COMPARISON OF LOCAL DESCRIPTORS FOR HUMANOID ROBOTS LOCALIZATION USING A VISUAL BAG OF WORDS APPROACH

NOÉ G. ALDANA-MURILLO, JEAN-BERNARD HAYET AND HÉCTOR M. BECERRA

Centro de Investigación en Matemáticas (CIMAT)
Jalisco S/N, Col. Valenciana, C.P. 36240, Guanajuato, Gto., Mexico.
E-mails: {noe.aldana, jbhayet, hector.becerra}@cimat.mx

ABSTRACT— In this paper, we address the problem of the appearance-based localization of a humanoid robot, in the context of robot navigation. We only use information obtained by a single sensor, in this case the camera mounted on the robot. We aim at determining the most similar image within a previously acquired set of key images (also referred to as a visual memory) to the current view of the monocular camera carried by the robot. The robot is initially kidnapped and the current image has to be compared with the visual memory. To solve this problem, we rely on a hierarchical visual bag-of-words approach. The contribution of this paper is two-fold: (1) we compare binary, floating-point and color descriptors, which feed the representation in bag-of-words using images captured by a humanoid robot; (2) a specific visual vocabulary is proposed to deal with the typical issues generated by the humanoid locomotion.

Key Words: Vision-based localization, humanoid robots, local descriptors comparison, visual bag of words.

1. INTRODUCTION

Visual memory mimics the human behavior of remembering key visual information when moving in unknown environments, to make the future navigation easier in a topological map. This metric-free methodology has been extensively studied in the context of wheeled mobile robots [1, 2, 3]. However, much fewer studies are reported in the context of humanoid robots. This kind of robots raises particular challenges that are not considered in the aforementioned works. In particular, because of the bipedal locomotion, sharp accelerations produce blurring effects on the images; also, the robot sway motion produces rotations around the optical axis.

The navigation of a robot based on a visual memory typically implies two distinct stages [1, 2]. First, the learning stage consists in making the robot build a representation of the initially unknown environment, by means of a set of key images that forms the so-called visual memory. Then, in the autonomous navigation stage, the robot has to reach a location associated to a desired key image by following a visual path. That path is defined by a subset of images from the visual memory that topologically connects the key image that is the most similar to the current robot view to the target image. The autonomous navigation stage may be addressed assuming that the visual path is given, like in [3, 4], where the control laws are formulated in terms of a geometric constraint, with nonlinear and fuzzy approaches, respectively.

A visual memory is built by selecting a subset of images from a sequence of images captured in a learning stage. The basic criterion for the selection of a key image in the learning stage is to satisfy a compromise between two aspects: a pair of consecutive key images must share enough visual information (for instance a minimum number of matched interest points) and, at the same time, we do not want too many images. The construction of the visual memory is a problem by itself. In this work, we assume that the visual memory is given and we focus on the problem of localizing a robot within this visual memory. Moreover, we assume that the robot is initially placed at an unknown location with no prior knowledge about this location or about where it was previously. This is known as the kidnapped robot problem [5].

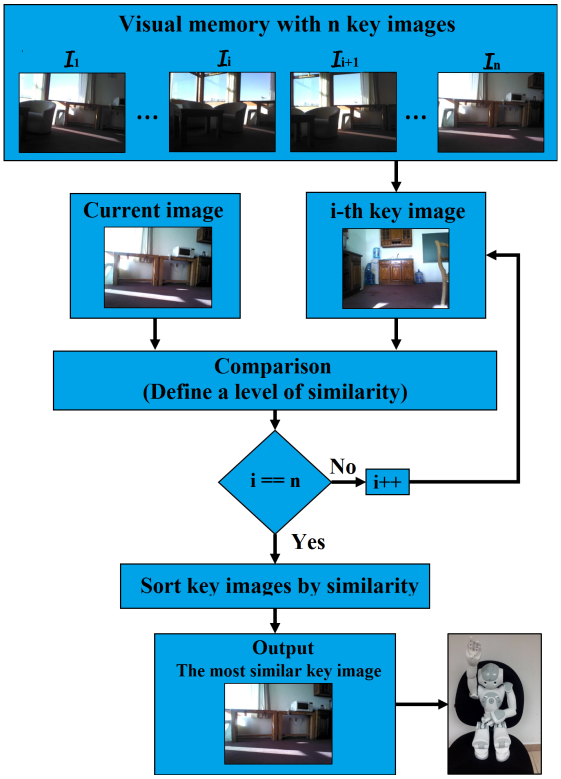

Figure 1: General diagram of the appearance-based localization from a visual memory.

Few works have been reported about the navigation of humanoid robots based on a visual memory [6, 7]. In the aforementioned works, the robot is not initially kidnapped, as in our setup, but instead starts its navigation from a known position. Hence, compared to these works, one of the main motivation in this paper is to propose a solution to the appearance-based localization problem, where the current image is matched to a known location only by comparing images [8]. In particular, we address the problem of the localization of humanoid robots using monocular vision only, by determining the key image in a visual memory that is the most similar in appearance to the current view of the robot (input image). Fig. 1 presents an overview of the problem. Consider that the visual memory consists of n key images $(\mathcal{I}_1^*, \mathcal{I}_2^*, ..., \mathcal{I}_n^*)$. The current view $\mathcal{I}$ has to be compared (a priori) with the n key images and the method should give us as an output the most similar key image $\mathcal{I}_o^*$ within the visual memory, from which the visual path to the target image could be defined.

Since a naive comparison of $\mathcal{I}$ with all the visual memory would take too much time, depending on the size of the visual memory, we use a method that compresses the visual memory into a compact, efficient-to-access representation: the visual bag of words (VBoW) [9]. A bag of words is a structure that represents an image as a numerical vector, allowing fast images comparisons. In robotics, the VBoW approach has been used in particular for loop-closure detection in Simultaneous Localization and Mapping (SLAM) [10, 11], where re-visited places have to be recognized to manage the proper closure of loops while building maps.

