Object recognition using region detection and feature extraction

E. Jauregi, E. Lazkano and B. Sierra
Robotics and Autonomous Systems Group
Computer Science and Artificial Intelligence Department
University of Basque Country, Donostia
ekaitzji@hotmail.com, http://www.sc.ehu.es/ccwrobot

***Abstract*— Feature extraction in images is an important issue in mobile robotics, as it helps the robot to understand its environment and fulfil its objectives. This paper summarises a new two-step algorithm based on region detection and feature extraction that aims to improve the relevance of the extracted features in order to reduce the superfluous keypoints to be compared and, at the same time, increase the efficiency by improving accuracy and reducing the computational time. Experiments are carried out for the task of door handle identification during robot navigation and extended to the traffic signal identification problem.**

I. INTRODUCTION

There are several applications where it is mandatory to recognise objects or scenes: image retrieval [11], mobile robot localisation [6][22] and SLAM [19], physical sign recognition [18] and automatic guidance of vehicles [17]. Although different sensors can be used, vision turns out to be the most appropriate one due to the rich information that can be extracted from images. However, object identification becomes a complex task specially due to varying environmental conditions and changes in object scale and camera viewpoint.

Recently, several methods have been developed for extracting invariant local image descriptors. SIFT (*Scale Invariant Feature Transform*) [13] is a method to extract features invariant to image scaling and rotation, and partially invariant to change in illumination and 3D camera viewpoint. Those properties make it suitable for being used in robotics applications, where changes in robot viewpoint distort the images taken from a conventional camera. But it is known that SIFT suffers from a high computational payload.

SURF and USURF[2] (*Speeded-Up Robust Features*) are other detector-descriptor algorithms developed with the aim of speeding up the keypoint localisation step without losing discriminative capabilities. The scale-space is analysed by up-scaling the filter size instead of by iteratively reducing the image size as occurs in the SIFT approach.

This paper presents a new two-step algorithm that aims to reduce the superfluous keypoints obtained by methods like SIFT and (U)SURF by firstly extracting the region of the image where it is likely that the object or objects to be identified are located.

This work has been supported by the Basque Country Government (Research Team Financing Program) and by the University of the Basque Country

II. TWO STAGE PROCEDURE

As mentioned before, instead of computing the invariant features of the whole image, the approach presented here aims to reduce the size of the image to be processed by extracting the portion of the image that, with high probability, will contain most of the crucial features. Next subsections summarise these two steps.

A. Extracting the region of interest

Several methods can be used for extracting the region of interest (ROI). A priori knowledge of objects to be identified can be used, for instance, shape or colour information [9], but this would make the method specific for a concrete environment or class of objects. Instead, the approach can be generalised by scanning the image for continuous connected regions or *blobs*. A blob (binary large object) is an area of touching pixels with the same logical state. Blob extraction, also known as region detection or labelling, is an image segmentation technique that categorises the pixels in an image as belonging to one of many discrete regions. The process consists of scanning and numbering any new regions that are encountered, but also merging old regions when they prove to be connected on a lower row. Therefore, the image is scanned and every pixel is individually labelled with an identifier which signifies the region to which it belongs (see [8] for more details).

Blob detection is generally performed on the resulting binary image from a thresholding step. Instead, we apply the SUSAN (Smallest Univalue Segment Assimilating Nucleus) edge detector [21], a more stable and faster operator.

The blob extraction process can give many different regions for a single image. In order for a blob to be confirmed as a candidate, the result of the blob detection process should be filtered and false positives should be discarded. Different filtering techniques can be used. For instance, in [9] colour information inside the candidate region is used for handle identification. Alternatively, in [23] location-related pixel information is used for blob discrimination where the aim is to count persons in images taken by a surveillance system. A similar approach is used in our proposal where blobs that are not consistent with the defined size restrictions are discarded.. The size restrictions depend on the proximity the images are taken at.

Depending on the task to be solved, images need not to be restricted to a single object to be identified and hence, could

contain more than one region to be detected and extracted. Once the most interesting blobs are located, blob's length and width values are used to find their centre and afterwards, a square subimage is extracted for each one. The size of the square is determined by the maximum value between the length and width of the candidate blob. The subimage is then scaled to a fixed size to obtain the portion of the image that, with high probability should contain the object of interest, i.e a ROI. Then, the keypoint extraction and matching procedure is performed to every extracted ROI.

