2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Assigning Visual Words to Places for Loop Closure Detection

Konstantinos A. Tsintotas, Loukas Bampis and Antonios Gasteratos

***Abstract*— Place recognition of pre-visited areas, widely known as Loop Closure Detection (LCD), constitutes one of the most important components in robotic applications, where the robot needs to estimate its pose while navigating through the field (e.g., simultaneous localization and mapping). In this paper, we present a novel approach for LCD based on the assignment of Visual Words (VWs) to particular places of the traversed path. The system operates in real time and does not require any pre-training procedure, such as visual vocabulary construction or descriptor-space dimensionality reduction. A place is defined through a dynamic segmentation of the incoming image stream and is assigned with VWs through the usage of an on-line clustering algorithm. At query time, image descriptors are converted into VWs on the map accumulating votes to the corresponding places. By means of a probability function, the mechanism is capable of identifying a loop closing candidate place. A nearest neighbor voting scheme on the descriptors' space allows the system to select the most appropriate image match at the chosen place. Geometrical and temporal consistency checks are applied on the proposed loop closing pair increasing the system's performance. Evaluation took place on several publicly available and challenging datasets offering high precision and recall scores as compared to other state-of-the-art approaches.**

I. INTRODUCTION

While a robot crosses a route, it is possible to misinterpret its trajectory due to field abnormalities, enviromental conditions or mis-accurate sensor measures. Such cases produce drifts affecting the incrementally constructed map, which introduces a great risk in completing robot's mission. The ability of the robot to localize itself and map its surroundings, widely known as Simultaneous Localization and Mapping (SLAM) [1], [2], is strengthened owing to visual place recognition functionality and more specifically the identification of pre-visited areas, also known as Loop Closure Detection (LCD) [3]. Accurate LCDs methods offer precise pose estimation and improved system performance.

In recent autonomous systems with increased computational powers, cameras have become the primary sensor modules for appearance-based place recognition due to the rich information they provide [6]. The acquired images are processed to detect keypoints described by methods such as SIFT [7], SURF [5], or binary equivalents like BRISK [8] or ORB [9]. Using a clustering procedure on a training sample of local features many LCD mechanisms quantize the descriptor space into Visual Words (VWs). This framework, known as Bag-of-Words (BoW) [10], is originated in text retrieval techniques [11]. According to the way by which the Visual Vocabulary (VV) is constructed, BoW models can be distinguished into two main categories, namely off-line and on-line ones. A popular quantization technique, which falls into the off-line approaches is k-means clustering, where k denotes the number of clusters and consequently the VV's size. When the incoming image enters the pipeline a VW histogram, of equal size to the VV, is created based on the widely used "term frequency-inverse document frequency" (tf-idf) technique [10]. Comparing the VW histograms of the query and the database image yields the most appropriate match. Besides the descriptor space discretization, many LCD approaches are based on voting techniques for place recognition. Utilizing the votes' aggregation density into a dimensionally reduced descriptor space [12], [13], pre-visited areas can be identified.

In this paper, we present an efficient image-to-sequence appearance-based LCD method. Using a voting scheme over the on-line generated VWs and coupling the method with a probability function, our approach is able to accurately detect revisited places. A dynamic sequence segmentation, performed on the incoming image stream formulates "places" on the robot's navigated path. Subsequently, the accumulated

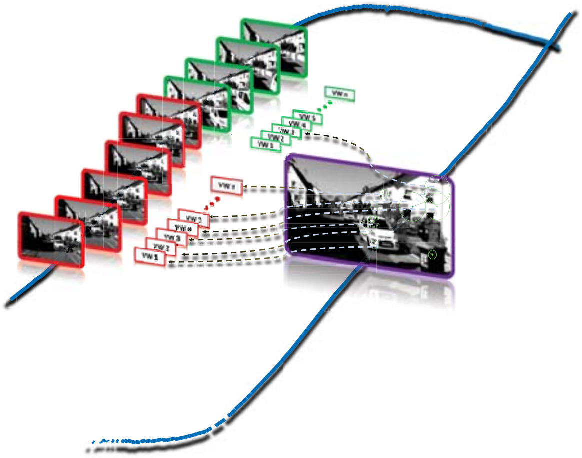

Fig. 1: A representation of the proposed loop closure detection method tested on KITTI 05 [4] dataset. Red and green outlined frames indicate different image sequences (places) the robot constructed during its autonomous mission. Each place contains a different and unique set of Visual Words (VWs). At query time, SURF [5] local feature descriptors are converted into their nearest neighboring VWs, while the votes' aggregation (density of dashed arrows) indicates an image-to-place match.

