APPEARANCE-BASED AND ACTIVE 3D OBJECT RECOGNITION USING VISION

F. Trujillo-Romero and M. Devy
CNRS; LAAS; 7 avenue du Colonel Roche, 31077, Toulouse, France
Université de Toulouse; UPS, INSA, INPT, ISAE; LAAS; Toulouse, France

Keywords: Recognition, Appearance-based, Active strategy, Robotics.

Abstract: This paper concerns 3D object recognition from vision. In our robotics context,an object must be recognized and localized in order to be grasped by a mobile robot equipped with a manipulator arm: several cameras are mounted on this robot, on a static mast or on the wrist of the arm. The use of such a robot for object recognition, makes possible active strategies for object recognition. This system must be able to place the sensor in different positions around the object in order to learn discriminant features on every object to be recognized in a first step, and then to recognize these objects before a grasping task. Our method exploits the Mutual Information to actively acquire visual data until the recognition, like it was proposed in works presented in (Denzler and Brown, 2000) and (Denzler et al., 2001): *color histogram*, *shape context*, *shape signature*, *Harris* or *Sift* points descriptors are learnt from different viewpoint around every object in order to make the system more robust and efficient.

1 INTRODUCTION

Object recognition is a task that a human being carries out in an instinctive way. Many factors make difficult such a task: illumination conditions, relative camera-object positions, occlusions, etc. So, endowing a robot of this capability is not easy.

Many researchers in Computer Vision have worked in this topic, providing many publications. During the last decade, many improvements have been provided by the appearance-based methods. Lowe et al. (Lowe, 1999; Lowe, 2001) propose to exploit points extracted from images because their photometric properties are invariant with respect to small camera motions: such points are extracted by Differences of Gaussians (DOG) or other scale-invariant detectors (e.g. the Scaled Salient Patches of Kadir), and then are characterized by a descriptor: the *SIFT* one has been proven to be the more discriminant. Hebert et al. (Johnson and Hebert, 1996) (Zhang and Hebert, 1996) have developed an approach for object recognition, using *Spin Images*, i.e. a map of images acquired when a camera is moved around an oriented point. Fergus et al. (Fergus et al., 2003) and Ke et al. (Ke and Sukthankar, 2004) have proposed independantly PCA methods in order to improve the original Lowe approach based on SIFT descriptors.

In the Computer Vision community, the typical strategy consists in exploiting only one image in order to recognize an object. Using robots to move sensors, allows active recognition methods, since the system can place the sensor in the scene in function of the current status of the recognition process, i.e. of what has been perceived and understood from previous images. In (Trujillo-Romero et al., 2004), we proposed an active recognition method based on the mutual information, by exploiting only color attributes of the analyzed objects. In (Jonquires, 2000), we proposed another active strategy for the recognition and the localization of polyhedral objects from a camera mounted on the wrist of a manipulator: Bayesian Belief Networks (BBN) was built during a preliminary step, to learn (1) how to select the best strategies along the recognition step, (2) the best perceptual groupings to provide hypothesis from an initial image and (3) the best camera positions to verify these hypothesis from next images.

Our problem concerns the recognition and localization of objects to be grasped by a robot: objects must be recognized by the system shownoin figure 1; learning and recognition functions must be performed on line, in a human environment, typically at home, where illumination conditions cannot be controlled. The illumination variability makes non efficient ap-

Trujillo-Romero F. and Devy M. (2009).
APPEARANCE-BASED AND ACTIVE 3D OBJECT RECOGNITION USING VISION.
In *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, pages 417-424
DOI: 10.5220/0001805604170424
Copyright © SciTePress

Figure 1: Our robotic system.

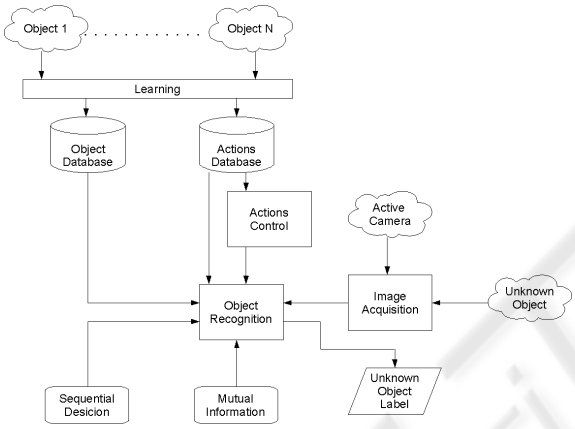


Figure 2: Flow diagram of object recognition system.

proaches of object recognition based on the color. It is the reason we must incorporate more attributes, ad specially invariant ones.

