Place Recognition using Near and Far Visual Information [⋆]

César Cadena [*] John McDonald [**] John J. Leonard [***]
José Neira [*]

[*] *Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Zaragoza 50018, Spain (e-mail: {ccadena, jneira} @unizar.es)*
[**] *Department of Computer Science, National University of Ireland, Maynooth Co. Kildare, Ireland (e-mail: johnmcd@cs.nuim.ie)*
[***] *Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA (e-mail: jleonard@mit.edu)*

Abstract: In this paper we show how to carry out robust place recognition using both near and far information provided by a stereo camera. Visual appearance is known to be very useful in place recognition tasks. In recent years, it has been shown that taking geometric information also into account further improves system robustness. Stereo visual systems provide 3D information and texture of nearby regions, as well as an image of far regions. In order to make use of all this information, our system builds two probabilistic undirected graphs, each considering either near or far information. Inference is carried out in the framework of conditional random fields. We evaluate our algorithm in public indoor and outdoor datasets from the RAWSEEDS project and in an outdoor dataset obtained at the MIT campus. Results show that this combination of information is very useful to solve challenging cases of perceptual aliasing.

Keywords: Place Recognition, Conditional Random Fields, Stereo cameras, Environment Modelling.

1. INTRODUCTION

Place recognition, the identification of places that a mobile robot has already visited, is a central problem in environment modelling algorithms. Place recognition allows to close loops, increasing the precision and usefulness of the model being built. In recent years, there has been extensive use of the visual bag of words technique (BoW) of Sivic and Zisserman (2003) for this purpose. BoWs suffer from perceptual aliasing, different scenes may contain the same features although in different configurations, but the system will consider them the same place. Several alternatives have been proposed to improve the robustness of this technique. Recently, Cadena et al. (2010) proposed a system that uses as a BoW combined with matching 3D information using Conditional Random Fields (CRFs), a technique first proposed by Ramos et al. (2008). When the system finds several alternative places that match the current image, 3D information is used to disambiguate. But the CRF ignores information in the image that is not 3D, the far regions, or background. In this paper we propose an improvement to that system which considers inference also about the features that are in the background of the images obtained. We show that by taking into account the two types of information, near and far, the system is able to achieve good performance in challenging scenes for which previous techniques failed. The complete system is evaluated on several public indoor and outdoor datasets from the Rawseeds project and on an outdoor dataset obtained at the MIT campus.

This paper is organised as follows. We begin with a discussion of the related work in Section 2. We then provide a brief description of the our previous system for place recognition in Section 3. In Section 4 we provide an overview of Conditional Random Fields applied to associate scenes using near and far information. In Section 5 we describe the decision algorithm followed to obtain the loop closures. Finally, we analyse in Section 6 experimental results on real data that demonstrate the improvement in very difficult scenes.

2. RELATED WORK

Recent advances to solve the loop closing problem include those based on map or image features (Williams et al., 2009), robot poses (Olson, 2009) and 3D range data (Steder et al., 2010). Williams et al. (2009) show that the image-to-map method for loop closing does not scale well in larger environments. The method proposed by Olson (2009) requires odometry in order to generate a hypothesis for loop closing, a weak technique for large loops. Steder et al. (2010) use 3D data from a laser scan, thus without texture and with limited range.

[⋆] This research has been funded by the Dirección General de Investigación of Spain under projects DPI2009-13710, DPI2009-07130 and by the Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan.

Appearance-based methods have become very popular since cameras provide rich scene information for a low price. These methods focus on place recognition, and mainly use the bag-of-words representation (Sivic and Zisserman, 2003), supported by some probabilistic framework (Angeli et al., 2008). On the issue of recognition of places using only visual information perhaps the state of the art is the FAB-MAP (Cummins and Newman, 2008), since it has proved very successful with a low proportion of false positives. However, that system presents problems in applications using front facing cameras (Piniés et al., 2010). Recently proposed by Paul and Newman (2010), the FAB-MAP 3D uses the same framework including 3D information provided by a laser scanner, but can only make inferences about visual features in its range.