Authors are with the Department of Production and Management Engineering, Democritus University of Thrace, 12 Vas. Sophias, GR-671 32, Xanthi, Greece ktsintot@pme.duth.gr, lbampis@pme.duth.gr, agaster@pme.duth.gr

978-1-5386-3081-5/18/$31.00 ©2018 IEEE

local feature descriptors are processed by a Growing Neural Gas (GNG) [14] clustering mechanism for the corresponding VWs generation. When new query images enter to the pipeline, the extracted descriptors assign votes to the database sequences including their nearest neighboring VWs (Fig. 1). The system uses a binomial probability density function to locate the proper candidate place and a nearest descriptor neighbor technique to identify image-to-image associations within the selected loop closing sequence. In addition, temporal and geometrical consistency checks are performed between the query and candidate image, providing a higher level of discrimination. The main contributions offered by the paper in hand are summarized as follows:

- An image-to-sequence LCD pipeline with robust results and low computational complexity, evaluated on six different challenging environments.
- A dynamic sequence segmentation method, where places are constructed through a feature matching criterion. During this procedure, we are not interested in the semantics of the environment, but rather in grouping camera measurements which contain common visual information.
- A novel "VWs to places assignment" capable of identifying pre-visited locations only by converting the query's local descriptors into database's VWs.

Using a pre-trained vocabulary can possibly lead to an inaccurate discretization of the descriptor space, especially when the robot traverses places with inconsistent to the vocabulary's training set visual information. Likewise, in cases where the descriptors' space is dimensionally reduced, a training procedure is required. To this end, the proposed algorithm achieves a standalone and "anytime-anywhere" ready system by adopting an on-line VV formulation of the observed world. The rest of the paper is organized as follows. In Section II, a brief discussion on the appearance-based LCD methods is provided. In Section III, the proposed algorithm is described in detail, while Section IV presents an evaluation of our method with comparative results. Conclusions and future work are discussed in Section V.

II. RELATED WORK

The off-line appearance-based method proposed in [15] utilizes the BoW model for image representation, while a Chow Liu tree [16] learns the co-visibility probabilities between VWs' occurrences. In a similar approach [17], a binary pre-trained VV is used, accompanied by a geometrical verification step for further performance enhancement. Since the above methods focus on image-to-image associations, in our recent work [18] we tackle the LCD problem by comparing sequences of images instead of single instances. The incoming data are segmented into fix-sized groups of images where each sequence is represented by a common VW histogram. Finally, using a quantitative interpretation of temporal consistency, sequence-to-sequence matches coherently advancing along time are enhanced.

Nicosevici and Garcia [19] proposed the formulation of an on-line VV by using an agglomerative clustering mechanism. A scalable and dynamic VV was created free from any restriction presented in traditional methods (e.g., number of clusters in k-means). Furthermore, Angeli et al. [20] costructed two parallel VVs (one from local color histograms and another from SIFT descriptors) using distance thresholds as merging criterion, while the final matching probability was estimated through a Bayesian filter. In addition, Khan and Wollherr [21] proposed an on-line incremental formulation of a binary VV, by tracking features between consecutive instances. Loop closures were detected using a likelihood function and temporal consistency checks.

In methods avert from BoW model, such in [12] and [13], LCD is achieved by adopting voting techniques. During a pre-processing stage, the descriptors' space is dimensionally reduced via PCA [22], while a classification mechanism, such as K-Nearest Neighbor (K-NN), assigns votes to the database images. In [12], the matching functionality between query's local descriptors and databases' ones, was implemented by a k-d tree [23], where the regions of high vote density are selected as loop closure candidates. Likewise, the authors in [13] used a probabilistic score in order to estimate the similarity between images. Temporal and geometrical verification checks were also applied for performance enhancement. Our approach differs from the aforementioned ones due to the fact that we adopt an image-to-sequence inference scheme, rather than querying against the entire image database. In addition, our method does not require any training process.

As a final note, recently appeared methods in the visual place recognition field, such as [24] and [25], utilized Convolution Neural Networks (CNNs) initially trained for object recognition. The output of specific convolutional layers are treated as image descriptors and revisited places are determined by measuring distance metrics amongst them. Despite their high performances, CNNs are viewpoint dependent, while the topological information is not provided in the higher networks' levels [26], [27]. Thus, they are still disconnected from the overall SLAM architecture and LCD, which is the main target of the proposed work.

