Learning to Associate Image Features with CRF-Matching

Fabio Ramos[1], M. Waleed Kadous[2], and Dieter Fox[3]

[1] ARC Centre of Excellence for Autonomous Systems, Australian Centre for Field Robotics, The University of Sydney, Sydney, NSW 2006, Australia. f.ramos@acfr.usyd.edu.au
[2] School of Computer Science and Engineering, University of New South Wales, Sydney NSW 2052, Australia. waleed@cse.unsw.edu.au
[3] Department of Computer Science & Engineering, University of Washington, Seattle, WA 98195, USA. fox@cs.washington.edu

Summary. This paper presents a supervised learning algorithm for image feature matching. The algorithm is based on *Conditional Random Fields* which provides a mechanism for globally reason about the associations. The novelty of this work is twofold: (i) the use of Delaunay triangulation as the graph structure for a probabilistic network to reason about image feature association; (ii) the combination of local and joint features to enforce consistency in a theoretically sound statistical learning procedure. Experimental results show that our approach outperforms RANSAC in our challenging datasets consisting of indoor and outdoor images, with significant occlusion, blurring, rotational and translational transformations.

1 INTRODUCTION

Data association remains a difficult and fundamental problem for many robotics tasks. Despite continuous efforts from the computer vision and the robotics communities, data association is still the key limitation for larger-scale problems in tracking, image registration, reconstruction, and simultaneous localisation and mapping (SLAM).

Most of the data association algorithms are unsupervised, i.e., given a set of possible associations, the algorithms try to find the correct matches without *a priori* information. However, they commonly rely on ad hoc heuristics which require the manual definition of thresholds. This is clearly not ideal since it can be difficult to specify thresholds in problems with large variability. Most algorithms are also limited by the independence assumption, where each possible association is considered as a separate problem, with no influence on the association of other data points located in the same vicinity. Finally, conventional algorithms only provide a deterministic result on the association. This

makes them less robust and difficult to incorporate in probabilistic filtering approaches since no uncertainty on the association is returned.

This paper builds upon the recently proposed algorithm for data association of laser scans known as CRF-Matching [15]. CRF-Matching is a supervised probabilistic model able to jointly reason about the association of points. This is obtained by overcoming the independence assumption through the use of Conditional Random Fields (CRFs) [8]. CRFs are an extremely flexible technique for integrating different features in the same probabilistic framework. The power of CRFs is enhanced through the possibility to use statistical measurements (such as the likelihood of the data given the model) to learn a parametrisation of the model given some training data.

CRF-matching was initially proposed for laser scan matching as an alternative to the Iterative Closest Point method [3], where no initialisation is necessary. In this paper, the approach is extended to reason about the association of image features. We propose the use of the Delaunay triangulation [13] as the graph structure for CRF-matching. The graph defines neighbour points by respecting interesting geometric constraints such as the *empty circle property*. We demonstrate how pairwise potential functions can be defined over edges to jointly reason about the associations.

The main contributions of this paper are: (i) A supervised learning algorithm for data association of image features; (ii) A probabilistic model defined over Delaunay triangulation to encode dependent and relational data; (iii) An experimental evaluation of image feature algorithms in challenging, yet realistic, datasets often found in field robotics.

2 RELATED WORK

Frequently in processing images from robots, there is a need to find correspondences between images. This has applications in visual SLAM, panorama creation, stereo vision and object recognition. The task is difficult because of the ego-motion of the robot, moving objects, the lighting changing, or objects becoming occluded. One approach is the use of local descriptions, the most popular of these being SIFT (scale invariant feature transform) [10]. However, this set of matches is putative and may contain errors because the location of matched descriptors (and hence the spatial consistency of the matches) is not taken into account. Consequently, computing an association such as a homography leads to errors, because mismatches are included in the calculations. Typically a secondary step to remove incorrect matches is applied such as RANSAC, (for example, in [5], [16] and [6]).

