly distributed in time, from an indoor session from April of 2010 to learn the weights for CRF-Matching. In the fig. 4 we sketch the trajectories (using Google Maps) and results. Both, our system and FAB-MAP obtain similar results in precision and recall. The results of our system spread more uniformly over the trajectory, see Fig. 5.



(a) Outdoor (start-final)

(b) Mixed (shadows)

(c) Mixed (start-final)

(d) Indoor (biggest loop)

Fig. 3. False negatives of FAB-MAP in the RAWSEEDS datasets. These scenes correspond to the biggest loop in the trajectories. In 3(a) the place was revisited 39 min later, and 36 min later in 3(c)

IV. Discussion and Future work

We have presented a system that combines a bag-of-words algorithm and conditional random fields to robustly solve the place recognition problem with stereo cameras. We have evaluated our place recognition system in public datasets and in different environments (indoor, outdoor and mixed). In all cases the system can attain 100% precision (no false positives) with higher recall than the state of the art (less false negatives), and detecting all (especially important) loop closure zones. No false positives mean that the environment model will not be corrupted, and less false negatives mean that it will be more precise. Our system also is more robust in situations of perceptual aliasing.

In the context of place recognition over time, our system performs well in multi-day sessions using parameters learned in different months, and this is also true of alternative systems such as FAB-MAP. The environment can also change during the operation in the same session, see Fig. 3(a-c). Our algorithm is also able to detect places revisited at different times of day, while alternative systems sometimes reject them in order to maintain high precision.

Several extensions are possible for operation in longer periods of time. The vocabulary for the BoW has shown to be useful in different environments, which suggests that a rich vocabulary needs not be updated frequently. The learned parameters in the CRF stage can be re-learned in sliding window mode depending on the duration of the mission. The system will then be able to adjust to changing conditions. In cases of periodical changes, such as times of day or seasons, incorporating a clock and calendar in the learning process would allow to maintain several environment models and select the most appropriate for a given moment of operation.

One issue to consider is the stability of the extracted descriptors in changing circumstances. In our case, SURF descriptors are useful when there is not much drift from its invariance properties [14]. For example they are still useful in the presence of outdoor seasonal changes [15], [16]. The same descriptors however do not seem useful to recognize places with totally different illuminations, e.g. an outdoor scene from day to night.

When object locations change in a scene, we think that the MSTs still allows to properly encode the scene. The MST codifies mainly local consistency (features belonging to the

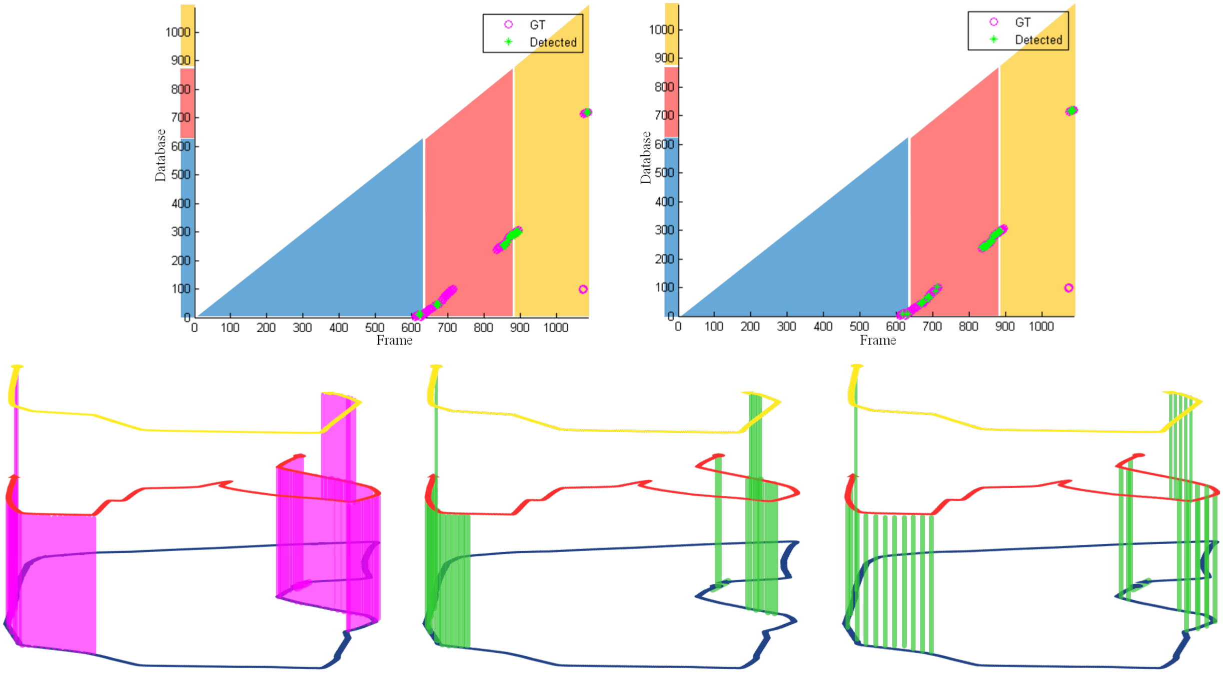


Fig. 5. Loops closure(green lines and stars) detected in the Stata Center multi-session dataset with FAB-MAP (top-left and bottom-middle) and our system (top and bottom right). Different colours correspond to different sessions (blue, red and yellow). On the top, we show the query of the current frame vs. the database with the frames already added. Ground truth (GT) is showed on bottom-left with magenta lines, on top with magenta circles.