We want to emphasize that our appearance-based localization is part of a more global visual memory approach, which relies on navigating with visual information extracted directly from images, without the need of a costly estimation machinery such as SLAM. The SLAM methods may provide a more accurate localization in terms of metric information, but it has been shown that appearance information is sufficient to navigate [7]. In this paper, a qualitative evaluation of the VBoW approach for the visual navigation problem is carried out. We apply the VBoW approach on real datasets captured by a camera mounted in the head of a small-size humanoid robot. An evaluation and comparison of the performance of different local descriptors are presented. The images taken by the humanoid robot are affected by the sway motion due to its

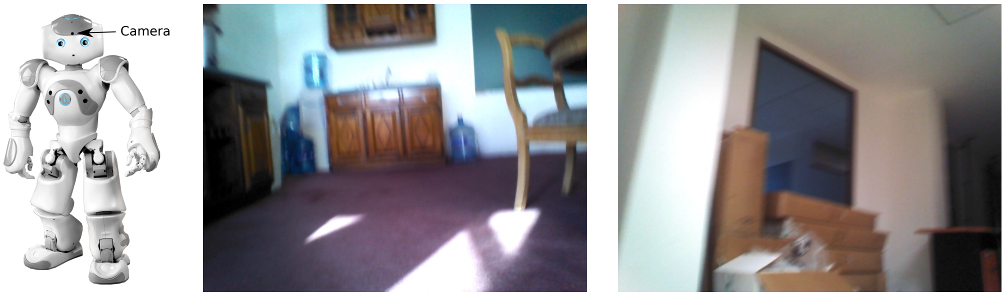

Figure 2: Humanoid robotic platform NAO and examples of images from its onboard camera.

locomotion. They undergo blurring and rotation around the optical axis. Hence, a specific visual vocabulary is proposed to tackle those issues. Fig. 2 shows two examples of 640×480 pixels images captured from our experimental platform: a NAO humanoid robot, where blur and rotation effects are visible.

This paper is an extension to a previously published conference paper, presented in the Mexican Conference on Pattern Recognition [12]. It is organized as follows. Section 2 introduces the local descriptors included in our evaluation. Section 3 details the VBoW approach as we implemented it. Section 4 presents the results of the experimental evaluation and Section 5 gives a few conclusions about the approach.

2. LOCAL DESCRIPTORS

Local features describe regions of interest of an image through descriptor vectors, defined around special image locations, given by local feature detectors. In the context of image comparison, the idea is that groups of local features should be robust against occlusions and viewpoint changes, in contrast to global methods, both in terms of the features locations and in terms of the local descriptors. Hence, from the existing local detectors/descriptors, we wish to select the best option for the specific task of appearance-based humanoids localization. Hereafter, we introduce the local descriptors selected for a comparative evaluation.

2.1 Real-valued descriptors

A popular system of combined keypoint detector/descriptor is SURF (Speeded Up Robust Features) [13]. It has good properties of invariance to scale and rotation. SURF keypoints can be computed and compared much faster than their previous competitors. Thus, we only selected SURF as a real-valued descriptor to be compared in our localization framework. The detection uses local extrema of the Hessian matrix and makes an approximation of the Hessian with integral images, to reduce the computation time. The descriptor combines approximate Haar-wavelet responses within the interest point neighborhood and also exploits integral images to increase speed. In our evaluation, the standard implementation of SURF (descriptor vector of dimension 64) included in the OpenCV library is used, with 4 octaves and 2 layers in each octave.

2.2 Binary descriptors

Binary descriptors represent image features by binary strings instead of floating-point vectors. Thus, the extracted information is very compact, occupies less memory and can be compared faster. Two popular binary descriptors have been selected for our evaluation: Binary Robust Independent Elementary Features (BRIEF [14]) and Oriented FAST and Rotated BRIEF (ORB [15]). Both use variants of FAST (Features from Accelerated Segment Tests) [16] as a detector, i.e. they detect keypoints by comparing the gray levels along a circle of radius 3 to the gray level of the circle center. In average, most pixels can be discarded soon, hence the detection is fast. BRIEF uses the standard FAST keypoints while ORB uses oFAST keypoints, an improved version of FAST including an adequate orientation component. The BRIEF descriptor is a binary vector of user-choice length where each bit results from an intensity comparison between some pairs of pixels within a patch around the detected keypoints. The patches are previously smoothed with a Gaussian

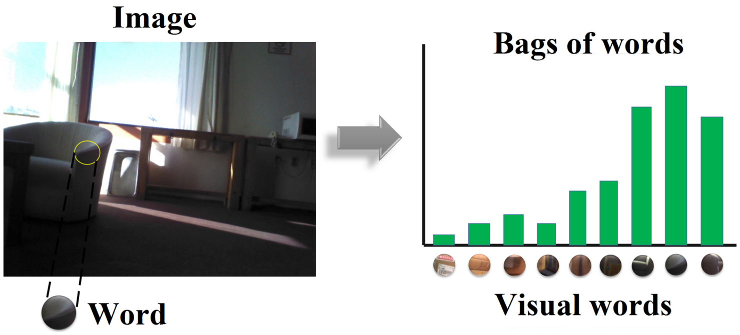

Figure 3: Representation of an image in visual words.

kernel to reduce noise. They do not include information of rotation or scale, so they are hardly invariant to them. This issue can be overcome by using the rotation-aware BRIEF descriptor (ORB), that computes a dominant orientation between the center of the keypoint and the intensity centroid of its patch. The BRIEF comparison pattern is rotated to obtain a descriptor that should not vary when the image is rotated in the plane. In our evaluation, we use oFAST keypoints given by the ORB detection method as implemented in OpenCV along with BRIEF with a patch size of 48 and a descriptor length of 256. The ORB implementation is the one of OpenCV with 256 bits descriptors.

2.3 Color descriptors

We also evaluate the bag-of-words image comparison approach with color information only. To do so, we use rectangular patches and a color histogram is associated to each patch as a descriptor. We select the HSL (Hue-Saturation-Lightness) color space because its three components are more natural to interpret and less correlated than in other color spaces. Also, only the H and S channels are used, in order to achieve as much robustness as possible against illumination changes. The color descriptor of each rectangular patch is formed by a two-dimensional histogram of hue and saturation and the length of the descriptor was set experimentally to 64 bits. Three different alternatives are evaluated to select the patches:

- Random patches: A number of 48×64 patches, randomly selected. Hereafter, this option is referred to as Color-Random.
- Uniform grid: A uniform grid of patches covering the image, with patches overlapped by half of their size. This option is referred to as Color-Whole.
- Uniform grid on half of the image: Instead of using the whole image, only the upper half is used. This is because the inferior parts, when taken by the humanoid robot, are mainly projections of the floor and do not discriminate well among possible locations. This option is referred to as Color-Half.

3. VISUAL BAG OF WORDS FOR HUMANOIDS LOCALIZATION

As mentioned above, this work relies on the hierarchical visual bag of words approach [17] to combine the highly descriptive power of local descriptors (see above) with the versatility and robustness of histograms. In Section 3.1, we recall the main characteristics of the work presented in [17], and then in Section 3.2, we introduce a novel use of the BRIEF descriptor suited within a VBoW approach in the context of humanoid robots navigation.