There are other works that join image and geometrical data, such as Newman et al. (2006), where an actuated laser scanner and a monocular camera are used. However, this system is not able to combine data from the two sensors, only camera information is used to detect loop closure events.

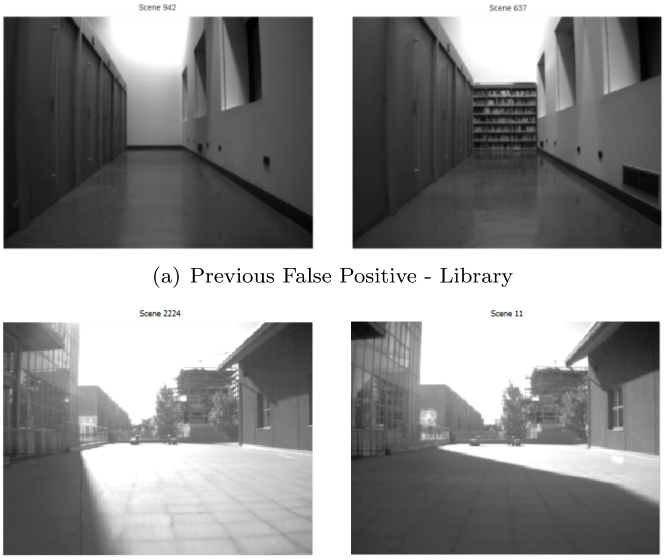
CRF-Matching was applied to associate images features in (Ramos et al., 2008), using a 2D Delaunay triangulation as a graph structure. However, the 2D relative location of features in an image is not robust to changes in camera location.

3. FRAMEWORK

In Cadena et al. (2010) we proposed a place recognition system with the following two components:

- The first component is based on the bag-of-words method (BoW) of Sivic and Zisserman (2003) which is implemented in a hierarchical way (Nister and Stewenius, 2006). In our implementation, λ_t is the BoW score computed between the current image and the previous one in the database. The minimum confidence expected for a loop closure candidate is α^-; the confidence for a loop closure to be accepted 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
  3. otherwise. the match is ignored.
- The second component consists of checking, when required, the previous candidates with CRF-Matching in 3D space. CRF-Matching is an algorithm based on Conditional Random Fields (CRF). CRF-Matching is a probabilistic model (more on CRF-Matching in the next Section). We compute the negative log-likelihood $\Lambda_{t,t'}$ from the maximum a posteriori (MAP) association between the current scene at time t and the candidate scene at time t'. We accept the match only if $\Lambda_{t,t'} \leq \Lambda_{t,t-1}$.

This system exploits the efficiency of BoW to detect revisited places candidates in real-time. CRF-Matching is a more computationally demanding data association algorithm because it uses much more information than BoW. For this reason, only the positive results of BoW are considered for CRF-Matching. The successful results of BoW, filtered by the CRF-Matching, determine the loop closures.



(a) Previous False Positive - Library

(b) Previous False Negative - Illumination

Fig. 1. Two of the challenging cases of errors that the previous system is not able to correctly judge. 1(a) belongs to the Rawseeds Indoor dataset, the central background is different, but the stereo system cannot compute 3D information about it. 1(b) belongs to the Rawseeds Outdoor, the background is the same, but the 3D information is greatly affected by the differences in illumination of the scenes.

A limitation of this system is that perceptual aliasing, or false positives, occurs in scenes where near information is very similar, although far or background information is not. Fig. 1(a) show an example in which near information (the image borders) is very similar, but far information (the image center) is different, resulting in false positives. These errors are catastrophic for the environment model being built, since it falsely connects unrelated areas. False negatives, fig. 1(b), or not being able to identify images from the same scene as correspondent, are not as catastrophic, although the precision of the resulting model is negatively affected when common areas are not identified. Ideally, all false positives as well as all false negatives should be avoided.

In this paper we modify the second component of this system by adding the CRF-Matching over the far or background information. As we will see, this results in the detection of these false positive and false negative cases, which improves the robustness of the place recognition system.