B. Feature extraction

Several invariant feature extraction techniques exist in the literature. **SIFT** features are extracted according to the following procedure:

1) Detect scale-space extrema, searching over all scales and image locations. Potential interest points invariant to scale and orientation are efficiently computed using a DoG (differential of Gaussian) function.
2) Localise keypoints, detecting local extrema and removing low contrast points or candidates located in edges.
3) Assign orientations to the candidate keypoints based on local image gradient directions.

After keypoints are localised, for each keypoint a descriptor is computed by calculating an histogram of local oriented gradients around the interest point and storing the bins in a 128 dimensional vector. These descriptors can be then compared with stored ones for object recognition purposes.

For on-line applications, each one of the three steps (detection of local extrema, keypoint description computation and keypoint matching) should be computed faster.

SURF [2] descriptors are computed in two steps. First, a reproducible orientation is found based on the information of a circular region around the interest point. This is performed using Haar-wavelet responses in the x and y directions at the scale the interest point was detected. The dominant orientation then is estimated by calculating the sum of all responses within a sliding window. Next, the region is split up into smaller square subregions and some simple features are computed (weighted Haar wavelet responses in both directions, sum of the absolute values of the responses). This yields a descriptor of length 64, half size of the original SIFT descriptor and hence, offers a less computationally expensive matching process.

The upright version of SURF, named **USURF**, skips the first step of the descriptor computation process, resulting in a faster version. USURF is proposed for those cases for which rotation invariance is not mandatory. A similar approach is applied to the SIFT algorithm by Ledwich and Williams [11], assuming that the viewpoint of the robot is relatively stable to rotation around the view axis.

In all the three methods, the matching process requires a reference database where the keypoints of the interesting objects are stored, together with a matching criteria. Figure 1 summarises the algorithm.

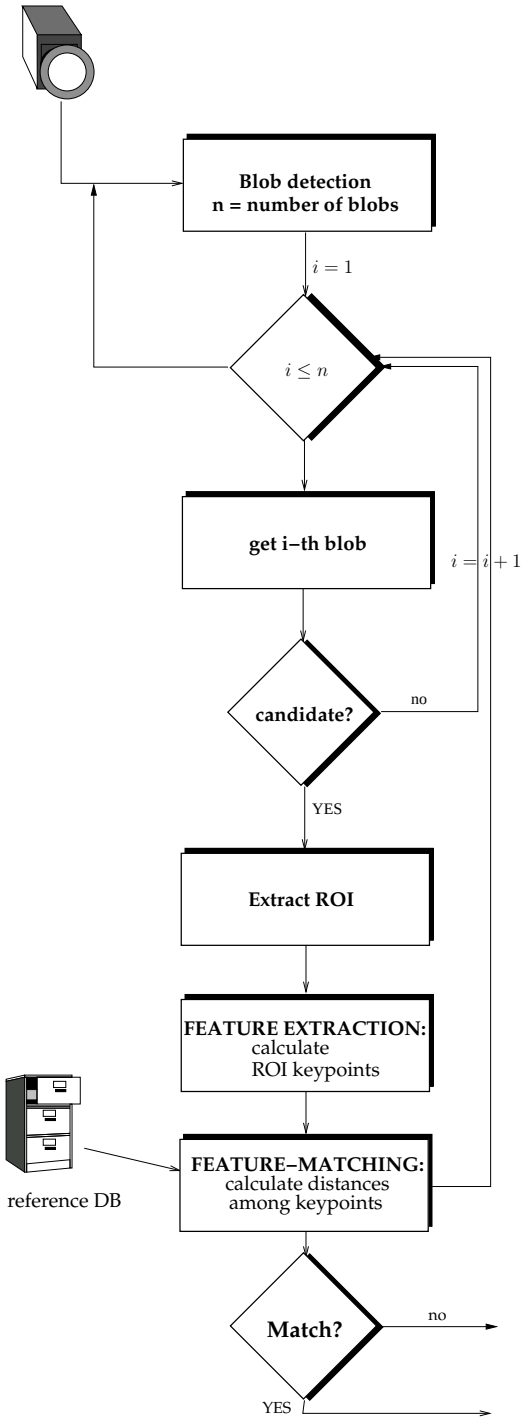
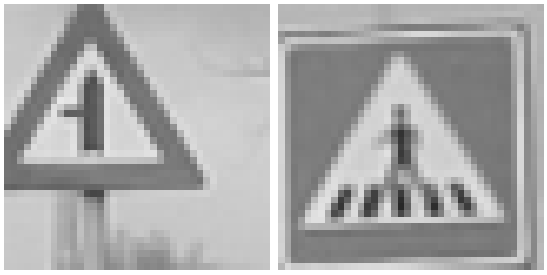


Fig. 1. Flow diagram

III. Experimental application I: Traffic signal identification

In order to test the capabilities of the proposed approach, it has been applied to the road signal identification problem using the database used in [7]. This database contains 360×270 sized 48 images and three signals are to be recognised: pedestrian, bicycle and crossing[1]. We used a reference database containing only three images, one per signal. The images contained very cluttered views and again, the blob properties have been restricted in order to reduce the blobs to be considered. Here, each image can contain more than one signal to be extracted. Figure 2 shows an example of the detected blobs and the subimages extracted afterwards.