III. METHODOLOGY

In this section an extended description of the proposed LCD pipeline is presented. As mentioned, the algorithm describes each place by a set of VWs formulated on-line. As a first step, images exhibiting time and content proximity are congregated, resulting in a sequence of images which determines a place. Subsequently, GNG clustering is performed over the sequence's accumulated descriptors, generating the corresponding VWs. In the course of a query, local features from the most recently acquired frame are extracted and converted to their most similar VWs in the trajectory. Each descriptor-to-VW conversion corresponds to a new vote for the sequence. The place which aggregates the most VW votes is determined through a binomial probability function. A nearest neighbor technique on the descriptors' space indicates the most similar image in the particular sequence. Finally, the chosen instance is propagated to geometrical and

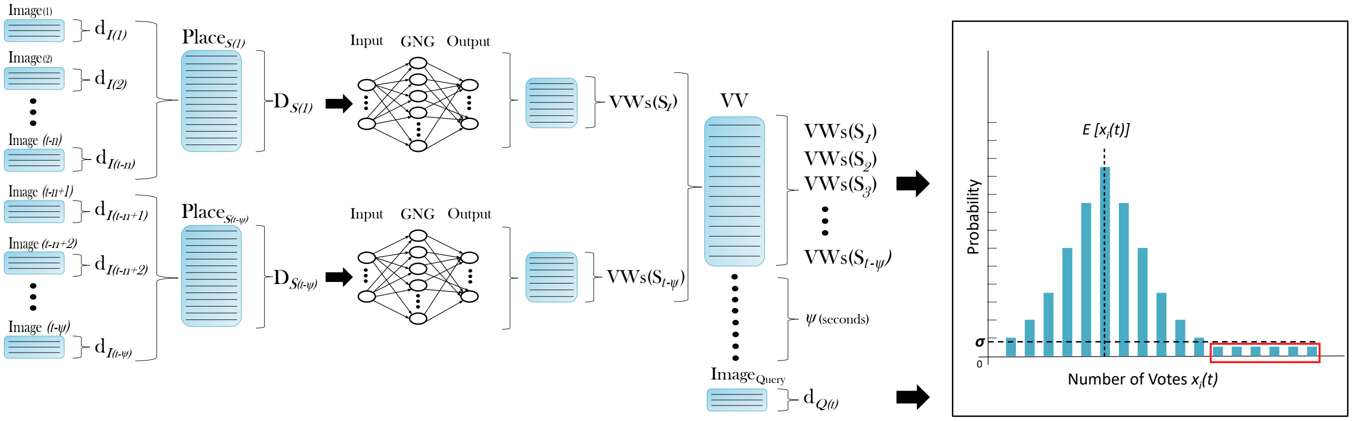

Fig. 2: An overview of the proposed loop closure detection method. As the incomming image stream enters the pipeline, dynamic "places" ($S_{(t)}$) are formulated through a feature matching technique (left). Subsequently, the accumulated local feature descriptors ($D_{S(t)}$) are fed into the Growing Neural Gas (GNG) [14] clustering mechanism, where the corresponding place's Visual Words (VWs(S_i)) are generated (center). At query time, image descriptors ($d_{Q(t)}$) are converted into their nearest neighboring VW from the on-line created Visual Vocabulary (VV). During conversion stage votes are shared among the places, where the VWs belong. Finally, through a binomial distribution function, candidate sequences are located according to their vote density (right). The highlighted red area indicates the density rarity a loop closure event would produce.

temporal consistency checks, in order to be accepted as a loop closure match. An outline of the proposed algorithm is shown in Fig. 2.

A. Places Formulation

The proposed place recognition system operates in a pipeline fashion; the incoming data being the image stream. For each instance I entering to the system, the ν most prominent SURF keypoints are detected. As the robot navigates through the field, some of the incoming camera measurements may be unable to produce enough visual information, e.g., the observation of a black plain. To avoid the creation of inconsistent places, images that contain less than ξ keypoints are rejected. During the on-line operation, the projected descriptor space is constantly updated by the detected feature vectors d_I. Instead of reducing the descriptors' dimensionality, the proposed algorithm utilizes the full SURF space.

During the course of the procedure, new places S are determined through a feature matching coherence check. More specifically, at time t, the incoming image stream $I(t-n), ..., I(t-2), I(t-1), I(t)$ is segmented when the correlation between the last n images's descriptors cease to exist:

$$\left| \bigcap_{i=0}^{i=n} d_{I(t-i)} \right| \leq 1 \tag{1}$$

where $|\mathbb{X}|$ denotes the cardinality of set $\mathbb{X}$. A descriptors database D_S is also retained for each place, via:

$$D_S = \bigcup_{i=0}^{i=n} d_{I(t-i)} \tag{2}$$

B. Representation of Places by Visual Words

In order to assign VWs to places local descriptors D_S are utlized as input to the GNG clustering algorithm. In contrast to other popular clustering methods, where the number of clusters is predefined, the GNG incrementally adds new nodes (VWs) until an error-minimization criterion is met. Since the proposed approach uses the GNG mechanism for quantizing the feature vectors, its main parameterization remains the same as originally implemented in [14]. The maximum allowed set of VWs (α), which will be created by the GNG, is defined as being equal to the images' extracted features ν. This analogy is chosen so as to provide a direct correspondence between VWs and image features. Thus, a new VW is generated when a frequency criterion φ is met, defined as the ratio between the maximum number of VWs per place and the mean of sequences' length μ ($\varphi = \nu / mean(\mu)$). Since the system aims for low computational complexity, the GNG iterations (ε) are selected to the lowest permissible. Finally, a VV database is retained during the procedure consisting of the generated VWs:

$$VV_{db} = \bigcup_{i=1}^{i=t} VWs(S_i) \tag{3}$$

where the term S_t is the latest formulated place in the trajectory. An inverted indexing list [17] is also maintained along the VV providing faster image-to-sequence associations during the inference procedure.