Our experimental setup and a typical recognition scenarios will be described first in section 2. Then the section 3 will present the main visual and decisional functions integrated in our system, e.g. feature extraction, hypothesis generation and verification. Experimental results will be commented in section 4, and finally conclusions and future works will be proposed in section 5.

2 ROBOTICS CONTEXT

2.0.1 Typical Scenario

Our robot has to execute the recognition task, using embedded sensors to acquire data, and motions to improve the recognition efficiency. The general robotics application concerns the *Companion Robot*, as it was called in the *COGNIRON* project [1] , i.e. a Robot used at home by a Human, typically an elder or disabled person, in order to execute services, like *Search Object*, *Pick and Place Object*, *Give Object to User*... The execution of such services involve the integration of many functions in the embedded system, i.e. navigation, docking, manipulation, docking, object perception, user perception, planning... These function executions are supervised by a decisionnal level.

Our Companion robot will execute a global scenario in an autonomous way, but always with the capability to interact with a Human. Let us describe a partial scenario centered on object recognition: one or several objects have been set on a table. Our curious robot decides to recognize these objects; at first it will detect that *Something* has been set on the table. *Something* is coarsely localized from a camera providing a large view of the environment, i.e. on our testbed, the stereo system mounted on the mast. It will compute an initial docking position near the table at a distance of 80cm approximately from *Something*. It controls the arm to put *Something* in the view filed of the cameras. Then in a first step the system executes the Recognition function, with two possible results: (1) if it fails in identifying what is there, even after several motions to evaluate different viewpoints, the robot interprets that it is a new object that must be learnt on line. (2) if it succeeds in recognizing one or several objects, at least an object-based environment model is updated, or the robot task continues depending on the global scenario.

So two steps have to be executed on line. At first, the robot must build an appearance-based representation of every object to be recognized, learning global or local characteristics on every view of every object; it is assumed here that objects are isolated when they are learnt. Afterwards, the robot will have to recognize learnt objects, either isolated or grouped in an object bunch; global attributes are only useful for simple scenes with isolated objets; local attributes are required if objects could be partially occluded by other ones.

During learning or recognition steps, it is mandatory to distinguish the object from its background in every image; the table has an uniform texture and color, in order to make simpler the image segmentation. The object can be placed anywhere on the table, but by now, it is supposed that all the learning or recognition steps could be executed from the same docking position near the table: once docked, the mobile platform stays in the same position until the task

[1] http://www.cogniron.org

ends.

2.0.2 The Experimental System

Figure 1 presents the mobile manipulator; our experiments are only based on the robotic arm and on one camera mounted on the wrist. By now stereovision is only used in order to acquire a dense geometrical model of these objects; in the future we could make profit of 3D characteristics extracted from stereo data in the same framework. Moreover, it is assumed here that the robot is docked along the table; in the future the platform could be moved in order to reach some view points or grasping position.

Our robot will have to grasp any object of common use: telephone, mug, cup, bottle ... Figure 3 presents different objects that have been used to validate our object recognition algorithm.

Figure 3: Objects to be recognized in our database.

3 SYSTEM DESCRIPTION

The learning and the recognition steps require the extraction of characteristics from every image; they will be recorded in a data base for the learning step, and they will be compared with the recorded ones during the recognition step. Figure 2 shows the flow diagram for the object recognition system. In order to extract characteristics from a view of the current scene, the first problem concerns segmentation: how to separate the object or an objects bunch from the background in the image? Considering the table intensity is uniform, the object silhouette can be easily found by a standard active contour, initialized around the barycenter of all edge points extracted in the image: this method assumes that few edge points are extracted on the table. Figure 4 gives a segmentation result with an isolated object on the table.

