Adaptive Appearance Based Loop-Closing in Heterogeneous Environments

András Majdik, Dorian Gálvez-López, Gheorghe Lazea and José A. Castellanos

***Abstract*— The work described in this paper concerns the problem of detecting loop-closure situations whenever an autonomous vehicle returns to previously visited places in the navigation area. An appearance-based perspective is considered by using images gathered by the on-board vision sensors for navigation tasks in heterogeneous environments characterized by the presence of buildings and urban furniture together with pedestrians and different types of vegetation. We propose a novel probabilistic on-line weight updating algorithm for the bag-of-words description of the gathered images which takes into account both prior knowledge derived from an off-line learning stage and the accuracy of the decisions taken by the algorithm along time. An intuitive measure of the ability of a certain word to contribute to the detection of a correct loop-closure is presented. The proposed strategy is extensively tested using well-known datasets obtained from challenging large-scale environments which emphasize the large improvement on its performance over previously reported works in the literature.**

I. Introduction

In recent years the mobile robot navigation algorithms based on image processing are getting more and more popular, because of various reasons: the relatively low cost of the vision sensors; the increase of processing speed, permitting a fast processing of the large amount of information; the intense development of the image processing algorithms; etc.

In [1] an image retrieval algorithm was presented, which was able to identify efficiently if a test image depicts an object already in the image database or not. The method is based on a hierarchical bag of words representation of the images, which allows a very fast search of an image in the database.

A very efficient visual data management is essential to be able to develop mobile robotic systems operating in large spaces based on visual sensors, because a vast amount of information, captured in various robot positions needs to be processed on board. The adaptation of the bag of words representation made possible the development of visual appearance based simultaneous localization and mapping (SLAM) algorithms.

This work was financially supported by PRODOC (Project for Doctoral Studies Development in Advanced Technologies), Technical University of Cluj-Napoca and by the Ministerio de Educación of Spain (scholarship FPU-AP2008-02272) and by the Spanish project MICINN-FEDER DPI 2009-13710.

András Majdik and Gheorghe Lazea are with the Robotics Research Group, Technical University of Cluj-Napoca, 71-73 Dorobantilor Street, 400609 Cluj-Napoca, Romania {andras.majdik, gheorghe.lazea}@aut.utcluj.ro

Dorian Gálvez-López and José A. Castellanos are with the Instituto de Ingeniería de Aragón, University of Zaragoza, c/ María de Luna 3, 50018 Zaragoza, Spain {dorian, jacaste}@unizar.es

The methods based on visual appearance have been applied successfully in the loop-closing problem [2], [3]. These methods are able to fast match similar images, but they need to be combined with other techniques to overcome some difficulties like perceptual aliasing, which cannot be completely eliminated [4]. The role of the visual appearance module becomes to provide loop-closure candidate positions, which are either verified by an additional visual method as in [5] or by an additional metric SLAM [6].

The major idea behind the work presented in this paper is as follows: if a false loop-closure is detected, then how can this information be used and how can it be propagated back to the visual similarity evaluation algorithm to avoid these false positives in the future, improving thereby the visual appearance recognition system.

In robotic applications better results are obtained if some other knowledge is also introduced in the system. In [2] the results of the visual similarity algorithm are improved by learning the co-appearance probability of the visual words in the scenes. In the current work we try to capture and model the probability that the use of a specific words is leading to correct or false loop-closures.

A similar issue is addressed in [7], where the problem of selecting the useful image features is investigated in image retrieval systems. In any application a very big number of features (or visual words) is detected, but many of these are unreliable or represent irrelevant clutter, these features are only error sources. Furthermore the usefulness of a feature can depend also on the environment and the goal followed in the application. In [8] are considered useful only those features that are detected on several images taken from different views of the same object and also obey a geometrical constraint.

In a mobile robot navigation framework an algorithm is presented in [6], which reduces the feature entropies considered in the computation of the similarity score of those visual words that lead to a false positive loop-closure.

The main contributions of the work presented in the paper over previously reported works are as follows:

- A novel probabilistic on-line weight updating strategy is presented. In the proposed approach an attribute is defined for each visual word representing the ability or skill of that visual word to contribute to correct loop-closure detection decisions.
- The new bag-of-word representation allows the introduction of prior knowledge in the weighting strategy regarding the usefulness of the word in detecting loop-closure situations by autonomous vehicles operating in

heterogeneous environments.