Nonetheless, others have found that in practice while RANSAC is efficient, it does have a number of drawbacks. Firstly, because it uses a random subset of the data, different runs of RANSAC will produce different models and inliers, especially when the number of matches is small (typically below 10) and the results can vary between runs. Secondly, it requires that the user set the appropriate threshold for outliers. Thirdly, it is subject to occasional

failure, when a totally incorrect model is fit. Fourthly, especially in images where there is textural ambiguity (e.g. buildings with recurring features such as windows), it may make incorrect choices. To remedy this issue, [17] suggests a generalised RANSAC algorithm that avoids committing to a single "best" match for the correspondence.

CRF-Matching overcomes most of these problems. As a probabilistic network, results are probabilistic distributions over the space of possible associations. This additional information can be used to compute the uncertainty of the robot movement in localisation tasks. CRF-Matching does not require initialisation and does not need to compute homography matrices. Thus it can still be employed even when the number of detected image features is as small as two. Additionally, as CRF-Matching parameters can be estimated in a supervised learning procedure, no threshold need to be manually specified.

3 CRF-MATCHING

3.1 Model Definition

CRF-Matching is based on Conditional Random Fields: undirected graphical models developed for labelling sequence data [8]. CRFs directly model $p(\mathbf{x}|\mathbf{z})$, the *conditional* distribution over the hidden variables $\mathbf{x}$ given observations $\mathbf{z}$. This is in contrast to generative models such as Hidden Markov Models or Markov Random Fields, which apply Bayes rule to infer hidden states [14]. Due to this structure, CRFs can handle arbitrary dependencies between the observations $\mathbf{z}$, which gives them substantial flexibility in using high-dimensional feature vectors.

We use the term *features* with two different meanings in the following description. Image feature refers to features detected by SIFT. Local or Pairwise features refer to functions defined over observations (local) and, observations and hidden states (pairwise). The appropriate meaning should be clear from the context. The nodes in a CRF represent hidden states, denoted $\mathbf{x} = \langle \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \rangle$, and data, denoted $\mathbf{z}$. The nodes $\mathbf{x}_i$, along with the connectivity structure represented by the undirected edges between them, 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 the clique c, respectively. 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 the potential value, the more likely the configuration. Using clique potentials, the conditional distribution over hidden states is written as

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

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 partition function can be exponential in the size of $\mathbf{x}$. Hence, exact inference is possible for a limited class of CRF models only.

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

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

To perform data association between image features detected in an image A and image features in an image B, CRF-Matching creates a graphical model over the features in A. Each hidden variable $\mathbf{x}_i$ has a multinomial distribution where each state j in $\mathbf{x}_i$ corresponds to the probability that feature i in image A is associated to feature j in image B. To create the graph structure we use the Delaunay triangulation algorithm [13]. This triangulation has the *empty circle property* meaning that no other image feature detected in A is inside any circumcircle defined over triangles computed by the algorithm. In practice, this guarantees that the neighbourhood information is properly encoded while not establishing direct relationship between features located far from each other. Also, it avoids long edges connecting parts of the image with less contextual relationship, while defining local neighbourhoods that are more appropriate for matching. The graph represents connections between the hidden variables $\mathbf{x}_i$ and ensures global consistence. Another reason for using the Delaunay triangulation is the existence of efficient open source implementations available in [9]. The local observations z_i describe local appearance properties of the image features represented as SIFT descriptors. Consistency is taken into account through the pairwise features indicated by edges. Outlier detection is also considered in the model as an additional state for x_i.

Parameter learning in CRF aims at determining the weights of the feature functions. CRFs learn these weights discriminatively by maximising the conditional likelihood of labelled training data. We resort to maximising the *pseudo-likelihood* of the training data, which is given by the sum of local likelihoods $p(\mathbf{x}_i \mid \mathrm{MB}(\mathbf{x}_i))$, where $\mathrm{MB}(\mathbf{x}_i)$ is the Markov blanket of variable $\mathbf{x}_i$: the set of the immediate neighbours of $\mathbf{x}_i$ in the CRF graph [1].