4. CRF-MATCHING

In this section we describe the CRF-Matching process for image features with 3D information (near), and for images features without 3D information (far).

4.1 Model definition

CRF-Matching is based on Conditional Random Fields, undirected graphical models developed for labeling sequence data (Lafferty et al., 2001). Instead of relying on Bayes rule to estimate the distribution over hidden states $\mathbf{x}$ from observations $\mathbf{z}$, CRFs directly model $p(\mathbf{x}|\mathbf{z})$, the *conditional* distribution over the hidden variables given observations. Due to this structure, CRFs can handle arbitrary dependencies between the observations, which gives them substantial flexibility in using complex and overlapped attributes or observations.

The nodes in a CRF represent hidden states, denoted $\mathbf{x} = \langle \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \rangle$, and observations, denoted $\mathbf{z}$. In our framework the hidden states correspond to all the possible associations between the n features in scene A and the m features in the scene B, i.e. $\mathbf{x}_i \in \{1, 2, \ldots, m+1\}$, where the additional state is the outlier state. Observations are provided by the sensors (e.g., 3D point cloud, appearance descriptors, or any combination of them). The nodes $\mathbf{x}_i$ along with the connectivity structure represented by the undirected graph define the conditional distribution $p(\mathbf{x}|\mathbf{z})$ over the hidden states $\mathbf{x}$. Let $\mathcal{C}$ be the set of cliques (fully connected subsets) in the graph of a CRF. Then, a CRF factorizes the conditional distribution into a product of *clique potentials* $\phi_c(\mathbf{z}, \mathbf{x}_c)$, where every $c \in \mathcal{C}$ is a clique in the graph, and $\mathbf{z}$ and $\mathbf{x}_c$ are the observed data and the hidden nodes in such clique. Clique potentials are functions that map variable configurations to non-negative numbers. Intuitively, a potential captures the "compatibility" among the variables in the clique: the larger a potential value, the more likely the configuration. Using the clique potential, the conditional distribution over hidden states is written as:

$$p(\mathbf{x}|\mathbf{z}) = \frac{1}{Z(\mathbf{z})} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{z}, \mathbf{x}_c) \tag{1}$$

where $Z(\mathbf{z}) = \sum_{\mathbf{x}} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{z}, \mathbf{x}_c)$ is the normalising partition function. The computation of this function can be exponential in the size of $\mathbf{x}$. Hence, exact inference is possible for a limited class of CRF models only, e.g. in tree-structured graphs.

Potentials $\phi_c(\mathbf{z}, \mathbf{x}_c)$ are described by log-linear combinations of *feature functions* $\mathbf{f}_c$, i.e., the conditional distribution (1) can be rewritten as:

$$p(\mathbf{x}|\mathbf{z}) = \frac{1}{Z(\mathbf{z})} \exp\left\{ \sum_{c \in \mathcal{C}} \mathbf{w}_c^T \cdot \mathbf{f}_c(\mathbf{z}, \mathbf{x}_c) \right\} \tag{2}$$

where $\mathbf{w}_c^T$ is the transpose of a weight vector, which represents the importance of different features for correctly identifying the hidden states. Weights can be learned from labeled training data.

4.2 Inference

Inference in a CRF estimates the marginal distribution of each hidden variable $\mathbf{x}_i$, and can thus determine the most likely configuration of the hidden variables $\mathbf{x}$ (i.e., the maximum a posteriori, or MAP, estimate). Both tasks can be solved using *belief propagation* (BP) (Pearl, 1988), which works by transmitting messages containing beliefs through the graph structure of the model. Each node sends messages to its neighbours based on messages it receives and the clique potentials. BP generates exact results in graphs with no loops, such as trees or polytrees. To create the graph structures we use the minimum spanning tree (MST), the first over the 3D metric coordinates of the SURF features extracted from the one of the image in the stereo pair (graph3D), and the second one over the pixel coordinates of the remaining SURF features (graphIm), see fig. 2.