- A dynamic model is developed to update the belief by the observations gathered online, similarly to a reinforcement learning approach, making the weighing algorithm an adaptive one. The skill to contribute to correct loop-closure detection is characterized also by a variance enabling to represent and update the strength of the belief and allowing a flexible weighting strategy.
- The results of extensive experiments made in a large heterogeneous environment characterized by the presence of buildings and urban furniture together with pedestrians and different types of vegetation are presented and discussed.

The visual appearance recognition system in which the probabilistic on-line weight updating strategy is integrated is presented in Fig. 1. The information flow is marked on the diagram between different parts of the robotic system.

The structure of the paper is divided on a theoretical and an experimental part. The emphasis of the theoretical part is on presenting the differences between a standard visual bag-of-word technique (section (II-A, B, C)) in contrast with the proposed one (section (III-A, B, C)). In section (IV) a loop-closure validation algorithm is presented based on stereo image processing. Further on in section (V, VI) the obtained results by applying the novel approach are exhibited. Finally the conclusions and future work are presented (section VII).

II. Standard Bag-Of-Words Algorithm

A. Image Representation

In [9] Sivic et al. originally presented the visual bag-of-words (BoWs in the sequel) technique which represents an image as a numerical vector quantising its salient local features. Their technique entails an off-line stage that performs hierarchical clustering[1] of the image descriptor space, obtaining a set of clusters arranged in a tree with a branch factor k and L depth levels. The set of k^L leaves of the tree conforms the so-called *visual vocabulary*, and each leaf is refered to as a *visual word*. Using BoWs, an image, gathered by the vision sensor, is represented by a vector $\mathbf{w} = (w_1, \ldots, w_j, \ldots, w_N)^T$ of weighting values with $N = k^L$.

B. Weighting Strategy

For each image a set of features is extracted (e.g. 64-D SURF features) and, by descending the previously learned tree structure, their nearest visual words are selected. A weight value w_j is then computed depending on the relevance of the j-th visual word of the visual vocabulary in the gathered image.

The term frequency – inverse document frequency (tf-idf) is a popular weighting method reported in the literature[2], where the concept of document refers to an image in our context. This strategy mathematically formulates the intuition that rare visual words should have a higher importance, and therefore a higher weight in the vector description of an image. Thus, the weight w_j associated to the j-th word is computed as,

$$w_j = \frac{n_{jd}}{n_d} \log\left(\frac{N}{N_j}\right) \tag{1}$$

where, n_{jd} is the number of occurrences of the j-th word in image d, n_d is the total number of words in image d, N represents the number of images and N_j the number of images containing word j.

As described latter this somehow *static* description of a certain image is improved by an on-line algorithm which updates the weighting factors based both on prior knowledge and the accuracy of loop-closure detection.

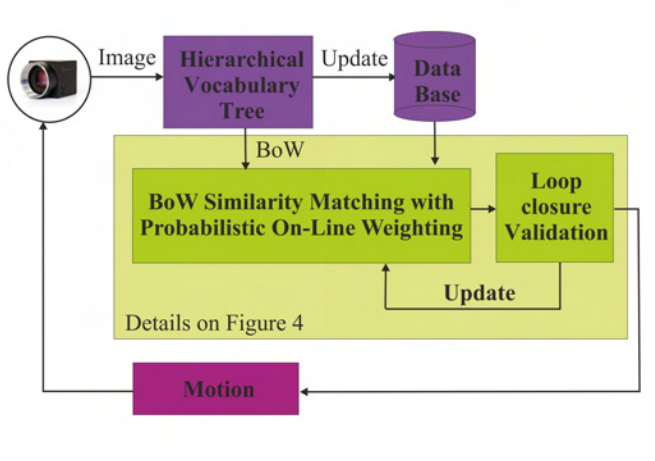

Fig. 1. Information flow diagram of the proposed probabilistic on-line weight updating strategy.

C. Image Similarity Computation

The similarity (or resemblance) between two images, described by the BoWs vectors $\mathbf{v}$ and $\mathbf{w}$ respectively, can be estimated by,

$$s(\mathbf{v}, \mathbf{w}) = 1 - \frac{1}{2}\left\| \frac{\mathbf{v}}{||\mathbf{v}||} - \frac{\mathbf{w}}{||\mathbf{w}||} \right\| \tag{2}$$

where $||.||$ stands for the L_1-norm. The score $s(\mathbf{v}, \mathbf{w})$ tends to zero for very different images and it tends to one for highly similar images.

Therefore, for an appearance-based SLAM algorithm a certain threshold is set on this score to discriminate when the vehicle has returned to previously visited places in the environment.