3.1 Hierarchical Visual Bag of Words Approach

The visual bag of words approach starts by discretizing the local descriptors space in a series of words, which can be considered as clusters in the local descriptors space. Obviously, the nature of this space (histograms, responses to filters) depends on the selected descriptors. Here, we follow the strategy of Nister and Stewenius [17], who performs this discretization step in a hierarchical way: in the set of n acquired key images $\mathcal{I}_1^*, \mathcal{I}_2^*, ..., \mathcal{I}_n^*$ forming the visual memory, a pool of D local descriptors is detected, as illustrated in Fig. 3, left. The local descriptors can be extracted by any of the methods mentioned above. Given a branch factor k, the idea is to form k clusters among the D descriptors by using the *kmeans++* algorithm. Then, the sets of descriptors associated to these k clusters are recursively clustered into other k clusters, and so on, up to a maximum depth of L levels, as depicted in Fig. 4. The leaves of this tree of recursively refined clusters correspond to the visual words, i.e., the clusters in the local descriptors space.

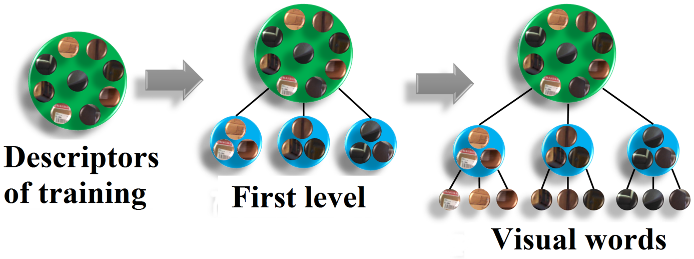

Figure 4: Hierarchical approach to build a visual bag of words.

When handling a new image $\mathcal{I}$, d descriptors are extracted, and each of these is associated to a visual word as explained above. This way, we obtain an empirical distribution of the visual words in $\mathcal{I}$, in the form of a histogram of visual words $v(\mathcal{I})$ (see Fig. 3, right). Now, the content of $\mathcal{I}$ can be compared with any of the key images $\mathcal{I}_i^*$ by comparing their words histograms. Of course, because n may be very high, it is out of question to compare the histogram of $\mathcal{I}$ with the n histograms of the key images. That is why an important element in this representation is the notion of inverse dictionary: for each visual word, one stores the list of images from the visual memory containing this word. Then, on a new image, we can easily determine, for each visual word it contains, the list of key images also containing this word. This way, to limit the number of comparisons, we restrain the search for the most similar images to the subset of key images having at least 5 visual words in common with the test image.

For an image $\mathcal{I}$, each histogram entry v_i (where i refers to the visual word) is defined as:

$$v_i(\mathcal{I}) = \frac{c_i(\mathcal{I})}{c(\mathcal{I})} \log(\frac{n}{n_i}), \tag{1}$$

where $c(\mathcal{I})$ is the total number of descriptors present in $\mathcal{I}$, $c_i(\mathcal{I})$ the numbers of descriptors in $\mathcal{I}$ classified as word i, and n_i the number of key images where the word i has been found. The $\log$ term allows to weight the frequency of word i in $\mathcal{I}$ in function of its overall presence: if a word is present everywhere in the database ($n_i \approx n$), then the information of its presence will not help in discriminating among images.

Last, we should choose how to compare histograms. After intensive comparisons made among the most popular metrics for histograms such as dot product, χ^2, Bhattacharyya coefficient, $L_1 - norm$ and $L_2 - norm$, we have observed that the best results were obtained with the χ^2 distance.

Hence, for any new image $\mathcal{I}$, we obtain a score against any image from the visual memory sharing at least 5 words with it. Fig. 5 sums up the whole methodology.

3.2 *A BRIEF-based vocabulary for humanoids localization*

We introduce a novel use of the BRIEF descriptor, suited for a VBoW approach in the context of humanoid robots. This is a specific vocabulary that we called BRIEFROT, which deals with the issues generated by the humanoid locomotion. As mentioned above, the humanoid locomotion generates rotated images within a certain rotation angle due to the lateral sway motion, as can be seen in Fig. 6. Due to this, the detected interest points may appear rotated relatively to how they have been learned in the visual memory.

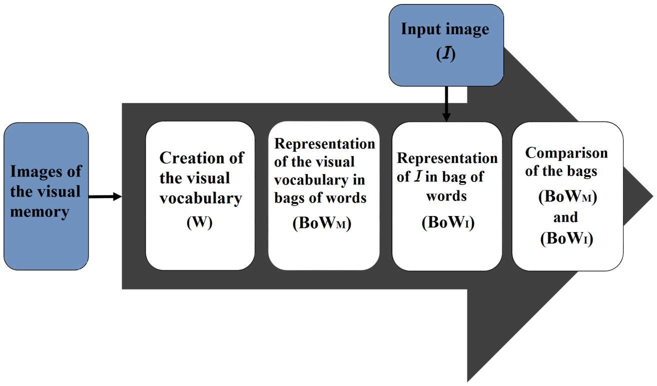
Input image
($\mathcal{I}$)
Images of the visual memory
Creation of the visual vocabulary
(W)
Representation of the visual vocabulary in bags of words
(BoW_M)
Representation of $\mathcal{I}$ in bag of words
(BoW_I)
Comparison of the bags
(BoW_M)
and
(BoW_I)

Figure 5: Complete method for image comparison based on visual bag of words.

However, much of detected points rotate with a similar angle with respect to their learned appearance. Hence, we apply rotations that span the possible values of this transformation to all the detected BRIEF points and generate a new vocabulary: BRIEFROT.

This comes as an alternative to ORB descriptors, that are invariant to rotation, as BRISK or FREAK. However, for these rotational invariant descriptors, each point has an intrinsic orientation that may correspond to a different rotation angle for each descriptor when it is not correctly estimated. This may lead to the non-detection of some of the visual words present in the image. Our proposed BRIEFROT vocabulary rotates all the detected descriptors to a certain fixed value by trying to imitate the real rotation of the points in the images because of the robot's sway motion and, therefore, more visual words can be retrieved. Also, we make the computation times lower, by avoiding the computation of the intrinsic orientation.

The humanoid locomotion also generates blurry images. To try to mitigate this problem, patches around the detected points are smoothed with a Gaussian kernel to reduce the difference between descriptors of sharp images and descriptors of blurry images.

The BRIEFROT vocabulary contains three independent internal BRIEF vocabularies, two of which are rotated at a fixed angle: one anti-clockwise, the other clockwise. To create these vocabularies, BRIEF points are extracted and they are rotated by an adequate angle to generate the rotated vocabularies. The third vocabulary is identical to the normal BRIEF. The angle of rotation was obtained experimentally by varying the angle to obtain the largest number of correct results, as will be described in Section 4.3. Formally speaking, we have three vocabularies, generated from the three rotations. Hence, for any new image $\mathcal{I}$, we can define $z_i^{(o)}(\mathcal{I})$, the score between image $\mathcal{I}$ and the i-th image of the vocabulary $o \in \{1,2,3\}$. Note that by construction, the key images are the same in all cases; what is different is the set of features contained in the vocabulary. Then, the overall score is simply defined as:

$$z_i(\mathcal{I}) = \max_{o \in \{1,2,3\}} z_i^{(o)}(\mathcal{I}). \tag{2}$$

These rotated vocabularies were implemented with the idea of coping with the slight rotation caused by the locomotion of these robotic systems. The rotated vocabularies represent the images of the visual memory as rotated features. When evaluating a new image, we evaluate it under the three vocabularies separately and the winner is the vocabulary that obtains the maximum result.