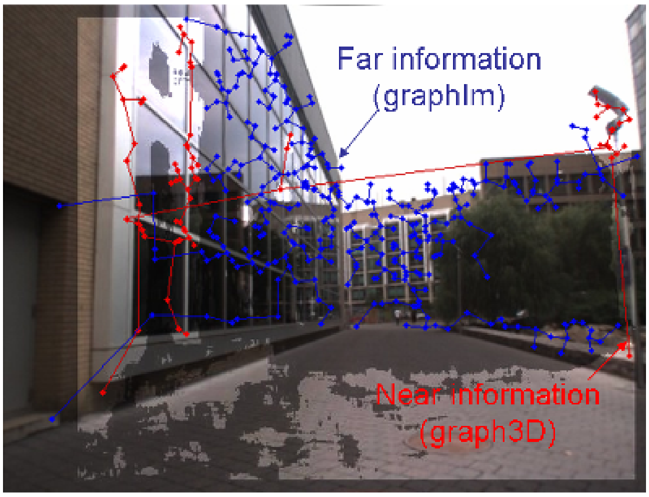


Fig. 2. We apply the CRF-Matching over both graphs. In blue the graph for far features (graphIm), in red the graph for near features (graph3D). The minimum spanning tree for graph3D is computed with the metric coordinates, and here is projected over the images only for visualisation.

Here is the major difference with Ramos et al. (2008): they used the 2D Delaunay triangulation as graph structure, containing loops. As Quattoni et al. (2007) showed, the MST allows exact inference with no loss of accuracy.

4.3 Parameter learning

The goal of parameter learning is to determine the weights of the feature functions used in the conditional likelihood (2). CRFs learn these weights discriminatively by maximising the conditional likelihood of labeled training data. We resort to maximising the *pseudo-likelihood* of the training data, which is given by the product of all local likelihoods $p(\mathbf{x}_i|\text{MB}(\mathbf{x}_i))$; $\text{MB}(\mathbf{x}_i)$ is the Markov Blanket of variable $\mathbf{x}_i$, which contains the immediate neighbours of $\mathbf{x}_i$ in the CRF graph. Optimisation of this pseudo-likelihood is performed by minimising the negative of its log, resulting in the following objective function:

$$L(\mathbf{w}) = -\sum_{i=1}^{n} \log p(\mathbf{x}_i|\text{MB}(\mathbf{x}_i), \mathbf{w}) + \frac{\mathbf{w}^T\mathbf{w}}{2\sigma_{\mathbf{w}}^2} \tag{3}$$

The rightmost term in (3) serves as a zero-mean Gaussian prior, with variance $\sigma_{\mathbf{w}}^2$, on each component of the weight vector.

For graph3D, the training data is labeled from the best rigid-body transformation using RANSAC after a SURF matching (Olson, 2008) of two consecutive scenes. In the case of graphIm, the training data is labeled from the fundamental matrix using RANSAC after a SURF matching of two consecutive scenes.

4.4 Feature description

CRF-matching can employ arbitrary local features to describe shape, image properties, or any particular aspect of the data. Our features describe *differences* between shape (only for graph3D) and appearance (for both graphs) of the features. The local features we use are the following:

Shape difference: These features capture how much the local shape of dense stereo data differs for each possible association. We use the geodesic, PCA and curvature distance.

The *geodesic distance*, defined as the sum of Euclidean distances between points in the minimum spanning tree, provides shape information of a scene. It can be calculated for different neighbourhoods representing local or long-term shape information. Given points $z_{A,i}$, $z_{B,j}$ and a neighbourhood k, the geodesic distance feature is computed as:

$$\mathbf{f}_{geo}(i,j,k,z_A,z_B) = \frac{\left\| \sum_{l=i}^{i+k-1} \|z_{A,l+1} - z_{A,l}\| - \sum_{l=j}^{j+k-1} \|z_{B,l+1} - z_{B,l}\| \right\|}{\sigma} \tag{4}$$

where i and j correspond to the hidden state $\mathbf{x}_i$ that associates the feature i of the scene A with the feature j of the scene B. The neighbourhood k of $\mathbf{x}_i$ in the graph corresponds to all the nodes separated by k nodes from $\mathbf{x}_i$. In our implementation, this feature is computed for $k \in \{1, 2, 3\}$. A similar feature is used to match 3D laser scans in Anguelov et al. (2005). Parameter σ controls the scale of the corresponding distance (as in subsequent equations).