III. Probabilistic On-Line Weight Updating

This section describes an important contribution of the paper where a probabilistic on-line weight updating strategy is described. Intuitively an attribute is defined for each visual word representing the ability or skill of that visual word to contribute to correct loop-closure detection decisions.

[1] Our implementation of hierarchical clustering is available from `http://webdiis.unizar.es/~dorian` where the *kmeans++* algorithm [5], [10] is used.

[2] See [11] for a comparison of different methods to measure the relevance of a word in a vocabulary.

Fig. 2. Samples from the positively labelled images representing structured, more distinguishable environment.

Fig. 3. Samples from the negatively labelled images representing unstructured, dynamic, misleading environment.

A. Probabilistic Image Representation

Extending the description presented earlier in section (II-A), an image is represented now by a stochastic vector $\mathbf{w} = (w_1, \ldots, w_j, \ldots, w_N)^T$ of weighting values with $N = k^L$, where the mean vector and the covariance matrix of its associated probability density function are given by $\hat{\mathbf{w}}$ and $\mathbf{P_w}$ respectively. In the present work, and aimed at reducing the complexity of the approach, the weights of a certain image description are considered independent with marginal Gaussian distributions, therefore $\mathbf{w}$ is understood as an N-dimensional Gaussian stochastic vector with diagonal covariance matrix $\mathbf{P_w}$.

In this new framework, the weight assigned to the j-th visual word is now computed as,

$$w_j = \alpha_j \frac{n_{jd}}{n_d} \log\left(\frac{N}{N_j}\right) \tag{3}$$

where α_j is a Gaussian random variable with mean $\hat{\alpha}_j$ and variance σ_j^2 and which satisfies $E[\alpha_i \alpha_j] = 0$, $\forall i \neq j$.

Further down in the paper, a dynamic model is provided for each α_j coefficient, initially defined by a certain prior $\alpha_{j,0}$ and subsequently updated at time k from the available knowledge at time $k-1$.

The introduced random variable α_j represents the ability to correctly detect a loop-closing situation, that is, it is interpreted as a probability that the use of the j-th word will lead to a true or false loop-closing detection. This extra weighting parameter is considered a belief factor by the weighting system for every word of the vocabulary. In the sequel the proposed algorithm is largely discussed and in the experimental section the performance of the method is illustrated by extensive results with real data.

B. Computing the Weight's Priors

The training set of images $\mathcal{I}$ used during the off-line learning stage which builds the visual vocabulary mentioned in section (II-A) can be splitted down into two labelled subsets of images, namely the subset of *positive* images $\mathcal{I}_p \subseteq \mathcal{I}$ and the subset of *negative* images $\mathcal{I}_n \subseteq \mathcal{I}$, complementary to the former.

A certain image of the training set is labelled as a positive image when it mainly represents a well-structured urban scene with static objects such as buildings, houses, fix urban furniture, etc. Figure 2 describes some representative positive images from the Zurich building dataset [12]. Conversely, an image is labelled as a negative image when it represents an unstructured scene with dynamic objects such as pedestrians and different types of vegetation. Figure 3 shows some selected examples of negatively labelled images[3].

From a loop-closing detection perspective, negative images could clearly corrupt the results due to scene clutter and perceptual aliasing, and therefore increase the number of false alarms which could drive the estimation of the vehicle location out of coherence and consistency. This intuition could be incorporated into the weight's computation by assigning priors to the random variables α_j defined in eq. (3) of the form,

$$\hat{\alpha}_{j,0} = \frac{C_j}{M_j} \tag{4}$$

where M_j is the total number of occurrence of the j-th word in the image training set from which the vocabulary was built and C_j is the number of its occurrences in positively labelled images. From eq. (4) we observe that the prior tends to zero when its associated word was mainly present in negatively-labelled images and it tends to one in the opposite situation where it mainly appeared in possitively-labelled images, therefore providing a measure of the ability of the given word to contribute to the correct detection of loop-closing situations.

[3]In the video attachment the complete training set that was used to build the hierarchical vocabulary tree in all our experiments is provided. These are isolated images which do not make a trajectory and differ from all of the images collected by the robot during the experiments.