C. Query to Place Assignment

Given a query image I_t^Q, a searching procedure is performed among the produced places in order to detect loop closure candidates. As opposed to the most BoW-based systems, where VW histogram comparison techniques are used, the proposed approach utilizes a voting scheme. A nearest neighbor technique projects the query's local features to the generated VWs belonging to VV_{db}. During the conversion of the query's feature descriptors, votes are assigned to places

TABLE I: Datasets' synopsis

Dataset	Description	Camera Position	Image Resolution	# Images	Frames Per Second
KITTI 00 [4]	Outdoor, dynamic	Frontal	1241 × 376	4551	10
KITTI 05 [4]	Outdoor, dynamic	Frontal	1241 × 376	2761	10
Biccoca 2009-02-25b [28]	Indoor, static	Frontal	640 × 480	26335	15
Malaga 2009 Parking 6L [29]	Outdoor, slightly dynamic	Frontal	1024 × 768	3474	7
New College [30]	Outdoor, dynamic	Frontal	512 × 384	52480	20
Euroc MH 05 [31]	Indoor, static	Frontal	752 × 480	2773	20

in accordance with the VWs' origin. The vote density $x_i(t)$ of each place i constitutes the factor that determines the probabilistic similarity score.

In cases where the robot's velocity decreases or the system remains still, it is possible for the query and database sets to observe the same scene. In such a scenario, the incoming camera measurements exhibit a strong spatial relationship resulting in false-positive LCDs. Yet, the equally strong temporal relationship of their posses dictates that such a loop closure is erroneous. A searching area (VV_{sa}) that rejects recently acquired input frames is defined based on a temporal constant ψ:

$$VV_{sa} = VV_{db} \cap [VWs(S_1), VWs(S_{t-\psi})] \tag{4}$$

When the voting procedure is completed, a binomial probability function [13] is employed to check the potential loop closuring places. If the robot visits a new sight (never encountered before), the voting procedure shall be random meaning that the votes density for each place in VV_{sa} would be low. Accordingly, when the robot navigates through a revisited area, the vote density of the specific place should be high. Based on the binomial distribution function's properties, the later case corresponds to a low probability event. Such instances are interpreted as loop closure candidates by the proposed system:

$$X_i(t) \sim Bin(n,p), n = N(t), p = \frac{\lambda}{\Lambda(t)} \tag{5}$$

$$N = \sum d_{Q(t)} \tag{6}$$

$$\lambda = VWs(S_i) \tag{7}$$

$$\Lambda = \sum VW(S_1) \sim VW(S_{t-\psi}) \tag{8}$$

where N denotes the multitude of query's VWs ($d_{Q(t)}$), λ corresponds to place's i VWs and Λ is the sum of VWs within the searching area VV_{sa}. The probability score is calculated for each place, while two conditions have to be satisfied before a place is recognized as pre-visited. The score has to satisfy a threshold value σ:

$$Pr(X_i(t) = x_i(t)) < \sigma < 1 \tag{9}$$

Additionally, the number of accumulated VWs for a specific place needs to be greater than the distribution's extended value:

$$x_i(t) > E[X_i(t)] \tag{10}$$

Equation 10 discards the cases where fewer votes are collected than a random voting.

D. Image to Image Association

Up to this point, our algorithm is capable of identifying a pre-visited location in the traversed map. As a final step, an image-to-image correlation is performed between the query image and the most similar member of the selected place S_m in the database. Based on a K-NN classifier ($K = 1$), the query's descriptors d_Q are matched with the ones ($D_{S_{(m)}}$) belonging to $S_{(m)}$. The image (I^S) which gathers the most matches is considered as loop closure candidate and is kept for further validation.

In order to avoid false positive loop closure matches, the pair of I_t^Q and I^S is subjected to a geometrical consistency check. Using a RANSAC-based scheme, a fundamental matrix T is estimated between the query and the proposed image. If the computation of T fails or the number of inlier points between the two images is less than a factor τ, the candidate instance is ignored. The parameterization of the applied RANSAC method follows the one in [17]. Finally, with a view to accept a matching pair, the method incorporates a temporal consistency check among the last β input frames. More specifically, a LCD is accepted when the aforementioned conditions are met for β consecutive camera measurements.