(a)



(b)

Fig. 2. Detected blobs and extracted signals

To calculate the classification accuracy we assume that an image can contain at most one signal of each type, i.e. the can not be more than three signals per image. In this way, the accuracy is computed summing up the true positive and true negative cases. But accuracy is considered a fairly crude score that does not give much information about the performance of a categorizer. Instead, F1 measure is employed as the main evaluation metric as it combines both precision and recall into a single metric and favors a balanced performance of the two metrics (see equation 1) [3].

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (1)$$

[1]http://www.cs.rug.nl/ imaging/databases/traffic_sign_database/traffic_sign_database.html

Figure 3 shows the obtained results. Both, accuracy and F1 measure are shown in order to evaluate the goodnesses/weaknesses of the approaches. For ROI sizes larger that 150 the number of false positives increases considerably when applying SIFT. Best results are obtained for small ROI sizes because the signal size in the images approximates that geometry. In spite of the difficulty of the task, the accuracy raised up to 95% for ROI size of 40×40. For larger ROI sizes, the scaling of the subimage seems to drasticaly affect the stability of the keypoints producing a degradation on the classification performance.

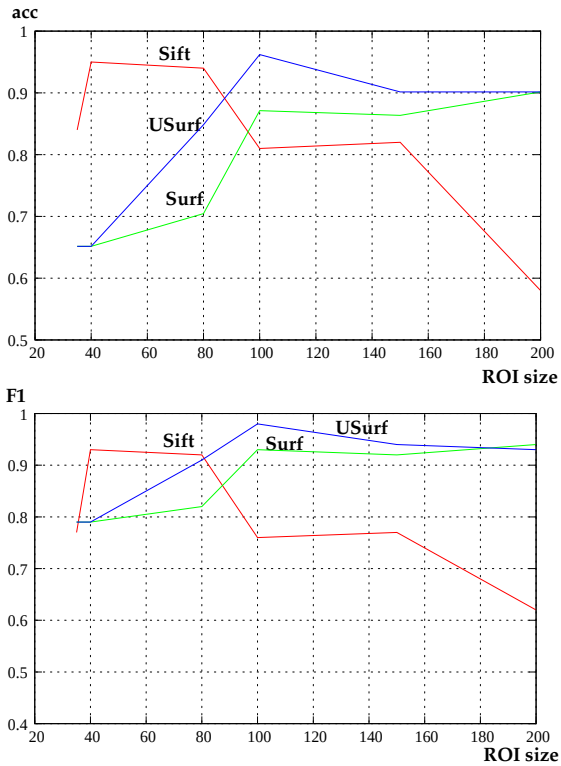


Fig. 3. Results over ROI size

However, both SURF and USURF need larger ROI sizes compared to SIFT to achieve their best classification performances. Small ROI sizes are not appropriate for the speeded up variants because the number of keypoints extracted is very small, as can be seen in table I.

TABLE I
AVERAGE NUMBER OF KEYPOINTS IN THE IMAGES OF THE SIGNALS REFERENCE DATABASE

	30	40	80	100	150	200
SIFT	0.66	18.66	49.66	72	119.33	159
(U)SURF	0.33	1.66	14	24	48.33	59.66

IV. Experimental application II: Handle recognition task

In order to test the adequacy of the proposed method, it has been applied to the door handle recognition task. Office-

like indoor environments are usually cluttered environments. Mobile robots have to navigate safely while recognising locations for task delivery. Many navigation tasks can be fulfilled by point to point navigation, door identification and door crossing [12]. Hence, endowing the robot with the door identification ability would undoubtedly increase its navigating capabilities.

Several references can be found that tackle the problem of door identification. For instance, in [10] doors are located in a map and do not need to be recognised, but rectangular handles are searched for manipulation purposes. The handles are identified using cue integration by consensus. However, most of the references we found are vision-based approaches [20] and focus on edge detection [4][14][15].

But navigating in narrow corridors makes it difficult to identify doors by line extraction due to the inappropriate viewpoint restrictions imposed by the limited distance to the walls. Therefore, the goal is to try to identify the doors by recognising the handles, extracting the necessary local features needed for robust identification.