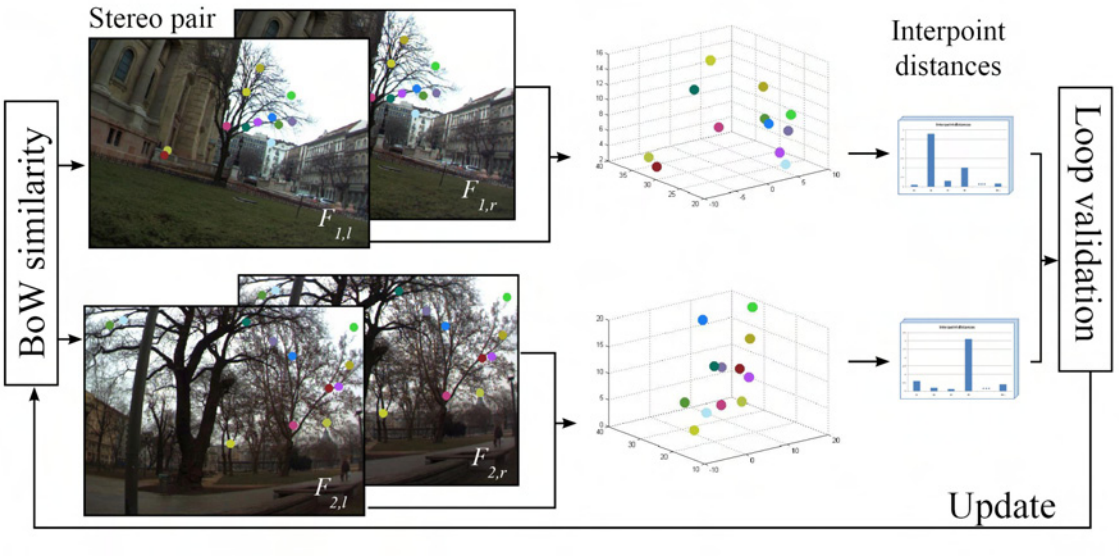

Fig. 4. Proposed loop-closure validation algorithm.

C. On-line Weight's Update

The ability of a certain word j to accurately detect a loop-closing situation highly depends on the information provided by the training set as considered by most of the reported works in the literature. Nevertheless, adaptation of the weights to the on-line behaviour of the algorithm is desirable providing greater flexibility and increasing the success rate as described latter in our experimental section.

Starting from its prior $\alpha_{j,0}$ we propose a simple dynamic model of the form,

$$\alpha_{j,k} = f(\alpha_{j,k-1},\ d_k) + \varepsilon_{j,k} \tag{5}$$

where additive zero mean Gaussian noise $\varepsilon_{j,k}$ is considered, and d_k refers to the correctness of the decision taken at time k concerning the detection of a loop-closing situation,

$$d_k = \begin{cases} 1, & \text{if loop-closure was correctly detected} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Similarly the variance is updated as,

$$\sigma^2_{j,k} = g(\sigma^2_{j,k-1},\ d_k) \tag{7}$$

As mentioned before, the random variable $\alpha_{j,k}$ can be interpreted as the ability of the j-th word to contribute to correctly detect a loop-closing situation. Therefore, the uncertainty in the knowledge of $\alpha_{j,k}$, which is related to its variance $\sigma^2_{j,k}$ should straightforwardly be updated depending on whether the decision was either correct or not. Intuitively, $\sigma^2_{j,k}$ should increase whenever the decision was wrong and it should decrease otherwise. Thus, the general function $g(\cdot)$ could be as simple as a multiplier with magnitude lower than one for $d_k = 1$ and greater than one otherwise.

Concerning the update of the mean value of $\alpha_{j,k}$ in the case a right decision (i.e. $d_k = 1$) was made, its ability to detect loop-closing situations should be increased provided that the mean value of $\alpha_{j,k-1}$ was above a certain threshold indicating that this j-th word has had a relevant role in the current decision. Conversely, if the mean value of $\alpha_{j,k-1}$ was below such a threshold, its ability should be reduced due to its minor contribution to the current decision.

Similarly, in the presence of a wrong decision (i.e. $d_k = 0$), the ability of the relevant words at time $k-1$ (i.e. those with mean value of $\alpha_{j,k-1}$ above the threshold) should be reduced because they have driven the system to a wrong decision, and conversely, those with a minor role in the current decision should be provided with an increase in its mean value.

The experimental section of the paper provides a more detailed description of the parametrization of the function $f(\cdot)$ and $g(\cdot)$ for the evaluation of the reported strategy.

IV. Loop-Closure Validation

This section presents a robust algorithm that validates whether or not the loop-closure was correctly detected by the visual appearance system and therefore it complements the strategy described above.

Usually a complementary sensor or a SLAM system [3], [13] is applied to capture metric information between the positions where the images were taken. We propose the use of a stereo camera system to validate and update the visual appearance system.