Inference in CRFs can estimate either the marginal distribution of each hidden variable $\mathbf{x}_i$ or the most likely configuration of all hidden variables $\mathbf{x}$ (*i.e.*, MAP estimation), as defined in (2). Both tasks can be solved using *belief propagation* (BP), which works by sending local messages through the graph structure of the model [12, 11].

3.2 Feature Description

CRF-Matching can employ arbitrary local potentials to describe image properties, or any particular aspect of the data. If other sensors are available, their

measurements can also be incorporated. Since our focus is on associating image features from two images, our local features describe *differences* between SIFT descriptors. The learning algorithm provides means to weight each of the resulting potentials to best associate the data. The local potentials are described as follows:

SIFT descriptor distance: This feature measures the difference between individual SIFT feature element in one image w.r.t. individual SIFT element in the other image. As opposed to the SIFT match procedure where only the Euclidean distance between the 128-dimensional descriptor is used, this feature provides distances for each element individually. This will allow an optimal combination of elements for matching during the learning procedure.

Euclidean distance of SIFT descriptor: This feature is essentially the same used by the SIFT Match algorithm [10]. It calculates the Euclidean distance between the 128 element vector, for each possible association.

The following feature is used to define the clique potentials of nodes connected in the CRF.

Pairwise distance: This feature computes distances between neighbour nodes in the CRF graph and compares with distances between two image features in the other image. The idea is to use the spatial arrangement of the image features to enforce consistency. This feature is defined of over *two* hidden nodes $\mathbf{x}_i$ and $\mathbf{x}_j$ and observations $z_{A,i}$, $z_{A,j}$ from image A and multiple observations $z_{B,m}$ and $z_{B,n}$ in image B.

4 EXPERIMENTS

We evaluate CRF-Matching in two different datasets. For each dataset 30 pairs of images were selected and manually labelled. CRF-Matching was implemented using the open source MATLAB SIFT package[4]

We compared the performance of RANSAC with SIFT features using homography as the fitting algorithm, and CRF-Matching with SIFT features. The CRF was trained on 29 images, and tested on the 30th, and this was repeated for each of the 30 images in each dataset (this is commonly known as leave-one-out cross validation). Because the homography computation used by RANSAC requires at least 4 matching points, RANSAC could not be used on 4 of the indoor images, therefore we excluded those four images from the results[5]. It must also be remembered that RANSAC is non-deterministic, and especially in the case where there are few SIFT matches, it can be sensitive to initial conditions. Therefore, RANSAC was run 10 times and the results averaged.

To compare matching performance empirically, information retrieval metrics were used to assess whether the correct matches were selected. This al-

[4] Available at http://vision.ucla.edu/ vedaldi/code/sift/sift.html. Other image feature detectors could also be used but no open source implementations where supplied by the authors.

[5] Inclusion of these results would underestimate the performance of RANSAC.

lowed us to evaluate correctness of the match in a model-independent manner. The three measures used are Precision, Recall and F_1 score.

4.1 Indoor Dataset

To evaluate the effectiveness of CRF-Matching on data collected from robots, we used a dataset that was collected at Robocup 2007 as part of the Rescue Robot League competition. The competition is administered by the National Institute of Standards and Technology (NIST) which constructs a mock disaster site the size of a small house. We gathered the data using a variant of CASTER Scorpion [7], which is equipped with a range imager, and a camera. As the range imager has a limited point of view (approximately 44°), it is mounted on a pan-tilt unit with some overlap (however, due to robot movement and flex in the mountings, the exact position is only known to within 5°). The images obtained have resolution of 176×144 pixels. In these experiments we do not use range or position information, and use only the image pairs.