Instead of processing the whole image, the aim of the new method is to first extract the region of interest (ROI) and process it afterwards.

Some size restrictions have been imposed to the detected blobs in order to ensure that the selected one corresponds, with high probability, to the door handle. Using the blob information, a squared subimage is obtained based on the centre of the blob and scaled to a fixed size.

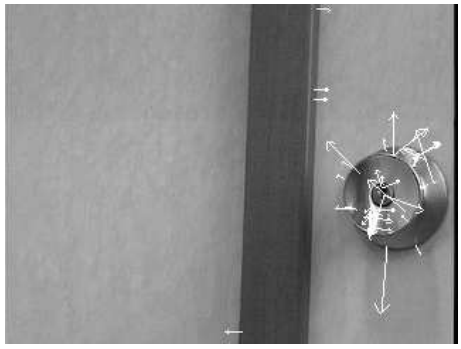
Door blades are poorly featured surfaces, generally speaking. In our robot's environment two kinds of handles are distinguished: circular and rectangular handles. Circular ones are located into pladour (a type of laminated surface) blades. These have the advantage that almost every keypoint is located at the door handle and only a few of them appear at the handle surroundings (see Figure 4(a)). On the other hand, rectangular handles are located into wooden door blades that are not textureless and therefore, keypoints appear outside the handle (see Figure 4(b)).

Figure 5 shows examples of the result of extracting the ROI for both handle types. Obviously, the rectangular handle identification problem is more difficult due to the non symmetrical property of the handle itself and the featured blades they are located into.

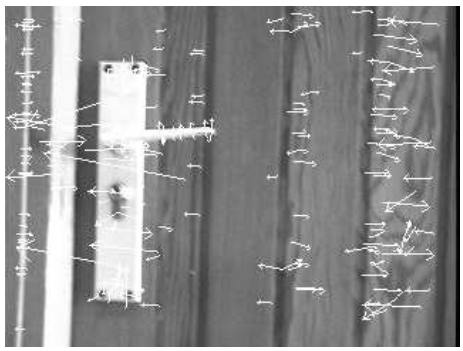
A. Off-line experiments and results

At this stage of the experimentation, the handle recognition was performed separately for each class of handle, using separate databases for each one.

All the images (references and testing DBs) were taken while the robot followed the corridors using its local navigation strategies and therefore, they were taken at distances at which the robot is allowed to approach the walls. Both databases contained positive and negative cases. It must be mentioned that the test cases were collected in a different environment from where reference cases were taken. Therefore, the test databases did not contain images of the handles in the reference database.

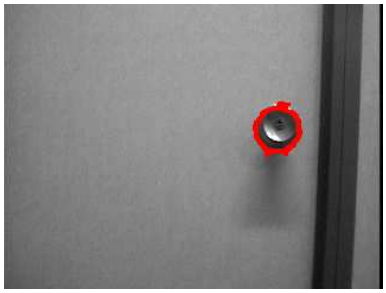


(a) Circular Handle



(b) Rectangular Handle

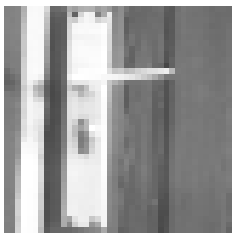
Fig. 4. SIFT keypoints in 320×240 images



(a) Circular handle: blob

(b) Circular handle: ROI

(c) Rectangular handle: blob



(d) Rectangular handle: ROI

Fig. 5. Blob extraction and ROI scaling

The databases characteristics were the following:

1) For circular handles, the database contained about 3000 entries and the reference database 32 images (scaled to the appropriated ROI size)
2) For rectangular handles, the size was bigger with about 5000 images and the reference database containing 19 images, also scaled to the adequate ROI size.

The keypoint matching criteria used in these experiments is the 1-NN, the same proposed in [13].