Roughly speaking, when a tentative loop-closure situation is detected by the BoW visual similarity system (Fig. 4) between two candidate images taken at different positions ($F_{1,l}$ and $F_{2,l}$), then the common visual words (marked with the same colors in the images and plots on Fig. 4) are searched in the right image of the stereo pairs taken at the different locations ($F_{1,r}$ and $F_{2,r}$ respectively).

Knowing the focal length of the camera and the baseline of the stereo system, local 3D-point maps are computed from the displacements of the visual words between the left and right image of the stereo pairs ($F_{1,l}$- $F_{1,r}$ and $F_{2,l}$-$F_{2,r}$)

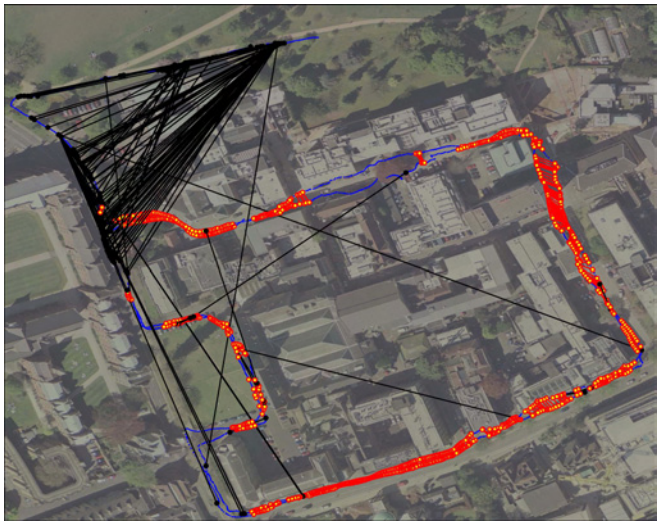

Fig. 5. Oxford City Center data set aerial view. A low similarity threshold was used to analyze the false positive loop-closure candidates.

Fig. 6. Obtained results on the Oxford City Center data set by applying the presented probabilistic on-line weight updating algorithm (aerial view).

representing the words in local real world coordinates in the scene. Thus it becomes possible to incorporate also rough depth information of these visual words in the scenes.

Next, the 3D inter-point distances are computed between the visual words, within the two local 3D-point clouds separately. In presence of a loop-closure situation the stereo image pairs would correspond to the same scene and, therefore, the computed inter-point distances would be similar, up to a certain error threshold. When at least 50% of the computed inter-point distances are below the predefined error threshold, the loop-closure is validated. A similar method, related to laser range data was reported in [14], however our proposal increases the portability of the sensor system and decreases the computational cost of the approach.

V. Single Camera Experiments

A first set of experiments was conducted by using the publicly available Oxford City Center dataset [2]. This dataset contains 2474 images gathered from a heterogeneous urban environment, obtained by a vision system mounted on a vehicle, with 1122 loop-closure events.

Fig. 5 depicts the trajectory of the vehicle (blue dots) overlaid to an aerial image[4] of the navigation area. Vegetation together with pedestrians and moving cars drove a classical appearance-based algorithm, with a low similarity threshold, to a large number of false positives (black links) concerning the detection of loop-closure situations. This effect is even more noticeable in the accompanying video. Opposite, on Fig. 6 the correctly recognized loop-closure situations (red links) are shown, obtained by applying the proposed adaptive appearance based weight updating algorithm without having any false positive detection. Please note that, in this case, the availability of ground-truth information guaranties the fully correct validation of loop-closure detection.

[4]The aerial images used are from `http://maps.google.com`

Using this dataset, the performance of different visual appearance algorithms have been evaluated in terms of:

1) Recall rate: Number of true positive detected over the total number of possible loop-closure events.
2) Precision rate: Number of true positive detected over the total number of loop-closure detected (both true and false).

For all the experiments presented in this section 64-D SURF features [15] were extracted from the images, with a Hessian threshold set to 500[5].

A. Results of the Off-line Weighting Strategy

Figure 7 compares the performance of two families of algorithms. First, those related to the classical tf-idf weighting strategy and then, those related to the concept of weight's priors described in section III-B. For each family, three different hierarchical vocabulary trees were built from the labeled data and the following parameterization: (1) 299 labeled images with $L = 4$ and $k = 11$ (red-colored); (2) 299 labeled images with $L = 5$ and $k = 11$ (black-colored); (3) 149 labeled images with $L = 5$ and $k = 11$ (blue-colored). The key performance metrics are summarized in Table I.