We also use Principal Component Analysis over the dense 3D point cloud that is contained within a radius given by the scale obtained by the SURF extractor for each node in the graph. Then *PCA distance* is computed as the absolute difference between principal components of a dense point cloud $z^{pca}_{A,i}$ in scene A and $z^{pca}_{B,j}$ in scene B:

$$\mathbf{f}_{PCA}(i,j,z^{pca}_A,z^{pca}_B) = \frac{\left| z^{pca}_{A,i} - z^{pca}_{B,j} \right|}{\sigma} \tag{5}$$

Another way to consider local shape is by computing the difference between the curvatures of the dense point clouds. This feature is computed as:

$$\mathbf{f}_{curv}(i,j,z^c_A,z^c_B) = \frac{\left\| z^c_{A,i} - z^c_{B,j} \right\|}{\sigma} \tag{6}$$

where $z^c = \frac{3s_3}{s_1+s_2+s_3}$, and $s_1 \geq s_2 \geq s_3$ are the *singular values* of the point cloud of each node.

Visual appearance: These features capture how much the local appearance from the points in the image differs for each possible association. We use the *SURF distance.* This feature calculates the Euclidean distance between the descriptor vectors for each possible association:

$$\mathbf{f}_{SURF}(i,j,z^{descr}_A,z^{descr}_B) = \frac{\left\| z^{descr}_{A,i} - z^{descr}_{B,j} \right\|}{\sigma} \tag{7}$$

Ramos et al. (2008) also includes the distance between the individual components of the descriptor. In our training and validations data we do not find a significant improvement in the accuracy of the labeling, and this in turn greatly increases the size of the weight vector.

All previous features described are unary, in that they only depend on a single hidden state i in scene A (indices j and k in the features denote nodes in scene B and neighbourhood size). In order to generate mutually *consistent* associations it is necessary to define features, over the cliques, that relate the hidden states in the CRF to each other.

Pairwise distance: This feature measures the consistency between the associations of *two* hidden states $\mathbf{x}_i$ and $\mathbf{x}_j$ and observations $z_{A,i}$, $z_{A,j}$ from scene A and multiple observations $z_{B,k}$ and $z_{B,l}$ in scene B:

$$\mathbf{f}_{pair}(i,j,k,l,z_A,z_B) = \frac{\left\| \|z_{A,i} - z_{A,j}\| - \|z_{B,k} - z_{B,l}\| \right\|}{\sigma} \tag{8}$$

The z_A and z_B are in metric coordinates for graph3D, and in images coordinates for graphIm.

5. DECISION

Our improved place recognition system can be summarized in the algorithm 1.

Algorithm 1 Pseudo-algorithm of our new 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 graph3D and graphIm
      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 negative log-likelihood Λ^{3D} of the MAP association for graph3D provides a measure of how similar two scenes are in terms of close range, and Λ^{Im} for graphIm in terms of far range. Thus, we compare how similar is the current scene with the scene in t', $\Lambda_{t,t'}$, with respect to how similar the current scene is with the scene in $t-1$, $\Lambda_{t,t-1}$. With the β parameter we can control the level we demand of similarity to $(t, t-1)$, lower means more demanding. By choosing different parameters for near and far information we can make a balance between the weight of each in our acceptance.