TABLE II: Parameter list

Minimum detected local features per image, ξ	:	5
Maximum prominent local features per image, ν Maximum generated visual words per place, α	:	300
Visual words' generation frequency, φ	:	25
Growing Neural Gas iterations, ε	:	1
Search area time constant, ψ	:	40 secs
Geometrical verification inliers, τ	:	12 [17]
Images' temporal consistency, β	:	2
Probability score threshold, σ	:	10^{-12}

IV. EXPERIMENTAL EVALUATION

The following section provides a brief description of the experimental procedure, an expansive evaluation of the proposed algorithm, as well as comparative results. The method is tested with six publicly available datasets, including indoor and outdoor environments, as shown in Table I. Comparisons performed against several state-of-the-art approaches, such as FAB-MAP 2.0 [15], IBuILD [21], Gálvez-López et. al. [17], Gehrig et. al. [13], Bampis et. al. [18].

A. Procedure

1) Datasets: Publicly available datasets are selected so that a variety of different camera measurement properties, e.g., image's resolution, robot's velocity and frame rate (FPS), can be achieved. In the cases of KITTI 00 [4], KITTI 05 [4] and Malaga 2009 Parking 6L (Malaga6L) [29] datasets, the incoming instances are obtained by a camera mounted on a moving car. Biccoca 2009-02-25b (Biccoca) [28] and New College [30] have been recorded by means of the vision stystem of a wheeled robotic platform, while in Euroc MH 05 [31] the incoming image stream is retrieved by sensors mounted on a hex-rotor helicopter. Most of the aforementioned datasets contain stereo information, though for the purposes of this evaluation only the left camera image stream is used.

KITTI 00 and KITTI 05 are preferred as outdoor datasets, since the recorded data provide considerable loop closure examples, accurate odometry and high resolution information (image's size and vehicle's velocity). The Biccoca and Euroc MH 05 are selected as indoor datasets, due to many strong perceptual aliasing examples presented in their trajectory. New College and Malaga6L are chosen for evaluation as they contain a wealth of visual information, as shown in Table I.

2) Ground Truth: A binary matrix, whose rows and columns correspond to images at different timestamps and its elements denote the presence $(GT_{ij} = 1)$ or absense $(GT_{ij} = 0)$ of a loop closure event, is defined as Ground Truth (GT) for each dataset. For the Biccoca, Malaga6L and New College instances, GT is provided by the authors in [17]. KITTI 00, KITTI 05 and Euroc MH 05 do not provide such information therefore, we manually constructed the corresponding GTs by considering the odometry information.

3) Precision-Recall: In order to compare the performance between different algorithms, the precision and recall metrics are used. Precision is defined as the number of true positive LCDs over the total systems identifications. Recall is the ratio between the true positive detections and the actual total of loop closure events defined by the GT.

B. Method Evaluation

The selected method's parameters are summarized in Table II. During the experiments, the average sequences' length μ was observed to be approximately 12 for all tested datasets.

In Fig. 3, we evaluate the effect of the local features' number preserved for every image. As shown by the precision and recall curves in cases where the extracted local features are less than 400, the system's achieved performance is resembling. As the number of accepted features increases, it is observed that the recall rate (corresponding to 100% precision) is decreasing. This is owed to the fact that weak features detected during the robot's first visit to a certain location are mainly noisy. Thus, it is less probable for them to be matched in the course of a loop closure events.

In Fig. 4, the effect of GNG iterations in execution time is evaluated in a dataset containing 2.5k images[1]. As illustrated by the red curve, the system's performance is slightly improved when the number of iterations increases.

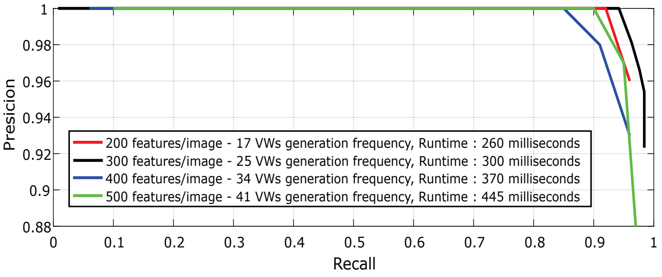

Fig. 3: Precision and recall curves evaluating the utilized local features ν per image. This corresponds to the number of VWs generated in each place (α). 300 features per image provide better recall rate, while the execution time remains low. Experiments are performed on the KITTI 05 [4] dataset.