The experimental results demonstrate the remarkable improvement on the performance metrics due to the consideration of prior knowledge on the weighting strategy, irrespective of the characteristics of the underlying vocabulary tree used. A significant result is that obtained with a poorly trained (only 149 images used) vocabulary, where the proposed algorithm doubles the metrics of the classical tf-idf approach.

B. Results of the On-line Weighting Strategy

To illustrate the benefits of the on-line weighting strategy described in section III-C further experiments with the Oxford City Center dataset were conducted. As a baseline, both

[5]We used the OpenCV implementation of the SURF descriptor.

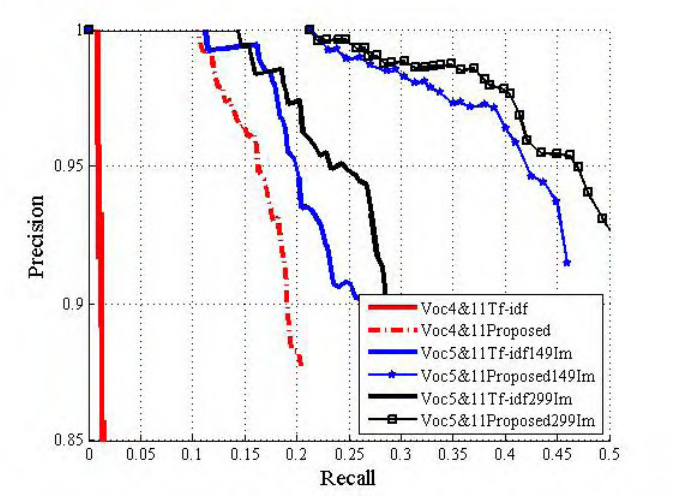

Fig. 7. Comparison of the results obtained with different vocabularies by applying the standard tf-idf versus the proposed algorithm based only on prior knowledge.

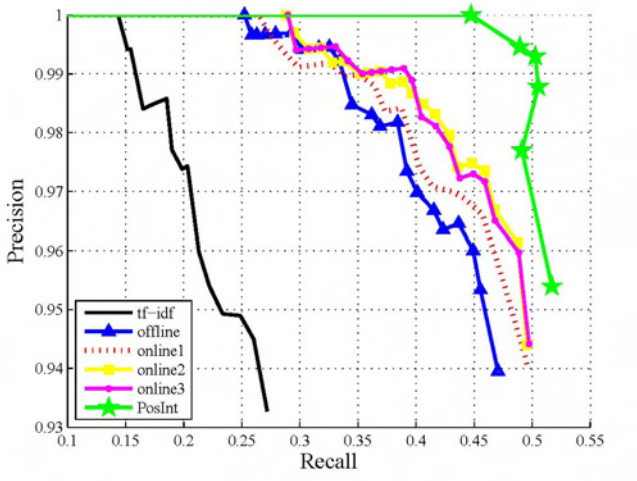

Fig. 8. Comparison of the results obtained by applying the standard tf-idf versus the proposed probabilistic on-line weight updating strategy.

the tf-idf and the best off-line weighting strategy reported in the previous section are considered, with a hierarchical vocabulary tree built from 299 labeled images with $L = 5$ and $k = 11$.

For the experimental setup a variance updating equation (eq. 7) of the form $\sigma^2_{j,k} = 0.01 d_k + \sigma^2_{j,k-1}$ with $\sigma^2_{j,0} = 0.05$ was considered. Additionally, the prior weights were initialized following eq. 4[6].

Figure 8 compares the on-line strategy with respect to the results commented above (tf-idf and offline) for different durations (in terms of the number of iterations of the algorithm throughout the complete dataset), namely, once (online1), twice (online2) and three times (online3), also in terms of the well-known performance metrics.

The analysis of the experimental results demonstrate the benefits on the on-line weighting strategy in terms of a significant increase in the performance metrics of the proposed algorithm over the baseline solutions considered. The adaptation of the hierarchical vocabulary tree parameters allow the algorithm to correct wrong decisions taken in the past which led to false-positive loop-closure detections. As observed from Fig. 8 subsequent trajectories on previously visited places gradually reduces the rate of wrong decisions.

Nevertheless, an important insight is that the adaptation of the weights reaches an upper limit from which further improvements in the performance metrics are negligible. In this particular experiment this upper limit is reached after three iterations on the complete dataset. For clarity those plots are not displayed in Fig. 8 but numerical results are provided in table II.