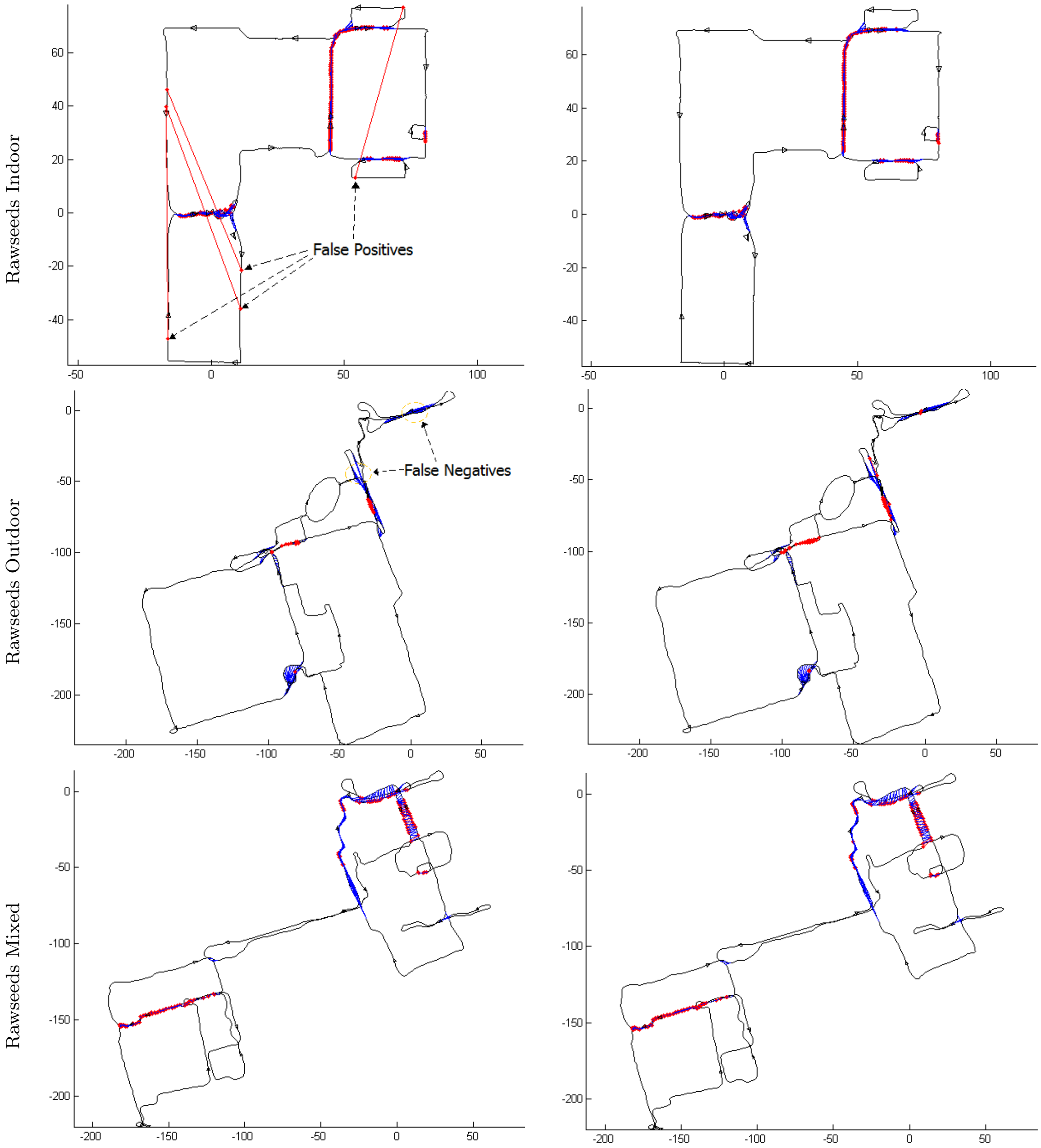


Fig. 3. Loops detected by each dataset with our previous system (left) and our improved system (right). Blue: ground truth loop closings. Red: detected loop closings.

6. EXPERIMENTS

We have evaluated our system with the public datasets from the RAWSEEDS Project [1]. 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. These are b/w images (640x480 px) taken at 15 fps. We used 200 images uniformly distributed in time, from a static mixed dataset, for training the vocabulary for BoW and for learning the weights for CRF-Matching. 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 for more details.