An illustrative comparison of the different techniques and stages in application of the algorithm is shown in Figure 1 for a typical pair of images from amongst the 30 pairs. As common for this dataset, there is no much texture. Figure 1(b) shows the results of applying SIFT matching. Note that while there are several correct matches, there are several extraneous matches. Figure 1(c) shows ground truth: the result of hand labelling the data by humans. Figure 1(d) shows the results of applying RANSAC. It does pick out three correct matches, but it also identifies two extraneous ones. Finally, Figure 1(e) shows the results of applying CRF with MAP inference, note that it missed the top match, but it is otherwise correct. These results are typical for our data; as can be seen, CRF outperforms RANSAC in this case.

	Rescue			Outdoor		
	F_1 score	Precision	Recall	F_1 score	Precision	Recall
CRF	0.6206	0.6546	0.6196	0.6869	0.6558	0.7391
RANSAC	0.5830	0.5380	0.6682	0.6696	0.6639	0.6993

Table 1. Information Retrieval measures on rescue data and outdoor dataset. In both cases RANSAC used an outlier threshold of 0.05.

Table 1 shows results for the rescue dataset. Because the homography computation used by RANSAC requires at least 4 matching points, RANSAC could not be used on 4 of the images, therefore we excluded those four images from the results[6]. This illustrates another advantage of CRFs: it can be used in situations where there are not many matches. CRF-matching attained a

[6] Inclusion of these results would underestimate the performance of RANSAC.

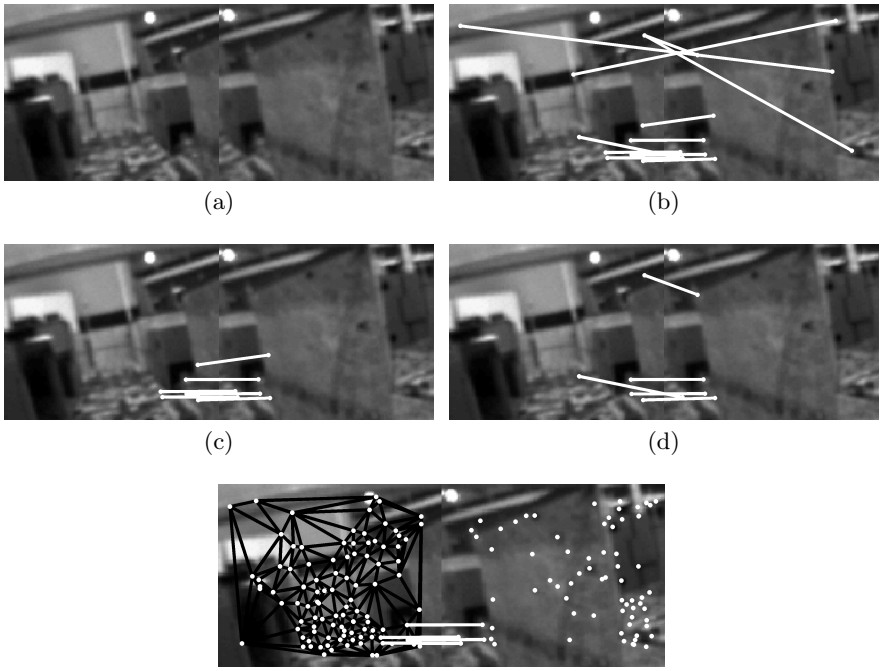

Fig. 1. A sequence of images showing the different results of the algorithms. Figure (a) shows a typical image pair with no matches. Note that there is no much texture in the images. Figure (b) shows the result of applying SIFT matching, with no subsequent filtering. Figure (c) shows hand-labelled results used for training and evaluating the matches. Figure (d) shows the results using RANSAC. Figure (e) shows the graphical model, the SIFT features and CRF-Matching results.

higher F_1 score overall across all of the images. It is also worth noting that the RANSAC results show a great deal of variability, over the ten runs, the range was 0.0852, with the minimum being 0.5474 for the F_1 score. We can see that where the number of matches is small, as it is on this data, RANSAC does not reliably perform well. CRF matching, on the other hand does not require tuning of parameters.