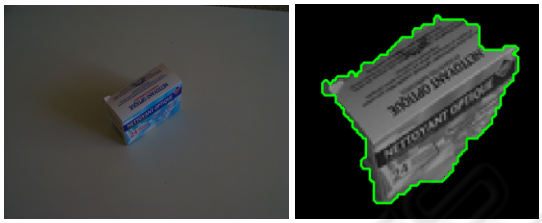


Figure 4: Objet segmentation for processing.

3.1 Learning Step

During the learning step of a class C_k, an object of this class is put alone on the table. Appearance-based characteristics must be learnt for all possible view point on the object. So, our camera is placed on preselected positions on a semisphere centered on the object. The learning trajectory of the camera is precalculated using a classical spherical discretization from an inscribed icosahedron: the advantage of this type of discretization is that the selected vertices on the sphere are equidistant. Every discretized vertex on which the optical center of the camera will be placed, has six neighbours; the camera will be only placed on reachable positions on the semisphere, with respect to the arm model.

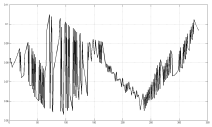
On every view, characteristics are evaluated on pixels inside the object silhouette. It is known from (Denzler et al., 2001) and (Trujillo-Romero et al., 2004), that color histograms are not discriminant enough, so it is proposed to characterize also an object view by the silhouette (shape signature), edge points (shape context) and interest points (Harris and Sift).

The shape signature gives a representation of a closed contour with respect to its barycenter. Shape signatures are commonly used as a fast indexing mechanism for shape retrieval. Since an object will be learnt from many images, only a raw image signature is extracted from the object silhouette, with a normalized radius for exemple $0, 10, 20...$ deg, generating a shape signature as a vector of 36 elements.

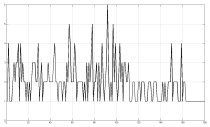
According to Belongie et al. (Belongie et al., 2002), shape context is the relative distribution of points in the plane relative to each point on the shape. In our case, in order to save computation time, the distance and orientation histograms are built only with respect to the barycenter of all edge points inside the object silhouette. We can take, by exemple, three bins

for the radius and eight bins for the orientations, generating a shape context of 24 elements.

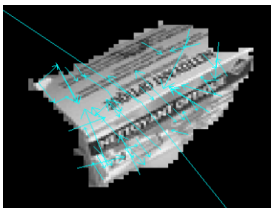
The silhouette presented on figure 4 defines an interest area: the figure 5 shows different characteristics extracted on this area.



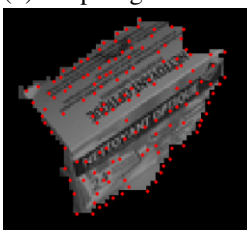
(a) Shape Context



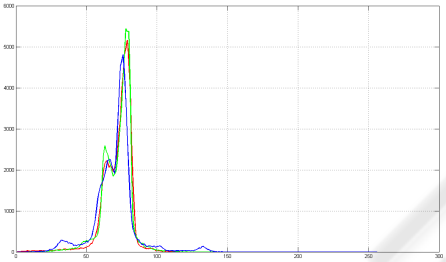
(b) Shape signature



(c) Sift points



(d) Harris points



(e) RGB histogram

Figure 5: Feature extraction of the Objet 1.

These new object characteristics are recorded in the robot database. Because every object is learnt through several views (typically, more than 50), it is represented by a set of conditional probability density functions (PDFs) $p(O_j|C_k,A_t)$, where O_j is the observed characteristic, C_k is the object class and A_t is the camera position on the semisphere. A position A_t can be encoded by the angle value of arm joints, because the camera is mounted close to the end effector, with a known hand-eye transform (especially, a tilt angle of 22.5 deg with respect to the effector reference frame). So the set of camera positions correspond to a set of discrete values for joint angles: each position a_t is defined as:

$$A_t = (q0_t, q1_t, q2_t, q3_t, q4_t, q5_t)^T$$

where q_i is the angle value for the joint i.

Characteristics O_j are learnt from every position A_t of every class C_k. Observations are given first by five histograms representing $(O_1)(O_2)(O_3)$ the chromatic intensity extracted from the normalized (r,g,b) components, (O_4) edge points, represented by the shape context and (O_5) the silhouette, represented by a contour signature. Every interest point is also represented by an array, the corresponding SIFT descriptor; we extract both the Harris points and the DOG ones.