The idea of using three vocabularies is that if the input image is rotated with respect to any image of the visual memory, then the image should be detected by one of the rotated vocabularies; if the input image is not rotated with respect to any image of the visual memory, then it is detected with the vocabulary without rotation.

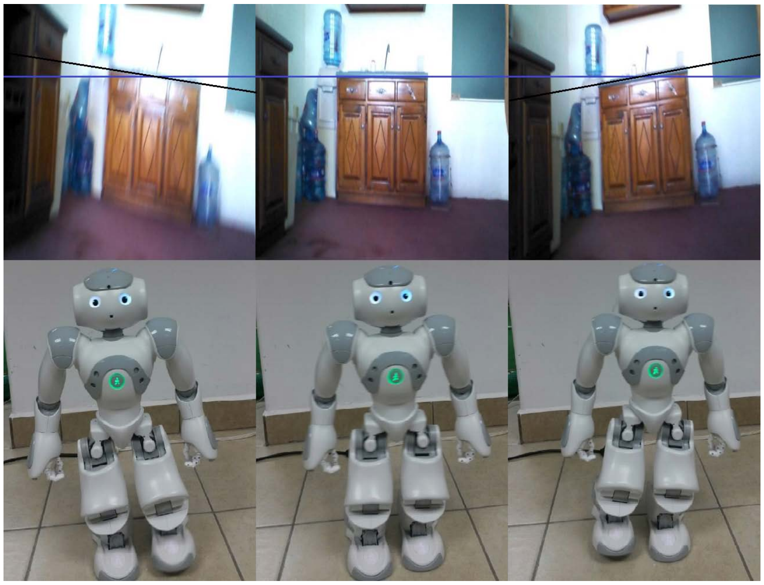

Figure 6: Motivation for a rotated BRIEF vocabulary. As the humanoid robot walks the images capture by the onboard camera are rotated with respect to the optical axis.

4. EXPERIMENTAL EVALUATION

We evaluated the local descriptors mentioned in Section 2 on 5 distinct datasets. Four of these datasets correspond to indoor environments (*CIMAT-NAO-A*[1], *CIMAT-NAO-B*[1], *Toulouse* and *Bicocca* [18]) and one corresponds to an outdoor environment (*New College* [19]). Table 1 summarizes the datasets characteristics. The tests were done with a laptop using Ubuntu 12.04 with 4 Gb of RAM and a 1.30 GHz processor.

Table 1: Datasets used for the evaluation of the local descriptors.

Dataset	Description	Image size (px × px)
CIMAT-NAO-A	Indoors/humanoid	640 × 480
CIMAT-NAO-B	Indoors/humanoid	640 × 480
Bicocca 2009-02-25b [18]	Indoors/Wheeled robot	320 × 240
New College [19]	Outdoors/Wheeled robot	384 × 512
Toulouse [20]	Indoors/humanoid	572 × 390

4.1 *Description of the evaluation datasets*

The *CIMAT-NAO-A* dataset was acquired with a NAO humanoid robot inside CIMAT. This dataset contains 640×480 images of good quality but also many blurry ones. Some images are affected by rotations introduced by the humanoid locomotion or by lighting changes. We have used 187 images, hand-selected, as a visual memory and 258 images for testing. Fig. 7 shows a map of the locations associated to the images of the visual memory for the *CIMAT-NAO-A* dataset. This metric information was obtained from the robot's odometry for visualization purposes and it is not used by our localization method. The *CIMAT-NAO-B* dataset was also captured indoors at CIMAT, with the same humanoid robot. It also contains good quality and blurry 640×480 images, but it does not have images with drastic light changes, as in the previous dataset. We have used 94 images as a visual memory and 94 images for testing.

[1] Datasets available at `http://personal.cimat.mx:8181/~hmbecerra/CimatDatasets.zip`

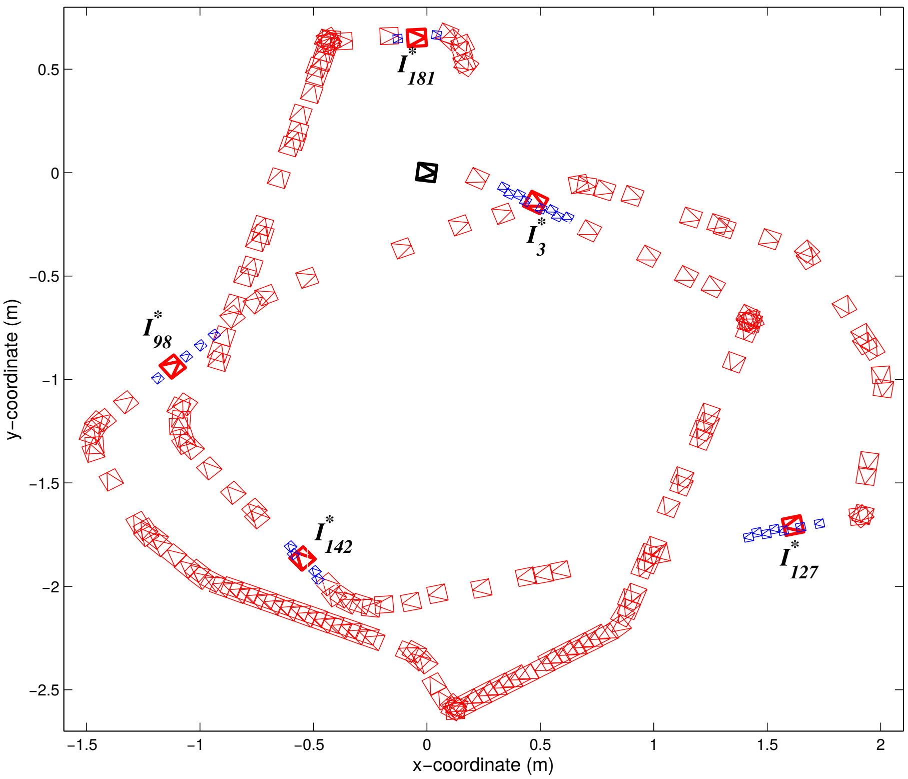

Figure 7: Map of the visual memory built from the *CIMAT-NAO-A* dataset. In bold red, five representative key images were selected. In blue, for each of the aforementioned selected key images, we show a few images that were selected for testing the localization, the closest key image being clear in those cases.