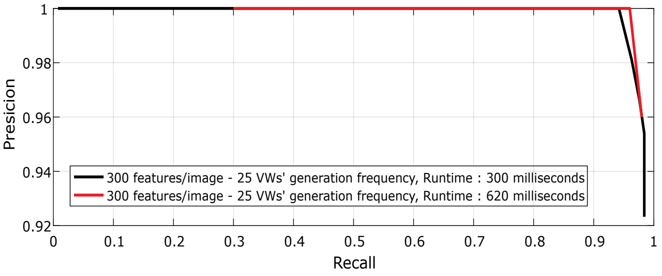

Fig. 4: Precision and recall curves evaluating the execution time of the prososed algorithm against the Growing Neural Gas (GNG) iterations. While the second iteration (red line) doubles the complexity, the recall rate (for 100% precision) indicates a similar performance to the first one (black line). Experiments are implemented on the KITTI 05 [4] dataset.

Moreover, the execution time is raising by a factor of two from the first to the second iteration (from 300 ms to 620 ms). Bearing that in mind, a small percentage of recall is sacrificed for a faster implementation capable of being utilized into large scale datasets (e.g., New College).

The searching offset ψ is selected to avoid LCDs with strong spatiotemporal relationship, as in the cases of KITTI 05 and New College. Moreover, given the chosen time constraint, the system is capable of including early pre-visited places in the navigated path, as demonstrated in the case of Biccoca.

The system's overall performance is presented in Fig. 5. Precision and recall curves are generated by selecting several different decision thresholds σ until false positives matches are eliminated. In order to evaluate the impact of our method, the proposed parameters were fixed in all tested scenarios.

C. Comparative Results

Table III compares the precision and recall metrics of the proposed method against the aforementioned state-of-the-art approaches. The cited methods' performance are obtained by the respective papers. It is notable that the proposed algorithm can achieve high recall rates for 100% precision in every tested environment. In outdoor datasets, such as KITTI 00, KITTI 05, New College and Magala6L, the system exhibits an improved performance with over 90% recall

[1] A Matlab-based implementation of the proposed approach was tested on a quad-core 2.6GHz system with 8GB RAM.

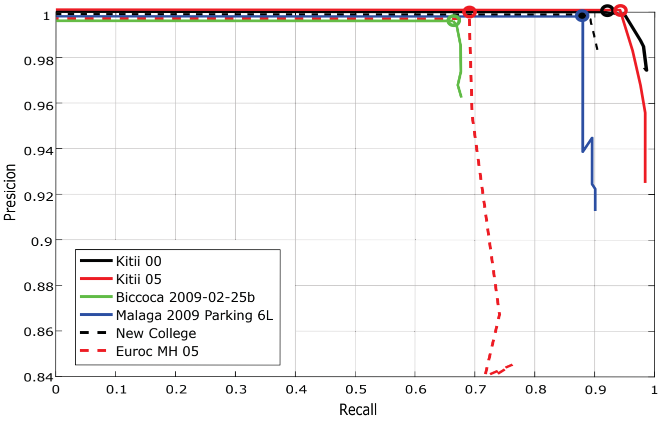

Fig. 5: Precision recall curves for the proposed approach. Color markers (cycles) on the top of the graphs highlight the highest recall for 100% precision. The used parameter setup is summarized in Table II.

in KITTI datasets and almost 90% in New College and Malaga6L. In Biccoca the method performs unfavorably, comparing to the rest of the approaches, the reason being the platform's low velocity (the robot navigates in areas with strong perceptual aliasing for a long period of time). In the case of Euroc MH 05, the system encounters many abruet velocity variations along the trajectory, resulting into a recall rate similar to [13]. Figure 6 demonstrates the LCDs provided by the system on to of the robot's trajectories.

TABLE III: Comparative results

Dataset	Approachs	Precision (%)	Recall (%)
KITTI 00 [4]	Gehrig et al. [13]	100	92
	Bampis et al. [18]	100	81.54
	Proposed	**100**	**93.18**
KITTI 05 [4]	Gehrig et al. [13]	100	94
	Bampis et al. [18]	100	84.80
	Proposed	**100**	**94.20**
Biccoca [28]	Gálvez-López et al. [17]	100	81.20
	Bampis et al. [18]	100	78.10
	Proposed	**100**	**66.19**
Malaga6L [29]	Gálvez-López et al. [17]	100	74.75
	FAB-MAP 2.0 [15]	100	68.52
	Bampis et al. [18]	100	76.78
	IBuILD [21]	100	78.13
	Proposed	**100**	**87.99**
New College [30]	Gálvez-López et al. [17]	100	55.92
	Bampis et al. [18]	100	77.55
	Proposed	**100**	**87.97**
Euroc MH 05 [31]	Gehrig et al. [13]	100	71
	Proposed	**100**	**69.21**