TABLE II
BEST RESULTS

	Circular		Rectangular	
	Acc.	F1	Acc.	F1
SIFT	62.39	59.25	**55.41**	**36.22**
SURF	72.5	69.17	42.34	32.41
USURF	**72.5**	**69.19**	47.0	34.44

a) No ROI extraction

	Circular			Rectangular		
	acc.	size	F1	acc.	size	F1
SIFT	91.82	150	86.33	**91.52**	80	**66.92**
SURF	92.94	100	89.05	87.77	40	40.6
USURF	**94.35**	150	**91.11**	87.77	40	40.6

b) Region labelling + feature extraction

	Circular			Rectangular		
	acc.	size	F1	acc.	size	F1
SIFT	94.48	240	90.98	**92.67**	80	**69.96**
SURF	95.70	80	93.13	89.87	100	66.84
USURF	**96.09**	150	**93.68**	89.71	100	66.62

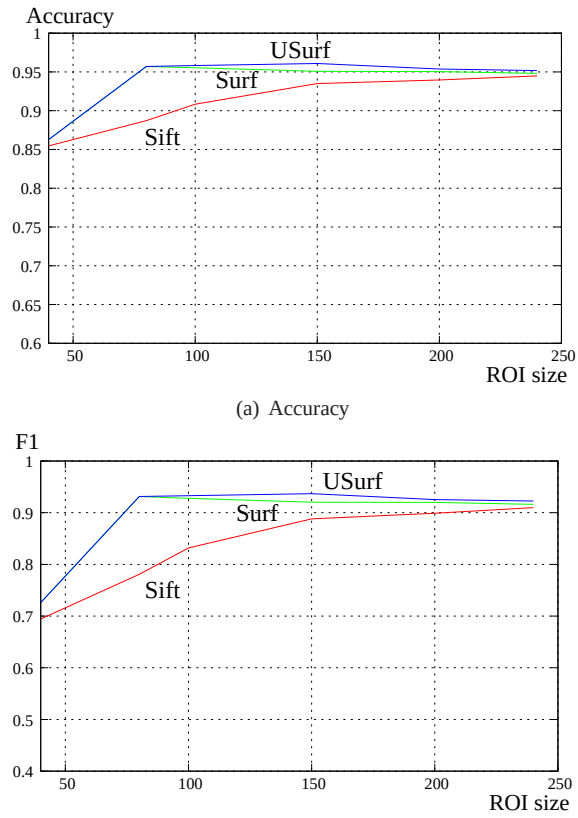
c) Region labelling + feature extraction + MR

Table II shows the best accuracy and corresponding F1 measure values obtained for the different approaches, together with the associated ROI size.

Table II-a) shows the results obtained applying the keypoint extraction methods to the original images (without ROI extraction). These performances are calculated for comparison purposes. USURF outperforms SIFT in both, accuracy and F1 measure in the case of circular handles, but its performance degrades for rectangular ones.

Table II-b) shows the improvement obtained applying the method proposed in the paper. Again, USURF seems to be the most adequate feature extraction method for circular handles but SIFT is clearly the best approach to be applied to the obtained ROI when non circular handles must be identified. The accuracy is increased in almost 30% for SIFT and about $20-22\%$ for SURF and USURF respectively for the case of circular handles. The improvement is better for rectangular handles, with an increase of 35% for SIFT and about 40% for SURF. Notice that this improvement is also reflected in the F1 measure, proving the adequateness of the methodology.

As mentioned in [16], vision may take advantage of the physical interaction of the agent with its environment. Taking into account the robot's morphology and the environmental niche, more specifically the height at which the camera is mounted on the robot, and the height at which the handles are located on the doors, these handles should always appear at a specific height on the image. The improvement introduced by this *morphological restriction* (MR) is also showed in Table II-c).



(a) Accuracy

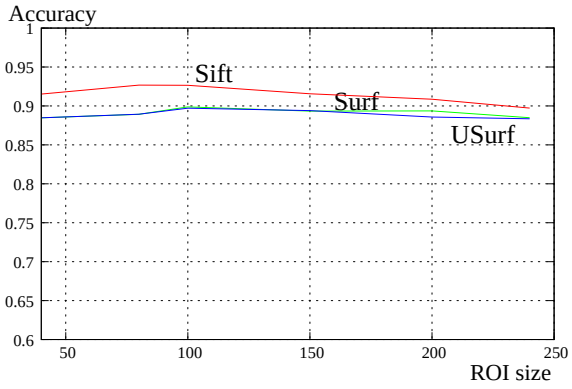
(b) F1

Fig. 6. Circular handles

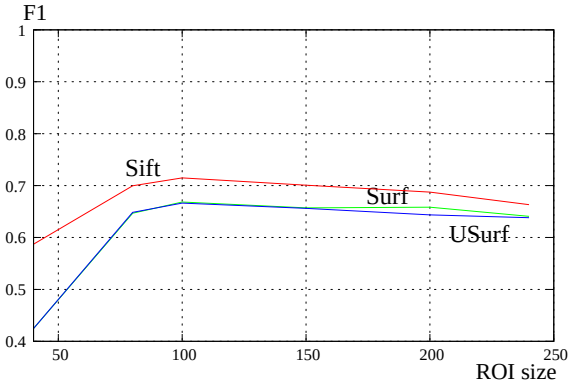
Figures 6 and 7 show the evolution of the performance, for circular and rectangular handles. Both, the classification accuracy and the F1 measure are ploted when increasing the ROI size. These figures correspond to the proposed two-step method together with the morphological restriction previously mentioned. It can be appreciated how USURF is the best approach for circular handles and, on the contrary, SIFT improves speeded-up variants when rectangular handles need to be detected, even for larger ROI sizes.