The *Toulouse* dataset was acquired by the humanoid robot HRP-2 inside a laboratory and it is available online [20]. It contains 572 × 390 grayscale images of good quality and very few blurry images. Moreover, compared with *CIMAT-NAO-A*, the images have almost zero rotation around the optical axis. The *Bicocca 2009-02-25b* dataset is also available online [18] and was acquired by a wheeled robot inside a university. The 320×240 images have no rotation around the optical axis nor blur. We used 120 images as a visual memory and 120 images for testing. Unlike the three previous datasets that were obtained indoors, the *New College* dataset was acquired outside the Oxford University by a wheeled robot [19], with important light changes. The 384×512 images are of good quality with no rotation nor blur. For this dataset, 122 images were chosen as a visual memory and 117 images for testing. The last two datasets were used to compare the visual vocabularies obtained by wheeled robots with the images obtained by humanoids robots, in order to make clearer the acquisition issues arising in the second case.

4.2 Evaluation metrics

Since the goal of this work is to evaluate different descriptors in a VBoW approach, it is critical to define adequate evaluation metrics to assess the quality of the result from our application. We propose two metrics that compare the ground-truth data to the algorithm output and generate a score; the first one is the ranking (according to the bag-of-words algorithm) of the theoretically most similar image:

$$\mu_1(\mathcal{I}) = \text{rank}(\bar{k}(\mathcal{I})), \tag{3}$$

where $\bar{k}(\mathcal{I})$ is defined as the ground truth index of the key image associated to $\mathcal{I}$. In the best case, the rank of the closest key image in the list of closest images given by our algorithm should be one, so $\mu_1(\mathcal{I}) = 1$

means that the retrieval is perfect, whereas higher values correspond to worse evaluations.

The second metric is similar in essence:

$$\mu_2(\mathcal{I}) = \sum_{l} \frac{z_l(\mathcal{I}^*_{\bar{k}(\mathcal{I})})}{\sum\limits_{l'} z_{l'}(\mathcal{I}^*_{\bar{k}(\mathcal{I})})} \operatorname{rank}(l), \tag{4}$$

where we recall that $z_l(\mathcal{I})$ is the similarity score between the key image l from the visual memory and the image $\mathcal{I}$. Hence, $z_l(\mathcal{I}^*_{\bar{k}(\mathcal{I})})$ refers to the similarity score between the key image l and the key image $\bar{k}(\mathcal{I})$ (the theoretically closest image to ours, according to the ground truth). This metric is "fair" in the sense that it handles the presence of close, similar key images within the dataset; hence, with this metric, the final score integrates weights (normalized by $\sum\limits_{l'} z_{l'}(\mathcal{I}^*_{\bar{k}(\mathcal{I})})$ to sum to one) from the key images l similar to the closest ground truth image $\bar{k}(\mathcal{I})$; this ensures that all the closest images are also well ranked.

4.3 Parameters selection

There are three free parameters in the VBoW algorithm: the number of clusters k at each level of the tree, the tree depth L, and the measure of similarity. Tests were performed by varying the value of k from $k = 8$ to $k = 10$ and varying the value of L from $L = 4$ to $L = 8$. Also, different similarity measures between histograms were tested: L1-Norm, L2-Norm, χ^2, Bhattacharyya and dot product (note that some of them are distances, other are similarity coefficients). To do the tests, we generated ground truth data, by defining manually the most similar key image to the input image, with which we can define the evaluation metrics μ_1 and μ_2 explained before. By examining the statistics of μ_1 scores, we obtained the best results with $k = 8$, $L = 8$ and the metric χ^2. The dataset used for the parameters selection was the *CIMAT-NAO-A*, since it is the most challenging dataset for the type of images it contains.

For interest-point feature detection, an important parameter to consider in the whole approach is the number of features to detect and to use to build the word histogram based on the vocabulary. In Fig. 8, we show effectiveness results (proportion of test images for which $\mu_1 = 1$) obtained for different features and different numbers of points, for the *CIMAT-NAO-A* dataset. Based on these results, we chose the number of 500 extracted features and discretized into words, in all the following experiments.

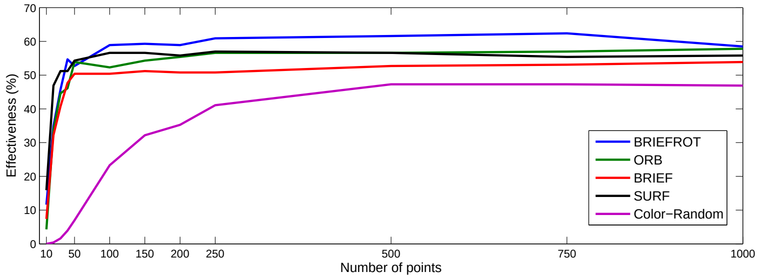

Figure 8: Performance evaluation for different number of point features.

A specific parameter to be adjusted for the BRIEFROT vocabulary is the angle of rotation for the rotated vocabularies. The tests were done by varying the angle from 0 to 20 degrees. In Fig. 9, the results are presented for different numbers of interest points, again in the case of the *CIMAT-NAO-A* dataset. The best results occur when the angle of rotation is 8 degrees. Therefore, the rotation value was set at this value and with 500 interest points. Note that for 750 points, better results were obtained, but we chose the 500 points because it requires less time of computation to realize the tests. Also, the effectiveness with respect to the results obtained for 750 points is very similar. This confirms the points of Fig. 8, discussed above.

Finally, to select the size of the color patches when using regular grids, we proceeded to similar effectiveness evaluations on the same *CIMAT-NAO-A* dataset, shown in Fig. 10. The x axis in this figure is the

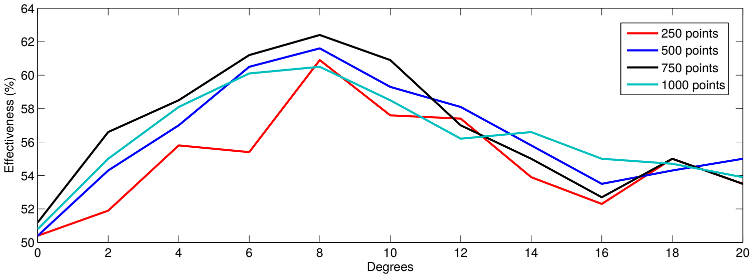

Figure 9: Effectiveness of BRIEFROT for different degrees of rotation.

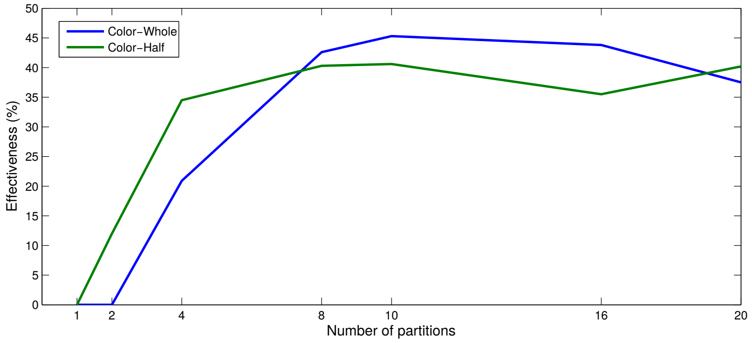

Figure 10: Performance evaluation of the Color-Half and Color-Whole for different number of partitions of the image.

number of partitions done in both x and y directions of the images to generate the grid. We selected this number as 10, which led us to extact 19×19 color patches.