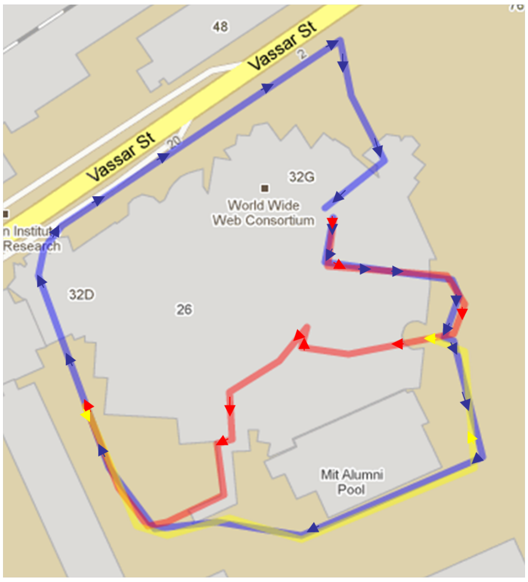


Fig. 4. Multisession experiment in the MIT campus. Different colours correspond to different sessions.

same object keep the same graph structure). Therefore, the inference process will still match features belonging to the same object. Some cases of perceptual aliasing are possible if the same objects appear in different localizations, but these cases will be much less frequent.

In our experiments, the β thresholds for acceptance of the CRF matching turned out to be clearly different for indoor and for outdoors scenarios. These parameters will also depend on the velocity of motion, mainly due to the fact that we use images from the previous second as reference in the comparisons. Incorporating the computation of these thresholds as part of the learning stage would also make the system more flexible.

Acknowledgement

We thank Dorian Gálvez-López for his collaboration in the hypothesis generation module of the place recognition system and for releasing his hierarchical bag-of-word (DBow) software library under an open source license available at http://webdiis.unizar.es/˜dorian. We also thank John McDonald for taking the datasets at the MIT campus with us.

References

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

[2] A. Angeli, D. Filliat, S. Doncieux, and J. Meyer, "A fast and incremental method for loop-closure detection using bags of visual words," *IEEE Transactions On Robotics, Special Issue on Visual SLAM*, vol. 24, pp. 1027–1037, 2008.

[3] M. Cummins and P. Newman, "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[4] ——, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *The International Journal of Robotics Research*, 2010. [Online]. Available: http://ijr.sagepub.com/content/early/2010/11/11/0278364910385483.abstract

[5] P. Piniés, L. M. Paz, D. Gálvez-López, and J. D. Tardós, "Ci-graph simultaneous localization and mapping for three-dimensional reconstruction of large and complex environments using a multicamera system," *Journal of Field Robotics*, vol. 27, pp. 561–586, 2010.

[6] R. Paul and P. Newman, "FAB-MAP 3D: Topological mapping with spatial and visual appearance," in *Proc. IEEE Int. Conf. Robotics and Automation*, may. 2010, pp. 2649 –2656.

[7] C. Cadena, D. Gálvez-López, F. Ramos, J. Tardós, and J. Neira, "Robust place recognition with stereo cameras," in *Proc. IEEE/RJS Int. Conference on Intelligent Robots and Systems*, Taipei, Taiwan, October 2010.

[8] C. Cadena, J. McDonald, J. Leonard, and J. Neira, "Place recognition using near and far visual information," in *18th World Congress of the International Federation of Automatic Control (IFAC)*, Milano, Italy, August 2011.

[9] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006, pp. 2161–2168.

[10] F. Ramos, D. Fox, and H. Durrant-Whyte, "CRF-Matching: Conditional Random Fields for Feature-Based Scan Matching," in *Robotics: Science and Systems (RSS)*, 2007.

[11] F. Ramos, M. W. Kadous, and D. Fox, "Learning to associate image features with CRF-Matching," in *ISER*, 2008, pp. 505–514.

[12] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 2001, pp. 282–289. [Online]. Available: citeseer.ist.psu.edu/lafferty01conditional.html

[13] "RAWSEEDS FP6 Project," http://www.rawseeds.org.

[14] P. Furgale and T. D. Barfoot, "Visual teach and repeat for long-range rover autonomy," *Journal of Field Robotics*, vol. 27, no. 5, pp. 534–560, 2010. [Online]. Available: http://dx.doi.org/10.1002/rob.20342

[15] C. Valgren and A. J. Lilienthal, "Sift, surf & seasons: Appearance-based long-term localization in outdoor environments," *Robotics and Autonomous Systems*, vol. 58, no. 2, pp. 149 – 156, 2010, selected papers from the 2007 European Conference on Mobile Robots (ECMR '07). [Online]. Available: http://www.sciencedirect.com/science/article/B6V16-4X908T5-6/2/679a6246b247d1b8329211a2b9df49f4

[16] T. Krajník, J. Faigl, V. Vonásek, K. Košnar, M. Kulich, and L. Přeučil, "Simple yet stable bearing-only navigation," *Journal of Field Robotics*, vol. 27, pp. 511–533, September 2010. [Online]. Available: http://dx.doi.org/10.1002/rob.v27:5