In fig. 3 we show the results of our whole system (right) and we compare with the results from Cadena et al. (2010) (left). We can see that we are now able to eliminate all false positive matches in the indoor dataset. There is also a reduction of false negatives in all cases (see table 1).

The system also was evaluated using a dataset taken in the MIT campus in multiple and different sessions, around of the Stata Center building, with indoor and outdoor routes. The stereo images were collected with a BumbleBee2, from PointGrey, with a 8cm of baseline. We used 200 images (512x384 px) uniformly distributed in time, from an indoor session to learn the weights for CRF-Matching. In the fig. 4 we sketch the results (using Google Maps) of our whole

[1] RAWSEEDS is an European FP6 Project, `http://www.rawseeds.org`

Table 1. Results for all datasets

	Precision	Recall
Rawseeds Indoor		
previous	97.53%	58.96%
improved	100%	58.21%
Rawseeds Outdoor		
previous	100%	5.73%
improved	100%	11.15%
Rawseeds Mixed		
previous	100%	33.83%
improved	100%	35.63%
Stata Center - MIT		
previous	100%	23.46%
improved	100%	38.27%

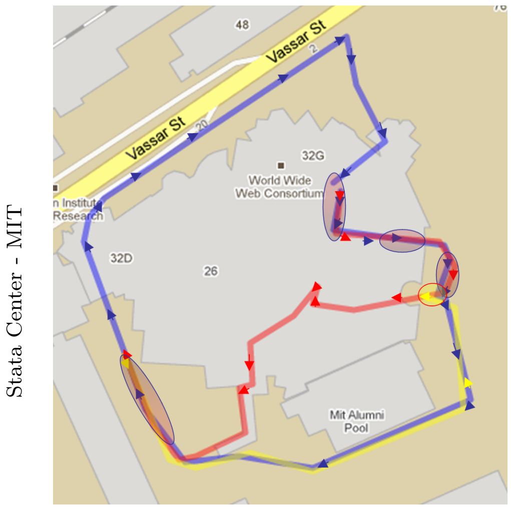


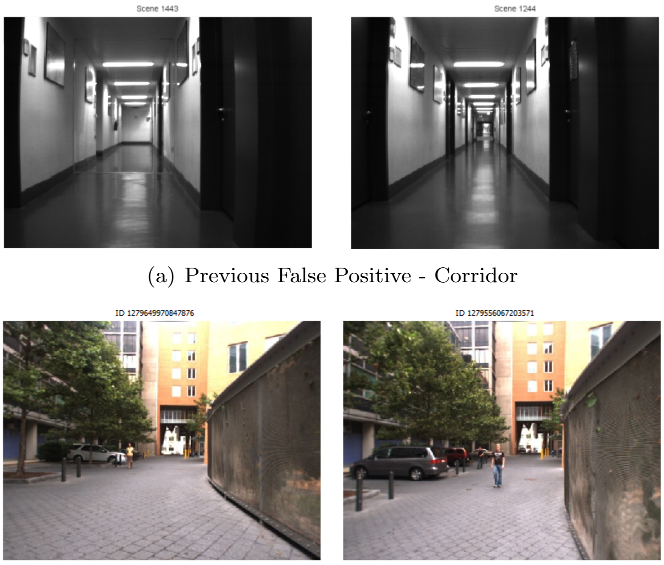
Fig. 4. All loops (oval regions) are detected in the Stata Center multisession dataset with our improved system. Different colours correspond to different sessions.

system in this experiment using 3 sessions (in different colours). The ovals indicate the zones with detected loop closures from our improved system. For this dataset we have no ground truth. In this dataset we also obtain a precision of 100% (see table 1), and recall is also increased.

Our last research C++ implementation of the system runs at $1fps$. Extracting SURF features is usually done in $0.15s$ per image, whereas running the BoW algorithm and maintaining the inverted file takes $11ms$ on average. For each loop closing candidate, the CRF stage takes, on average, $0.3s$ to process and infer over both near ($0.15s$) and far ($0.15s$) information. Our whole system takes on average $0.47s$ per frame with a maximum of $1.04s$.