4.4 Analysis of the results obtained on the evaluation datasets

In this section, we present evaluations obtained for the seven vocabularies created.

Computational times analysis. First, we compare the computational times (in mili-seconds) of the different features to be used in the vocabularies. In Tables 2 to 7, last column, we indicate the times needed to compute 500 interest points (or randomly selected color patches) in an image and to compare them to the visual memory. For the color patches grid, 361 patches are computed. These are times averaged over all the test images, for experiments run on a laptop using Ubuntu 12.04 with 4 Gb of RAM and a 1.30 GHz processor. All the approaches allow a close-to-real-time localization of the test image. Clearly, because of the multiple scales, SURF takes longer times than the other approaches, ORB being the fastest one for most datasets after the color approaches with regular grid.

4.4.1 Effectiveness analysis.

On the one hand, the effectiveness of the representation of the images through the chosen vocabularies is evaluated by using the score μ_1. In this case, the threshold chosen for the test to be classified as correct was 1 (i.e., a perfect match). On the other hand, for the level of confidence μ_2 we choose a threshold of 2.5, so that all tests showing scores below 2.5 were considered as correct. This second level of confidence is looser but takes into account the similarity between images within the visual memory.

CIMAT-NAO-A dataset

In Table 2, we present results obtained for the *CIMAT-NAO-A* dataset. In this case, the BRIEFROT vocabulary obtained the best behavior for both levels of confidence. For the case of μ_1, it has an efficiency of

Table 2: Percentages of correct results for the *CIMAT-NAO-A* dataset.

Descriptor	Number	Correct	Effectiveness	Correct	Effectiveness	Average
	of tests	tests μ_1	μ_1 (%)	tests μ_2	μ_2 (%)	time (ms)
BRIEF	258	132	51.16	185	71.71	122.6
BRIEFROT	258	157	60.85	194	75.19	132.4
Color-Random	258	110	42.64	117	45.35	129.4
Color-Half	258	104	40.31	160	62.01	93.1
Color-Whole	258	110	42.64	162	62.79	101.8
ORB	258	144	55.81	194	75.19	107.5
SURF	258	135	52.32	187	72.48	296.5

Table 3: Percentages of correct results for the *CIMAT-NAO-B* dataset.

Descriptor	Number	Correct	Effectiveness	Correct	Effectiveness	Average
	of tests	tests μ_1	μ_1 (%)	tests μ_2	μ_2 (%)	time (ms)
BRIEF	94	63	67.02	82	87.23	87.23
BRIEFROT	94	65	69.14	81	86.17	112.0
Color-Random	94	62	65.96	86	91.49	109.9
Color-Half	94	64	68.09	78	82.98	63.2
Color-Whole	94	68	72.34	83	88.3	73.1
ORB	94	69	73.40	83	88.3	77.7
SURF	94	70	74.46	86	91.49	267.9

60.85 % (measured as the proportion of correct test cases, i.e. such that $\mu_1 = 1$) and for μ_2 of 75.19 % (proportion of correct test cases, i.e. such that $\mu_2 < 2.5$). Also, the ORB vocabulary offered good performance for μ_2 and was the second best for μ_1. The Color-Half vocabulary obtained the worst results.

CIMAT-NAO-B dataset

In Table 3, we present results for the *CIMAT-NAO-B* dataset. It can be noticed that the SURF vocabulary behaved better than BRIEFROT, but with higher computation times. The times reported were measured from the stage of features extraction to the stage of comparison. ORB, again, behaved well and was the second best vocabulary. The Color-Random vocabulary obtained the worst performance for μ_1, but for μ_2 it was one of the best vocabulary; this means that it tends to put the correct key image in the second rank. Color-Half had the worst results for μ_2.

Toulouse dataset

In the *Toulouse* dataset, we have images with radial distortion due to the robot cameras and we undistort them for testing. We do not use color vocabularies because the images are grayscale. In Table 4, the results for this dataset, after image un-distortion, are presented. The BRIEFROT vocabulary achieved the best results for both evaluation metrics. The BRIEF vocabulary also obtained good performance, very similar to BRIEFROT. This is because the images of this dataset have almost zero rotation with respect to the camera optical axis. This can be seen in Fig. 11, in which we evaluate different BRIEFROT rotation angles. It can be seen that there is no clear angle value for which optimal results are obtained. This is due to the fact that the zero-degree vocabulary of BRIEFROT can explain all the images by itself. In Fig. 12, this is illustrated by a plot showing the frequencies of vocabularies (among the three) having the largest response, on average for this dataset. In Fig. 13, we show the same information for the *CIMAT-NAO-A* dataset. The vocabularies with rotation have more detections for this dataset. Also, the vocabulary rotated by θ degrees has significantly more detections than the vocabulary rotated by $-\theta$ degrees. This is because there are more images in the dataset rotated in that direction.

Table 4: Percentages of correct results for the *Toulouse* dataset with rectified images.

Descriptor	Number of tests	Correct tests μ_1	Effectiveness μ_1 (%)	Correct tests μ_2	Effectiveness μ_2 (%)	Average time (ms)
BRIEF	770	562	72.99	743	96.49	87.23
BRIEFROT	770	564	73.25	744	96.62	112.0
ORB	770	554	71.95	734	95.32	77.7
SURF	770	536	69.61	733	95.19	267.9

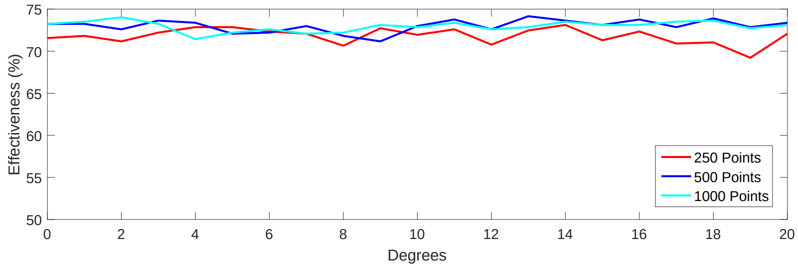

Figure 11: Effectiveness of BRIEFROT for different degrees of rotation for *Toulouse* dataset with rectified images.