Table III reflects how the average number of keypoints increases together with the ROI size. These values were calculated using the reference databases used in each experimental trial and hence, are only tentative values. Although the number of keypoints increases together with the ROI size, stable keypoints are found on small image sizes. Increasing the scale of the ROI gives a higher number of keypoints but the repeatability of these new keypoints seemingly is not good.

Table IV shows the average time needed to process a single image (blob location, keypoint extraction and matching against the reference database). Note that the time needed is sensibly higher for larger ROI sizes due to the larger number



(a) Accuracy



(b) F1

Fig. 7. Rectangular handles

TABLE III
AVERAGE NUMBER OF KEYPOINTS

Ref. DB	40	80	100	150	200	320 × 240
Circular	10.09	21.43	41.5	74.03	105.53	49.28
Rectangular	5.95	23.47	38.79	72.10	97.78	139.89

of keypoints that occur in those images.

TABLE IV
COMPUTATIONAL PAYLOAD (S) FOR PROCESSING AN IMAGE USING SIFT

Database	40	80	100	150	200	No ROI
Circular	0.06	0.18	0.22	0.46	0.75	0.31
Rectangular	0.06	0.13	0.17	0.42	0.80	0.69

B. Robot Behaviour

Tartalo is a PeopleBot robot from MobileRobots, provided with a Canon VCC5 monocular PTZ vision system, a Sick Laser, several sonars and bumpers and some other sensors. Player-Stage [5] is used to communicate with the different devices and the software to implement the control architecture is SORGIN [1], a specially designed framework that facilitates behaviour definition. To evaluate the robustness of the handle identification system developed, it has been integrated in a behaviour-based control architecture that allows the robot to travel across corridors without bumping into obstacles. When the robot finds a door, it stops, turns to face the door and knocks it with its bumpers a couple of times asking for the door to be opened and waiting for someone to open it. If after a certain time the door is still closed, Tartalo turns again to face the corridor and continues looking for a new handle. On the contrary, if someone opens the door the robot detects the opening with its laser and activates a door crossing behaviour module that allows it to enter the room. All the computation is carried out in its on-board Pentium (1.6GHz).

Although the off-line experimental step showed a degraded accuracy for the ROI of size 40 extracted using SIFT for circular handles, the short time needed to compute the identification (see Table IV) and the better performance of SIFT for the rectangular handle identification problem that, as stated before, showed to be more difficult, makes it more appealing for the real time problem stated in this paper. Hence, experiments within the real robot/environment system were performed using a ROI size of 40 and applying the SIFT feature extraction method. Also, to make the behaviour more robust, instead of relying on a single image classification, the robot will base its decision upon the sum of the descriptor matches accumulated for the last five consecutive images.

Experiments were carried out in three different environments.

a) Environment 1: circular handles: Figure 8 shows the environment together with the evolution of the sum of the matching keypoints over time. The horizontal line represents the value at which the threshold was fixed. The 18 doors present in the environment were properly identified and no false positives occurred.

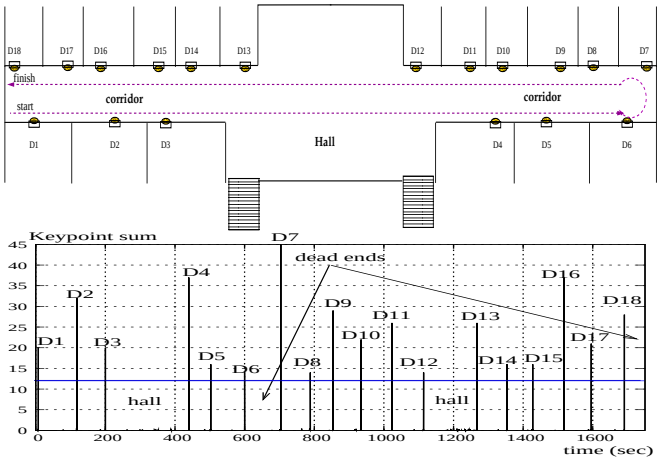


Fig. 8. First floor results

b) Environment 2: rectangular handles: Second floor of the building, lower corridor. 39 rectangular handles were consecutively to be identified. Figure 9 shows the original environment and the evolution of the keypoint sum over time.