V. CONCLUSIONS

In this paper, an online image-to-sequence probabilistic voting framework has been presented. The mechanism achieved each place's description by unique VWs, while the binomial probability function provides the final loop closure decisions. As the incoming data arrive to the system, a feature matching technique undertakes the image stream segmentation for constructing new places. Through a GNG clustering algorithm, the feature-members of each segmented sequence are converted into VWs. A probabilistic voting scheme between the query image and the database places is

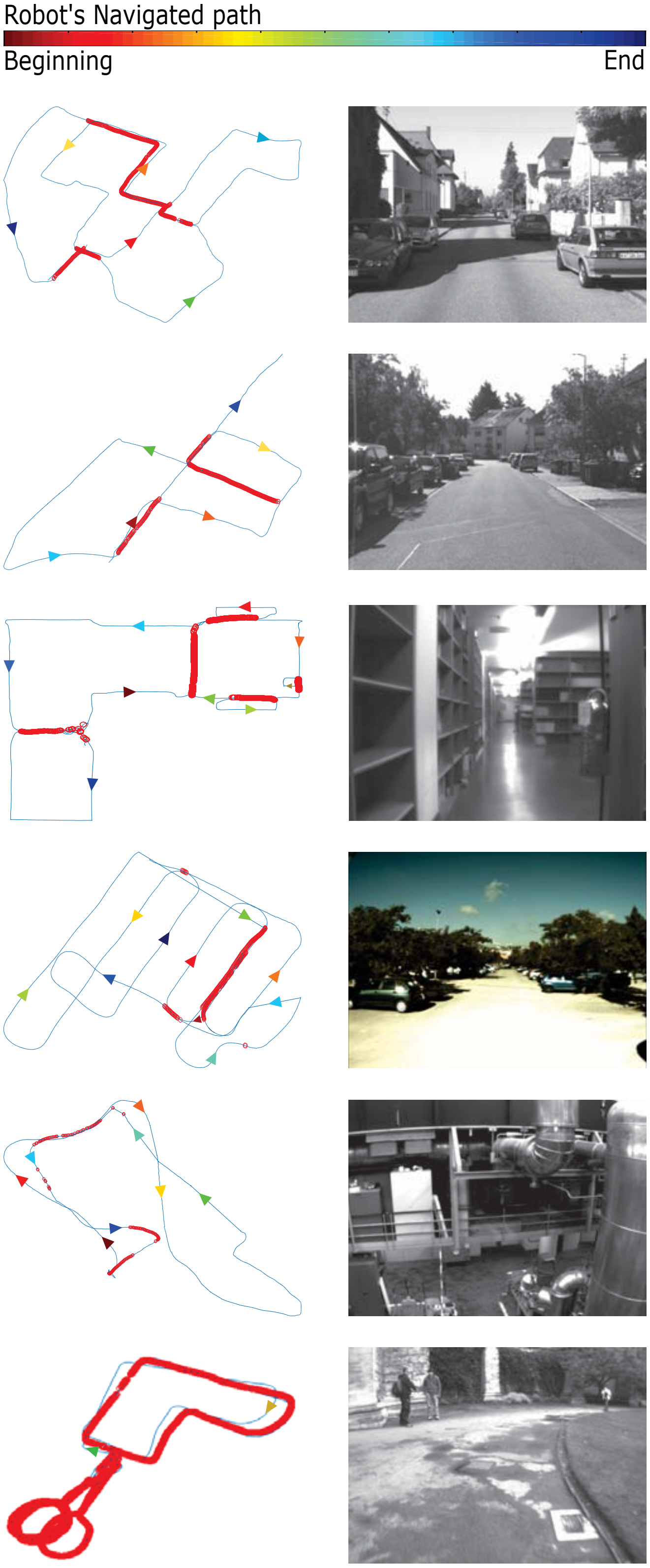

Fig. 6: Datasets' trajectories (left) and example images (right). From top to bottom: KITTI 00 [4], KITTI 05 [4], Biccoca 2009-02-25b [28], Malaga 2009 Parking 6L [29], Euroc MH 05 [31], New College [30]. Red circles indicate the system's loop closure detections, while colored arrows illustrate the temporal evolution according to the color-bar.

performed for loop closure candidate identifications. Finally, a nearest neighbor technique is used for image-to-image matching. The method is independent of any prior knowledge of the working environment demonstrated by its ability to perform robustly in six different datasets (indoor, outdoor, static and dynamic). As compared to several state-of-the-art approaches, the proposed algorithm offers high recall rates for 100% precision in most of the evaluated datasets. As future work the authors plan to evaluate the proposed method within a well-established SLAM architecture [32].

ACKNOWLEDGMENT

The authors express their appreciation to Dorian Gálvez-López for kindly offering GT information for many of the used datasets. The paper was partially supported by project ETAA, DUTH Research Committee 81328.