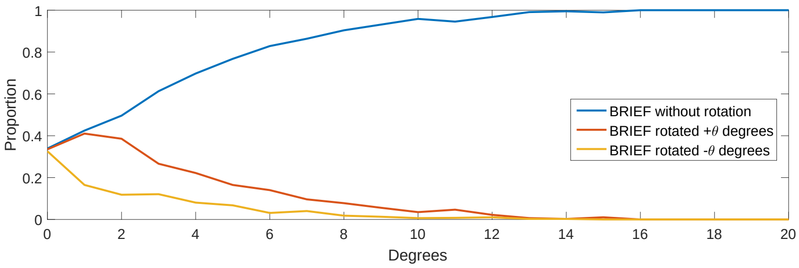

Figure 12: Frequency that a vocabulary gets a minimum distance for *Toulouse* dataset with rectified images.

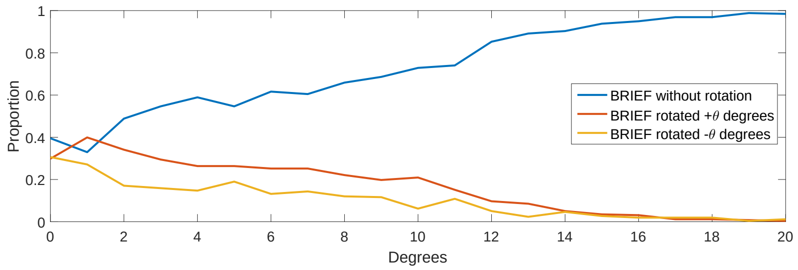

Figure 13: Frequency that a vocabulary gets a minimum distance for *CIMAT-NAO-A* dataset.

Tests were also performed on the images with radial distortion. In this case, the effectiveness decreases significantly with respect to the results obtained without radial distortion, as it can be seen in Fig. 14. The Table 5 shows that the BRIEF vocabulary obtained a higher number of correct results, followed closely by the BRIEFROT vocabulary with 0.26% difference. But, for μ_2, the BRIEFROT vocabulary obtained the best results with 0.26% difference compared to the second best vocabulary (BRIEF).

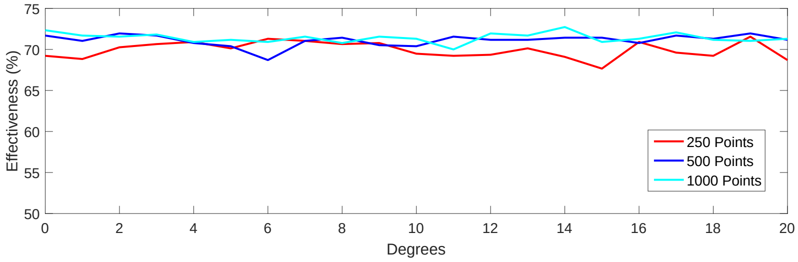

Figure 14: Effectiveness of BRIEFROT for different degrees of rotation for *Toulouse* dataset with distorted images.

Table 5: Percentages of correct results for the *Toulouse* dataset with distorted images.

Descriptor	Number of tests	Correct tests μ_1	Effectiveness μ_1 (%)	Correct tests μ_2	Effectiveness μ_2 (%)	Average time (ms)
BRIEF	770	549	71.3	744	96.62	87.23
BRIEFROT	770	547	71.04	746	96.88	112.0
ORB	770	546	70.91	737	95.71	77.7
SURF	770	538	69.87	725	94.16	267.9

Bicocca 2009-02-25b dataset

For the *Bicocca 2009-02-25b* dataset, in Table 6, three vocabularies obtained the best results for μ_1: BRIEFROT, Color-Random and SURF. The difference between these three vocabularies is in the computation time: SURF consumes much more time. For μ_2, BRIEFROT was the best.

New College dataset

With the *New College* dataset, in Table 7, the SURF vocabulary obtained the best behavior for both evaluation metrics. In both cases, the BRIEFROT vocabulary obtained a good behavior, close to SURF, but BRIEFROT consumes less than half of the time required by SURF.

As an illustration, we give in Figs. 15 and 16 a few examples of test images for the BRIEFROT vocabulary from the views represented in Fig. 7 as blue robot locations. In Fig. 15, images evaluated with $\mu_1 = 1$ (i.e., perfect match) to have the image $\mathcal{I}^*_3$ as closest key image are shown; similarly, in Fig. 16 we present images associated by the localization method with the key image $\mathcal{I}^*_{142}$ with $\mu_1 = 1$. In both cases, it can be appreciated that the appearance of the sequence of input images is similar, but with sometimes significant rotations or light effects. Despite of that, the localization method gives the correct most similar key image.

4.4.2 *Repeatability analysis*

An important aspect of the method is that it has a random component, at the level of the construction of the words tree: in the *kmeans++* algorithm, the center of each cluster is chosen randomly. Hence, it

Table 6: Percentages of correct results for the *Bicocca 2009-02-25b* dataset.

Descriptor	Number of tests	Correct tests μ_1	Effectiveness μ_1 (%)	Correct tests μ_2	Effectiveness μ_2 (%)	Average time (ms)
BRIEF	120	111	92.5	116	96.67	73.4
BRIEFROT	120	111	92.5	116	96.67	98.5
Color-Random	120	60	50	69	57.50	72.7
Color-Half	120	57	47.5	67	55.83	36.0
Color-Whole	120	59	49.17	65	54.17	79.4
ORB	120	110	91.67	114	95.00	60.2
SURF	120	111	92.5	114	95.00	120.0

Table 7: Percentages of correct results for the *New College* dataset.

Descriptor	Number of tests	Correct tests μ_1	Effectiveness μ_1 (%)	Correct tests μ_2	Effectiveness μ_2 (%)	Average time (ms)
BRIEF	117	70	59.83	85	72.65	105.9
BRIEFROT	117	70	59.83	87	74.36	134.6
Color-Random	117	48	41.02	69	58.97	110.9
Color-Half	117	34	29.06	64	54.70	60.4
Color-Whole	117	44	37.61	74	63.25	110.5
ORB	117	69	58.97	85	72.65	107.0
SURF	117	74	63.25	89	76.07	302.3

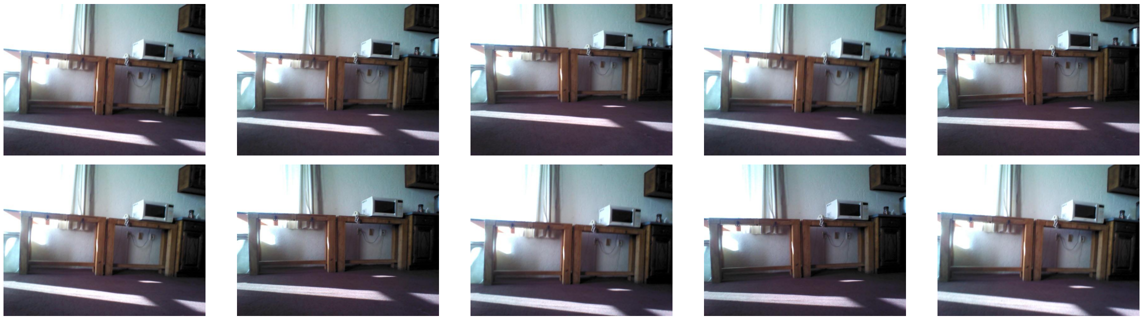

Figure 15: Example of eight similar images $\mathcal{I}$ associated by our method to $\mathcal{I}_3^*$ (first column) of the visual memory (see Fig. 7), i.e., $\mathcal{I}$ such that $\mu_1(\mathcal{I}) = 1$.

is important to evaluate the repeatability of the evaluation results, when generating new vocabularies. We repeated 100 times the building of visual vocabularies, with different features, and tested them for 10 different input images. We collected the average value of the proportion of highest frequency output considering $\mu_1 = 1$. As seen in Table 8, BRIEF, BRIEFROT and SURF have particularly high repeatability rates.