The main parameter in tuning RANSAC is the normalised distance threshold before a point is considered an outlier; in order to evaluate the sensitivity of RANSAC's performance to this distance, we repeated the tests at different threshold values. The results demonstrate that there is some variation in the performance of RANSAC with different thresholds, and thus some parameter tuning would be necessary. However, RANSAC seems to reach a pick of performance at approximately 0.58 for all values above 0.05. We averaged the results of 30 runs for these experiments, as the results with 10 runs exhibited too much noise to discern a pattern.

4.2 Outdoor Dataset

CRF-Matching was also tested in an outdoor, urban dataset. The dataset was collected with a Pioneer 2 AT equipped with a colour camera while navigating at a university campus. 30 pairs of images were selected and manually labelled. The images have resolution of 320×240 pixels and contain significant changes in illumination and viewpoint, along with occlusions and blurring.

Since the outdoor dataset has much more texture than the indoor, the number of SIFT features detected per image was around 500. This makes the process of manual labelling very time consuming in practice. To overcome this issue, we reduce the number of features detected by using the SIFT match criterion i.e., a feature is only included in the set if $\frac{d_2}{d_1} > t$, where t is a threshold and d_1 and d_2 are the Euclidean distances to the first and second nearest neighbours respectively. Note that the threshold is used only to reduce the number of features and does not need to be precisely determined.

An illustrative example is shown in Figure 2. Figure 2(a) shows the image pair. Note the image on the right is blurred as a result of the robot movement. The result of the standard SIFT match algorithm is shown in Figure 2(b). Note that there is one extraneous match associating the bottom of a tree with the top of a car. Figure 2(c) shows the ground-truth obtained with manual labelling. The RANSAC result is in FIgure 2(d). The incorrect match still remains and RANSAC eliminated three correct matches. Figure 2(e) shows the result from CRF-Matching using MAP inference. The matches are almost identical to the ground-truth except by one extra outlier.

Information retrieval metrics for the outdoor dataset are presented in Table 1. As, in general, there are more features per image than with the indoor dataset, CRF-Matching outperforms RANSAC but not by a large margin. Note, however, that results from RANSAC are variable due to the sampling nature of the algorithm and does require the definition of a threshold. Additionally, as in many outdoor applications, image feature matching can be used for localisation where uncertainty estimation is crucial. As a probabilistic approach, CRF-Matching returns a probabilistic distribution over the space of possible matches which can be used in a standard filter.

5 CONCLUSION AND FUTURE WORK

This paper presented a probabilistic network to perform image feature data association. Unlike other approaches where thresholds need to be manually specified, the proposed framework can learn parameters from data through a statistical learning procedure. Image feature association is performed as a joint probabilistic inference where spatial constraints are taken into account to ensure consistency. We demonstrate how the Delaunay triangulation can be used to build the graph structure of the CRF-Matching model which obeys geometric constraints. The experimental results reported indicate that CRF-Matching can be an interesting alternative to RANSAC for challenging problems usually encountered in practical robotics applications.