REFERENCES

[1] H. Durrant-Whyte and T. Bailey, "Simultaneous Localization and Mapping: part I," *IEEE Robotics & Automation Magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[2] S. Thrun and J. J. Leonard, "Simultaneous Localization and Mapping," in *Springer Handbook of Robotics*, 2008, pp. 871–889.

[3] K. L. Ho and P. Newman, "Loop closure detection in SLAM by combining visual and spatial appearance," *Robotics and Autonomous Systems*, vol. 54, no. 9, pp. 740–749, 2006.

[4] J. Fritsch, T. Kuehnl, and A. Geiger, "A New Performance Measure and Evaluation Benchmark for Road Detection Algorithms," in *Proc. IEEE International Conference on Intelligent Transportation Systems*, 2013, pp. 1693–1700.

[5] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Proc. European Conference on Computer Vision*, 2006, pp. 404–417.

[6] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual Place Recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[7] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[8] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary Robust Invariant Scalable Keypoints," in *Proc. IEEE International Conference on Computer Vision*, 2011, pp. 2548–2555.

[9] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," in *Proc. IEEE International Confernece on Computer Vision*, 2011, pp. 2564–2571.

[10] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," in *Proc. IEEE International Confernece on Computer Vision*, 2003, p. 1470.

[11] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern Information Retrieval*, 1999, vol. 463.

[12] S. Lynen, M. Bosse, P. Furgale, and R. Siegwart, "Placeless place-recognition," in *Proc. IEEE International Conference on 3D Vision*, 2014, pp. 303–310.

[13] M. Gehrig, E. Stumm, T. Hinzmann, and R. Siegwart, "Visual Place Recognition with Probabilistic Vertex Voting," in *Proc. IEEE International Conference on Robotics and Automation*, 2017, pp. 3192–3199.

[17] D. Gálvez-López and J. D. Tardos, "Bags of Binary Words for Fast Place Recognition in Image Sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[14] B. Fritzke, "A Growing Neural Gas Network Learns Topologies," in *Proc. Advances in neural information processing systems*, 1995, pp. 625–632.

[15] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[16] C. Chow and C. Liu, "Approximating Discrete Probability Distributions with Dependence Trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[18] L. Bampis, A. Amanatiadis, and A. Gasteratos, "Encoding the Description of Image Sequences: A Two-Layered Pipeline for Loop Closure Detection," in *Proc. IEEE International Conference on Intelligent Robots and Systems*, 2016, pp. 4530–4536.

[19] T. Nicosevici and R. Garcia, "Automatic Visual Bag-of-Words for Online Robot Navigation and Mapping," *IEEE Transactions on Robotics*, vol. 28, no. 4, pp. 886–898, 2012.

[20] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "Fast and Incremental Method for Loop-Closure Detection Using Bags of Visual Words," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1027–1037, 2008.

[21] S. Khan and D. Wollherr, "IBuILD: Incremental Bag of Binary Words for Appearance Based Loop Closure Detection," in *Proc. IEEE International Conference on Robotics and Automation*, 2015, pp. 5441–5447.

[22] S. Wold, K. Esbensen, and P. Geladi, "Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[23] J. L. Bentley, "Multidimensional Binary Search Trees Used for Associative Searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[24] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the Performance of ConvNet Features for Place Recognition," in *Proc. IEEE International Conference on Intelligent Robots and Systems*, 2015, pp. 4297–4304.

[25] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera, "Fusion and Binarization of CNN Features for Robust Topological Localization across Seasons," in *Proc. IEEE International Conference on Intelligent Robots and Systems*, 2016, pp. 4656–4663.

[26] E. Sizikova, V. K. Singh, B. Georgescu, M. Halber, K. Ma, and T. Chen, "Enhancing Place Recognition using Joint Intensity - Depth Analysis and Synthetic Data," in *Proc. European Conference on Computer Vision*, 2016, pp. 901–908.

[27] X. Fei, K. Tsotsos, and S. Soatto, "A Simple Hierarchical Pooling Data Structure for Loop Closure," in *Proc. European Conference on Computer Vision*, 2016, pp. 321–337.

[28] RAWSEEDS. (2007-2009) Robotics Advancement through Web-publishing of Sensorial and Elaborated Extensive Data Sets (Project FP6-IST-045144). [Online]. Available: http://www.rawseeds.org/rs/datasets

[29] J.-L. Blanco, F.-A. Moreno, and J. Gonzalez, "A Collection of Outdoor Robotic Datasets with centimeter-accuracy Ground Truth," *Autonomous Robots*, vol. 27, no. 4, pp. 327–351, 2009.

[30] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman, "The New College Vision and Laser Data Set," *The International Journal of Robotics Research*, vol. 28, no. 5, pp. 595–599, 2009.

[31] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[32] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.