The threshold to consider a positive identification was set to 8. The robot started on the left side of the corridor, with its camera pointing to its right and travelled all the way along

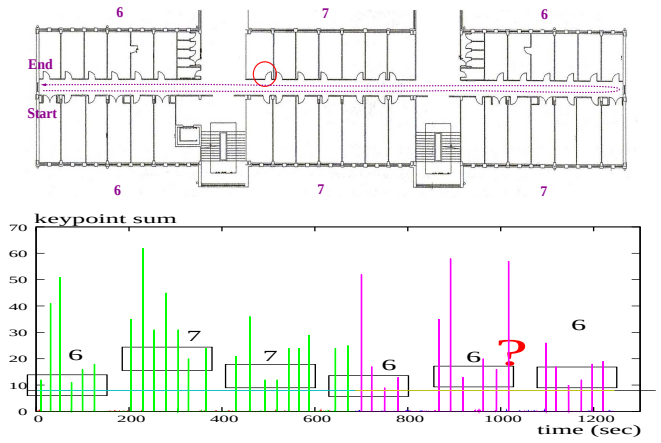


Fig. 9. Second floor results

the corridor successfully fulfilling the sequence of 6+7+7 handle identification, and then turned at the dead end. In its way back, only one of the handles of the central sector of the corridor, the one marked with a circle in the upper image of figure 9, failed to be recognised (6+6+6) and again, no false positives were given. Hence, a success ratio of 0.97 was achieved.

c) Environment 3: mixed identification: Mixed handle identification over time. Experiments were performed on a part of the third floor of the faculty (figure 10). Three circular handles and three rectangular handles were to be identified. The robot was left running for three rounds in the corridor.

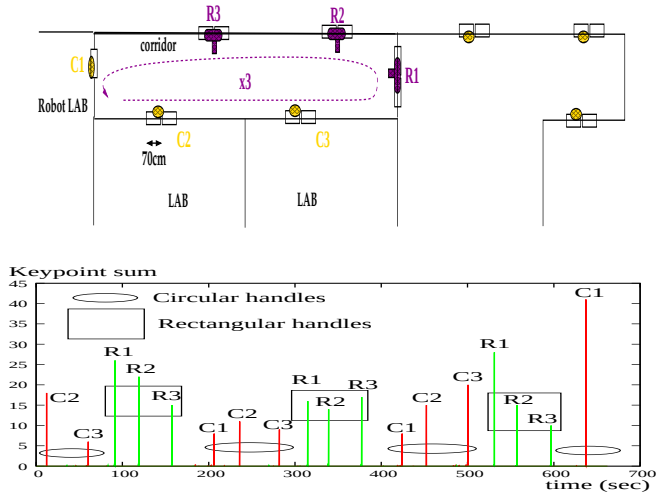


Fig. 10. Mixed environment

In each round, three circular handles and three rectangular ones were positively identified and no false positive was given by the system. The total amount of handles correctly recognised was 18.

Summing up, during the experiments performed within the real robot/environment system, 74 handles were identified, achieving a success of 98.66% and no false positives at all occurred.

V. CONCLUSIONS AND FUTURE WORKS

The paper summarises a new two-step algorithm based on feature extraction that aims to improve the extracted features in order to reduce the superfluous keypoints to be compared and, at the same time, increase the efficiency by improving accuracy and reducing the computational time. The system showed a very low tendency to give false positives while providing a robust identification. Authors' opinion is that the presented approach is appropriate for situations where the background varies a lot in such a way that background characteristics are either irrelevant for object identification or just add noise to the recognition process.

In order to evaluate the method, experiments were performed for road signal recognition and for identifying door handles during robot navigation. Using region detection, the area corresponding to the handle is extracted. ROI extraction improves handle identification procedure and depending on the ROI size, the computational time to classify an image can be considerably reduced.

The developed system outperforms the performances obtained without extracting the ROIs and experiments carried out in a real robot-environment system show the adequateness of the approach. Opposite to the method proposed by the authors in [9] where the door recognition method was mainly based on the color segmentation of the door blades, the method can easily be generalised to other kind of handles. The same blob extraction method could be applied to obtain the region of interest and only the reference database should be adapted to contain the correct reference keypoints.