Fig. 1 shows example where the previous system failed (false positive, false negative) because of not using far information. Fig. 5 shows two additional example scenes in which, thanks to weight in the decision of far information, the output is true positive and true negative, in contrast with the previous version.

The parameters α were constant in all the experiments with $\alpha^- = 0.15$ and $\alpha^+ = 0.6$. The β parameters were different for indoor and outdoor datasets, the values as shown in table 2.



(a) Previous False Positive - Corridor

(b) Previous False Negative - Point of view

Fig. 5. Two more examples of errors that the new system is able to detect thanks to inference being carried out also on the background image. 5(a) belongs to the Rawseeds Indoor dataset, the central background is different, but the stereo system cannot compute 3D information about it. 5(b) belongs to the MIT campus dataset, the background is the same, but the 3D information is affected by the differences in point of view.

Table 2. Parameter β

	β_{3D}	β_{Im}
Rawseeds Indoor	1	1.3
Rawseeds Outdoor	1.5	1.7
Rawseeds Mixed	1.5	1.7
Stata Center - MIT	1.5	1.7

7. DISCUSSION AND FUTURE WORK

We have presented a robust place recognition system based on stereo. By using jointly the CRF-Matching algorithm over visual near and far information, we have demonstrated that challenging false loop closures can be rejected. We have evaluated this improved place recognition system in public datasets, with different stereo camera systems, and in different environments. In all cases the precision obtained was 100%, and recall was increased when compared to our previous system. No false positives mean that the environment model will not be corrupted, and less false negatives mean that it will be more precise.

The tuning of the β parameters suggests a pattern for outdoors and another for indoors, but this is not conclusive. 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. As immediate future work we will investigate the automatic computation of the β parameters. One promising idea is including them in the learning process of each CRF-Matching.

ACKNOWLEDGEMENTS

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.

REFERENCES

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

Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., and Davis, J. (2005). SCAPE: shape completion and animation of people. *ACM Trans. Graph.*, 24(3), 408–416. doi:http://doi.acm.org/10.1145/1073204.1073207.

Cadena, C., Gálvez, D., Ramos, F., Tardós, J., and Neira, J. (2010). Robust place recognition with stereo cameras. In *Proc. IEEE/RJS Int. Conference on Intelligent Robots and Systems*. Taipei, Taiwan.

Cummins, M. and Newman, P. (2008). FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research*, 27(6), 647–665. doi:10.1177/0278364908090961.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, 282–289. Morgan Kaufmann, San Francisco, CA. URL `citeseer.ist.psu.edu/lafferty01conditional.html`.

Newman, P., Cole, D., and Ho, K.L. (2006). Outdoor SLAM using visual appearance and laser ranging. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. Orlando Florida USA.

Nister, D. and Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, 2161–2168. doi:10.1109/CVPR.2006.264.

Olson, E. (2009). Recognizing places using spectrally clustered local matches. *Robotics and Autonomous Systems*, 57(12), 1157–1172.

Olson, E. (2008). *Robust and Efficient Robotic Mapping*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

Paul, R. and Newman, P. (2010). Fab-map 3d: Topological mapping with spatial and visual appearance. In *Proc. IEEE Int. Conf. Robotics and Automation*, 2649 –2656.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

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

Quattoni, A., Wang, S., Morency, L.P., Collins, M., and Darrell, T. (2007). Hidden conditional random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10), 1848–1852. doi:10.1109/TPAMI.2007.1124.

Ramos, F., Kadous, M.W., and Fox, D. (2008). Learning to associate image features with CRF-Matching. In *ISER*, 505–514.

Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, 1470–1477.

Steder, B., Grisetti, G., and Burgard, W. (2010). Robust place recognition for 3d range data based on point features. In *Proc. IEEE Int. Conf. Robotics and Automation*, 1400 –1405.

Williams, B., Cummins, M., Neira, J., Newman, P., Reid, I., and Tardós, J. (2009). A comparison of loop closing techniques in monocular slam. *Robotics and Autonomous Systems*.