An additional improvement (PosInt plot in Fig. 8) on the key performance metrics is achieved by adopting a straightforward assumption on the vehicle position: if a loop-closure was detected from position p_{k-1} then, is highly probable that it would also be detected from position p_k. Moreover if p_{k-1} closed the loop with p_{j-1} then the current position p_k would probably close the loop with a neighbor position of the previous one such as $p_{(j-1)\pm\delta}$. Clearly, the value of δ would depend on the speed of the vehicle (in our experiments it was set to 2 frames). As depicted in the figure, a remarkable recall rate of around 45% was obtained, which constitute a success in view of the challenging, heterogeneous, outdoor environment considered for the experimentation.

[6]Please refer to the video attachment for simulated evolutions of the different parameters.

VI. Stereo Vision Experiments

A second set of experiments was performed from a dataset collected by the authors in the historical city center of Budapest (Hungary). A hand-held stereo camera gathered lateral images along a 2 km long human trajectory throughout the environment. A total of 562 stereo pair images were recorded, from which a maximum of 200 loop-closure events

TABLE I
Comparison of the Recall Rates without False Positives
(Off-line Weighting Strategy)

Vocabulary	tf-idf	proposed
Voc4L11k299Im	0.01	0.1
Voc5L11k149Im	0.1	0.21
Voc5L11k299Im	0.14	0.22

TABLE II
Comparison of the Recall Rates without False Positives
(On-line Weighting Strategy)

tf-idf	off-line	on-line first	on-line second
0.14	0.22	0.265	0.289
on-line third	**on-line fourth**	**on-line fifth**	**pos. int.**
0.29	0.291	0.29	0.447

Fig. 9. Obtained results on the Budapest City Center data set by applying the presented probabilistic on-line weight updating algorithm.

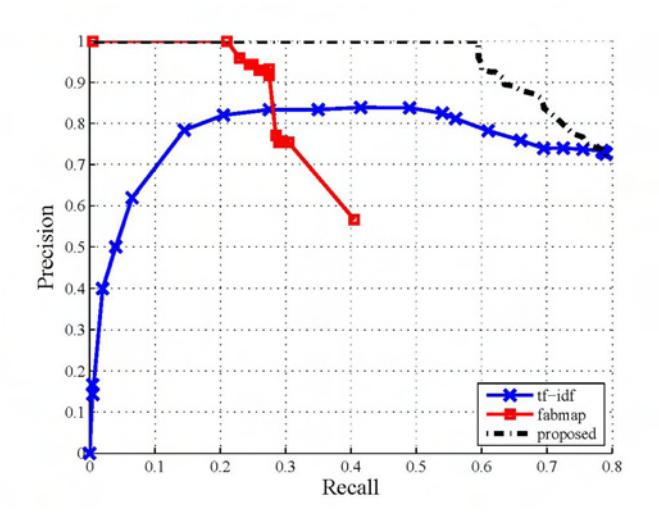

Fig. 10. Obtained results on the Budapest City Center data set by applying the presented probabilistic on-line weight updating algorithm.

TABLE III
COMPARISON OF THE RECALL RATES WITHOUT FALSE POSITIVES (BUDAPEST CITY CENTER DATASET)

tf-idf	FAB-MAP	Proposed
0	0.21	0.595

were identified by the human operator. The experiment was carried out in a heterogeneous environment characterized by the presence of buildings and urban furniture together with pedestrians, cars and different types of vegetation. The covered area was composed of a multitude of different urban scenes: park, narrow and wide streets with car traffic, pedestrian streets and public squares.

Fig. 9 represents the observation points along the trajectory (blue-dots) overlaid to an aerial image of the area. Also, correctly detected loop-closure locations are shown (yellow-points).

In this case 128-D SURF features [15] were used, with a Hessian threshold also set to 500. A hierarchical vocabulary tree with $L = 4$ and $k = 10$ was constructed from the labeled images, which can be considered small in comparison with previously reported work [3]. The dimension of the visual vocabulary is important because it has a direct influence on distinctiveness between words: a higher distinctiveness is obtained by using more words, resulting in a larger dictionary size.

The performance of the proposed on-line weighting algorithm was compared with the FAB-MAP algorithm [2][7] and the obtained results are shown in Fig. 10. As expected from the experiments described in the previous section, the proposed method outperforms current state-of-the-art algorithms. Table III also compares the performance of both methods in terms of recall rate in the cases of maximum precision.

The experimental results indicate that, roughly, 60% of the real loop-closure events were correctly detected, without false positive detections. We emphasize that these results were obtained using only visual information, without considering any smoothing parameter or assumption neither about the robot motion nor about consequent true positive loop-closure.