Still, the performance of the raw ROI extraction technique alone should be calculated in order to measure the improvement more soundly. Also, the keypoint matching criteria has to be analysed more deeply. More sophisticated and efficient algorithms remain to be tested and the performance of different distance measures still needs to be studied. Of course, the proposed method remains open to any other feature extraction method and to any improvement that could be made to the tested ones.

REFERENCES

[1] A. Astigarraga, E. Lazkano, I. Rañó, B. Sierra, and I. Zarautz. SORGIN: a software framework for behavior control implementation. In *CSCS14*, volume 1, pages 243–248, 2003.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *Proceedings of the 9th European Conference on Computer Vision*, 2006.

[3] Shu-Heng Chen. Elements of information theory : Tomas m. cover and joy a. thomas, (john wiley & sons, new york, ny, 1991). *Journal of Economic Dynamics and Control*, 20(5):819–824, May 1996.

[4] C. Eberset, M. Andersson, and H. I. Christensen. Vision-based door-traversal for autonomous mobile robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 620–625, 2000.

[5] B. P. Gerkey, R. T. Vaughan, and A. Howard. The Player/Stage project: tools for multi-robot and distributed sensor systems. In *Proc. of the International Conference on Advanced Robotics (ICAR)*, pages 317–323, 2003.

[6] A. Gil, O. Reinoso, A. Vicente, C. Fernández, and L. Payá. Monte Carlo localization using SIFT features. *Lecture Notes in Computer Science*, 3522:623–630, 2005.

[7] C. Grigorescu and N. Petkov. Distance sets for shape filters and shape recognition. In *IEEE Trans. on Image Processing*, volume 12 (10), pages 1274–1286, 2003.
[8] Berthold K. P Horn. *Robot Vision*. MIT Press, 1986.
[9] E. Jauregi, J. M. Martınez-Otzeta, B. Sierra, and E. Lazkano. Door handle identification: a three-stage approach. In *IAV-07: International Conference on Intelligent Autonomous Vehicles*, volume I, 2007.
[10] D. Kragic, L. Petersson, and H. I. Christensen. Visually guided manipulation tasks. *Robotics and Autonomous Systems*, 40(2-3):193–203, 2002.
[11] L. Ledwich and S. Williams. Reduced SIFT features for image retrieval and indoor localisation. In *Australian Conference on Robotics and Automation*, 2004.
[12] W. Li, H. I. Christensen, and A. Orebäck. An architecture for indoor navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1783–1788, 2004.
[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–100, 2004.
[14] I. Monasterio, E. Lazkano, I. Rañó, and B. Sierra. Learning to traverse doors using visual information. *Mathematics and Computers in Simulation*, 60:347–356, 2002.
[15] R. Muñoz-Salinas, E. Aguirre, and M. García-Silvente. Detection of doors using a genetic visual fuzzy system for mobile robots. Technical report, University of Granada, 2005.
[16] R. Pfeifer and J. Bongard. *How the body shapes the way we think. A new view of intelligence*. MIT Press, 2006.
[17] K. Primdahl, I. Katz, O. Feinstein, Y. Mok, H. Dahlkamp, D. Stavens, M. Montemerlo, and S. Thrun. Change detection from multiple camera images extended to non-stationary cameras. In *Proceedings of Field and Service Robotics*, Port Douglas, Australia, 2005.
[18] C. Roduner and M. Rohs. Practical issues in physical sign recognition with mobile devices. In Thomas Strang, Vinny Cahill, and Aaron Quigley, editors, *Pervasive 2006 Workshop Proceedings (Workshop on Pervasive Mobile Interaction Devices, PERMID 2006)*, pages 297–304, Dublin, Ireland, May 2006.
[19] S. Se, D. G. Lowe, and J. Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *International Journal of Robotics Research*, 21(9):735–758, 2002.
[20] M. W. Seo, Y. J. Kim, and M. T. Lim. *LNAI*, chapter Door Traversing for a Vision Based Mobile Robot using PCA, pages 525–531. Springer-Verlag, 2005.
[21] S. M. Smith and J. M. Brady. SUSAN: a new approach for low level image processing. *International Journal of Computer Vision*, 23(1):45–78, May 1997.
[22] H. Tamimi, A. Halawani, H. Burkhardt, and A. Zell. Appearance-based localization of mobile robots using local integral invariants. In *In Proc. of the 9th International Conference on Intelligent Autonomous Systems (IAS-9)*, pages 181–188, Tokyo, Japan, 2006.
[23] W. Ye and Z. Zhong. Robust people counting in crowded environment. In *Proceedings of the 2007 IEEE International Conference on Robotics and Biomimetics*, pages 1133–1137, 2007.