5. CONCLUSIONS

The problem of appearance-based localization of humanoid robots is tackled in this paper, which consists in determining the most similar image among a set of previously acquired images (visual memory) to the current robot view. We used a hierarchical visual bag of words (VBoW) approach to achieve this goal. We evaluated and compared the performance of different local descriptors used to feed the VBoW method: real-valued, binary and color descriptors were compared on real datasets captured by humanoids robots,

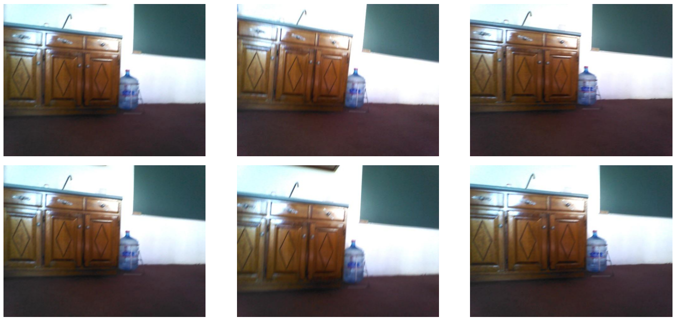

Figure 16: Example of four similar images $\mathcal{I}$ associated by our method to $\mathcal{I}^*_{142}$ (first column) of the visual memory (see Fig. 7), i.e., $\mathcal{I}$ such that $\mu_1(\mathcal{I}) = 1$.

Table 8: Repeatability of visual vocabularies.

Descriptor	Repeatability $(\%)$
BRIEF	100
BRIEFROT	98
Color-Random	87.6
ORB	89.1
SURF	97.2

in particular with a small-size humanoid platform (a NAO robot). We presented a novel use of the BRIEF descriptor suited to the VBoW approach for humanoid robots: BRIEFROT. According to our evaluation, the BRIEFROT vocabulary is very effective in this context, as reliable as SURF to solve the localization problem, but in much less time. We also show that keypoints-based vocabularies performed better than color-based vocabularies.

In this article, we have assumed that the visual memory is given and the construction of the visual memory is left as future work. However, the construction of the visual memory is critical and may have a great impact on the performance of the localization stage and on the autonomous visual navigation. The visual memory should come with no gaps, i.e., every pair of consecutive key images should have a minimum amount of common visual information. However, we stress that the localization stage is not as dependent on the visual memory quality as the navigation stage, since the localization will give an approximate location whenever some similarity in appearance is detected. Also as future work, we will implement the method onboard the NAO robot using a larger visual memory. We will explore the combination of visual vocabularies to robustify the localization results. We also wish to use the localization algorithm in the construction of the visual memory to identify revisited places.

REFERENCES

[1] J. Courbon, Y. Mezouar, and P. Martinet. Autonomous navigation of vehicles from a visual memory using a generic camera model. *IEEE Trans. on Intelligent Transportation Systems*, 10(3):392–402, 2009.

[2] A. Diosi, S. Segvic, A. Remazeilles, and F. Chaumette. Experimental evaluation of autonomous driving based on visual memory and image-based visual servoing. *IEEE Trans. on Intelligent Transportation Systems*, 12(3):870–833, 2011.

[3] H. M. Becerra, C. Sagüés, Y. Mezouar, and J. B. Hayet. Visual navigation of wheeled mobile robots using direct feedback of a geometric constraint. *Autonomous Robots*, 37(2):137–156, 2014.

[4] H. M. Becerra. Fuzzy visual control for memory-based navigation using the trifocal tensor. *Int. Journal of Intelligent Automation and Soft Computing (AutoSoft)*, 20(2):245–262, 2014.

[5] W. Burgard S. Thrun and D. Fox. *Probabilistic Robotics*. The MIT Press, 2006.

[6] J. Ido, Y. Shimizu, Y. Matsumoto, and T. Ogasawara. Indoor navigation for a humanoid robot using a view sequence. *Int. Journal of Robotics Research*, 28(2):315–325, 2009.

[7] J. Delfin, H. M. Becerra, and G. Arechavaleta. Visual path following using a sequence of target images and smooth robot velocities for humanoid navigation. In *IEEE Int. Conf. on Humanoid Robots*, pages 354–359, 2014.

[8] I. Ulrich and I. Nourbakhsh. Appearance-based place recognition for topological localization. In *IEEE Int. Conf. on Robotics and Automation*, pages 1023–1029, 2000.

[9] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE Int. Conf. on Computer Vision*, pages 1–8, 2003.

[10] T. Botterill, S. Mills, and R. Green. Bag-of-words-driven, single-camera simultaneous localization and mapping. *Journal of Field Robotics*, 28(2):204–226, 2011.

[11] D. Galvez-López and J.D. Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Trans. on Robotics*, 28(5):1188–1197, 2012.

[12] N. G. Aldana-Murillo, J. B. Hayet, and H. M. Becerra. Evaluation of local descriptors for vision-based localization of humanoid robots. In *Proc. of the Mexican Conf. on Pattern Recognition*, 2015.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346 – 359, 2008.

[14] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: Binary robust independent elementary features. In *Eur. Conf. on Computer Vision*, pages 778–792, 2010.

[15] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In *IEEE Int. Conf. on Computer Vision*, pages 2564–2571, 2011.

[16] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *Eur. Conf. on Computer Vision*, pages 430–443, 2006.

[17] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 2161–2168, 2006.

[18] A. Bonarini, W. Burgard, G. Fontana, M. Matteucci, D.G. Sorrenti, and J.D. Tardos. Rawseeds: Robotics advancement through web-publishing of sensorial and elaborated extensive data sets. In *Int. Conf. on Intelligent Robots and Systems*, 2006.

[19] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman. The new college vision and laser data set. *Int. Journal of Robotics Research*, 28(5):595–599, 2009.

[20] P.F. Alcantarilla, O. Stasse, S. Druon, L. M. Bergasa, and F. Dellaert. How to localize humanoids with a single camera? *Autonomous Robots*, 34(1-2):47–71, 2013.