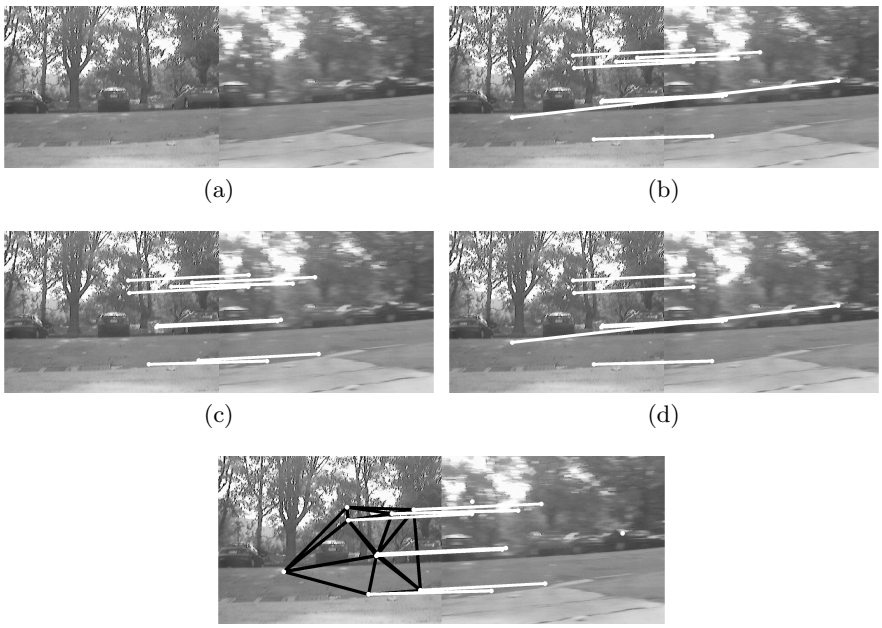

Fig. 2. A sequence of images showing the different results of the algorithms. Figure (a) shows a typical image pair with no matches. Note that while there is some texture, there is not much. Figure (b) shows the result of applying SIFT matching, with no subsequent filtering. Figure (c) shows hand-labelled results used for training and evaluating the matches. Figure (d) shows the results using RANSAC. Figure (e) shows the results of CRF-Matching and the graphical model.

The main caveat of CRF-Matching is the computational complexity of the inference process. In our datasets MAP inference was performed in 0.1 to 2 seconds depending on the number of image features detected. This includes the computation of SIFT features. The code is implemented in Matlab and is executed in a desktop machine. Offline learning takes about 5 minutes in the same machine.

As future work we will investigate alternative ways to yield faster MAP inferences. Two main algorithms will be given special attention: Graph Cuts [4] and the Iterated Conditional Modes [2]. Additionally we plan to extend CRF-Matching to 3D range and camera data registration.

ACKNOWLEDGMENTS

This work is partly supported by the ARC Centres of Excellence programme funded by the Australian Research Council (ARC) and the New South Wales State Government, and by DARPA's ASSIST and CALO Programmes (contract numbers:

NBCH-C-05-0137, SRI subcontract 27-000968). Our thanks also to Raymond Sheh for assistance collecting the rescue dataset.

References

1. J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24, 1975.
2. J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B*, 48:259–302, 1986.
3. P. J. Besl and McKay N. D. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 14(2):239–256, 1992.
4. Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, November 2001.
5. M. Brown and D. G. Lowe. Recognising panoramas. In *Proc. of the International Conference on Computer Vision (ICCV)*, pages 1218–1227, 2003.
6. M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comms. Assoc. Comp. Mach.*, 24(6):381–395, 1981.
7. M. W. Kadous, R. K. Sheh, and C. Sammut. CASTER: a robot for urban search and rescue. In *Proceedings of the 2005 Australasian Conference on Robotics and Automation*, 2005.
8. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the International Conference on Machine Learning (ICML)*, 2001.
9. Computational Geometry Algorithms Library. http://www.cgal.org/.
10. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
11. K. Murphy, Y. Weiss, and M. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999.
12. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., 1988.
13. F. R. Preparata and M. I. Shamos. *Computational Geometry: An Introduction*. Springer-Verlag, New York, USA, 1985.
14. L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*. IEEE, 1989.
15. F. Ramos, D. Fox, and H. Durrant-Whyte. Crf-matching: Conditional random fields for feature-based scan matching. In *Proc. of Robotics: Science and Systems*, 2007.
16. S. Se, D. Lowe, and J. Little. Global localization using distinctive visual features. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3698–3705, 2002.
17. W. Zhang and J. Kosecka. Generalized ransac framework for relaxed correspondence problems. In *Third International Symposium on 3D Data Processing, Visualization*, volume 0, pages 854–860, 2006.