Finally, Fig. 11 shows some of the examples of correctly matched images, when the re-visited scene changed considerably. Even in these challenging situations for a visual recognition system correct loop-closure situations were detected by the proposed algorithm. Important illumination and field-of-view changes occurred between the observation of related images as can be seen on the right hand side images of Fig. 11.

VII. CONCLUSIONS AND FUTURE WORKS

The paper presented a vision-based algorithm that improved the performance metrics of state-of-the-art methods in the task of correctly detecting loop-closure events in outdoor heterogeneous environments.

The advantages of the proposed approach are two-fold: first, and during an off-line learning step, it computes a prior belief from labeled image data, attributed to the word weights of a hierarchical vocabulary tree; second, and during an on-line step, if updates the word weights according to their reliability and usefulness in the joint decision of loop-closure detection.

Thus, an on-line adaptive scheme is proposed which profits from the experimental evidence gathered during early stages of its performance to remarkably increase the rate of true-positive loop-closure events.

Experimental results with both a well-known publicly available dataset and an own dataset illustrate the benefits of

[7]The parameters of the FAB-MAP algorithm to give best possible performances were set as: P_OBSERVE_GIVEN_EXISTS = 0.7; LIKELIHOOD_SMOOTHING_FACTOR = 0.99; FORWARD_MOTION_BIAS = 0.9

Fig. 11. Examples of correctly matched images captured at the same place at different visits, despite considerable scene change (left, right) or illumination and field of view change (right).

the proposed approach for key robotics navigation tasks. As observed from the results up to 60% recall rates are achieved in a challenging urban environment.

Future work aims at extended the experimental evidence on the performance of the described algorithm in even more challenging settings. The introduced framework enables to train, with minor changes the belief of the visual words as positive or negative ones by using automatic word labeling tools and labeled databases (such as LabelMe [16]).This task is intended to be integrated as a future development. Also, the concept of co-appearance probability of the visual words will be studied.

Another future interest is the use of supervised learning methods for object classification and localization in images collected by the mobile platform, in order to further improve appearance based loop-closing in heterogeneous environments.

REFERENCES

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

[2] M. Cummins and P. Newman, "FAB-MAP: probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[3] P. Piniés, L. M. Paz, D. Gálvez-López, and J. Tardós, "CI-Graph SLAM for 3D reconstruction of large and complex environments using a multicamera system," *Journal of Field Robotics*, vol. 27, no. 5, pp. 561– 586, Sept/Oct 2010.

[4] E. Olson, "Recognizing places using spectrally clustered local matches," *Robotics and Autonomous Systems*, vol. 57, no. 12, pp. 1157–1172, 2009.

[5] C. Cadena, D. Gálvez-López, F. Ramos, J. Tardós, and J. Neira, "Robust Place Recognition with Stereo Cameras," in *IEEE International Conference on Intelligent Robots and Systems*, 2010, pp. 5182–5189.

[6] D. G. Sabatta, D. Scaramuzza, and R. Siegwart, "Improved appearance-based matching in similar and dynamic environments using a vocabulary tree," in *IEEE International Conference on Robotics and Automation*, 2010, pp. 1008–1013.

[7] P. Turcot and D. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops),*, 2009, pp. 2109 –2116.

[8] A. Kawewong, N. Tongprasit, S. Tangruamsub, and O. Hasegawa, "Online and incremental appearance-based slam in highly dynamic environments," *International Journal of Robotics Research*, vol. 30, pp. 33–55, January 2011.

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

[10] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[11] J. Yang, Y. Jiang, A. Hauptmann, and C. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*. ACM, 2007, p. 206.

[12] T. S. Hao Shao and L. V. Gool, "Zubud - zurich buildings database for image based recognition," Computer Vision Laboratory, ETH Zurich, Tech. Rep. 260, 2003.

[13] A. J. Glover, W. P. Maddern, M. Milford, and G. F. Wyeth, "FAB-MAP + RatSLAM: appearance-based slam for multiple times of day," in *IEEE International Conference on Robotics and Automation*, 2010, pp. 3507–3512.

[14] R. Paul and P. Newman, "FAB-MAP 3D: Topological mapping with spatial and visual appearance," in *Proc. IEEE International Conference on Robotics and Automation (ICRA'10)*, Anchorage, Alaska, May 2010, pp. 2649–2656.

[15] T. T. Herbert Bay and L. V. Gool, "SURF: Speeded up robust features," in *Proceedings of the 9th European Conference on Computer Vision*, vol. 3951, no. 1. Springer LNCS, 2006, pp. 404–417.

[16] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, pp. 157–173, May 2008.