Interest points are local and could be matched even if the object is occluded; histograms are computed in the whole interest area, are global, so that they will be useful only if the object is seen alone.

$$I_p = \begin{pmatrix} I_r(p) \\ I_g(p) \\ I_b(p) \end{pmatrix} = \frac{1}{n}\sum_{i=0}^{n-1} I_i \begin{pmatrix} I_r(i) \\ I_g(i) \\ I_b(i) \end{pmatrix} \quad (1)$$

For example, for the color attributes, equation 1, I_p represents the mean of color features on an image p. I_r,I_g,I_b are the normalized color values for each pixel of that image. And n is the total number of pixels of the image. We use these values to compute the probability of observing a given object characteristic O_j on an object from the class C_k when the camera has been set to a given position A_t.

3.2 Recognition Step

This section presents the active strategy to cope with 3D object recognition. It is well known that active perception is very efficient when a robot tries to recognize an object. J.Denzler and C.Brown (Denzler et al., 2001) proposed to select successive sensor configurations in order to identify from images, objects known by characteristics recorded in a learnt database. In the same way, mutual information will be used here in order to reduce the uncertainty of the recognition task. Let us note x_t the state after t iterations, of the recognition process applied on a static scene.

At each iteration t, an action A_t is executed; it consists in moving the camera on a new view point; so actions and camera positions are noted A_t. From this position a new observation o_t will be extracted from acquired data. Let us recall that the entropy on x_t measures the uncertainty of a random experiment based on x_t. When exploiting a new observation o_t on x_t, the mutual information measures the impact of o_t in decreasing the uncertainty on x_t. So the optimal action A_t must optimize the mutual information between x_t and o_t. It can be defined as:

$$I(x_t;o_t|A_t) = H(x_t) - H(x_t|o_t,A_t) \quad (2)$$

where $H(\cdot)$ denotes the entropy on a continuous distribution

$$H(x_t) = -\int_{x_t} p(x_t)\log p(x_t)dx_t \quad (3)$$

So, the mutual information is expressed by

$$I(x_t;o_t|A_t) = \int_{x_t}\int_{o_t} p(x_t)p(o_t|x_t,A_t)log\frac{p(o_t|x_t,A_t)}{p(o_t|A_t)}do_t dx_t \quad (4)$$

where $p(o_t|x_t,A_t)$ is the perception model, or the likelihood function to observe o_t from the state x_t, and $p(o_t|A_t)$ is the probability to extract o_t whatever the state x_t.

An optimal action A_t^* that maximizes mutual information is given by

$$A_t^* = max_{A_t} I(x_t;o_t|A_t) \quad (5)$$

Then, the action A_t is executed (here the camera is placed in the corresponding position), an image is acquired from A_t, the true o_t is extracted, and the state estimate can be updated by the Bayes law:

$$p(x_{t+1} = p(x_t|o_t,A_t) = \frac{p(o_t|x_t,A_t)p(x_t)}{p(o_t|A_t)} \quad (6)$$

So this iterative framework is applied in our active recognition method. Let us begin with the simplest scenario: an object from the learnt class k, is presented to the system. The camera is placed randomly on an initial position A_0, the first observation O_0 is made, and the firt update is made from equ.6. The first optimal action is searched from equ. 5: here in order to save time, the maximisation is done only on the close positions from A_0, typically on the 6 neighbours.

Here distributions are discrete; if N classes, $k \in 1,n$, have been learnt, then:

$$x_t = (P_{t,1},P_{t,2}....P_{t,k}....P_{t,N})^T \quad (7)$$

where $P_{t,k}$ is the probability that an object from the class k is present in the analyzed scene. Initially, without any contextual information, uniform probabilities are assigned to each learnt class $k, k = 1,N$; so $P_{0,k} = \frac{1}{N}$. At each step, probabilities $P_{t,k}$ are updated for all classes, reenforcing probability of possibles ambiguous classes and in the other side, decreasing probabilities for non similar classes. This procedure iterates in a sequential way until that the probability of the most probable class exceeds a given threshold.

4 EXPERIMENTS AND RESULTS

This section presents the evaluation of our method. We used 8 different objects, shown in Figure 3. Every object is learnt from a set of images acquired by moving the camera on the semisphere. In Figure 6 we can see severals images of the object that we use for make this test.

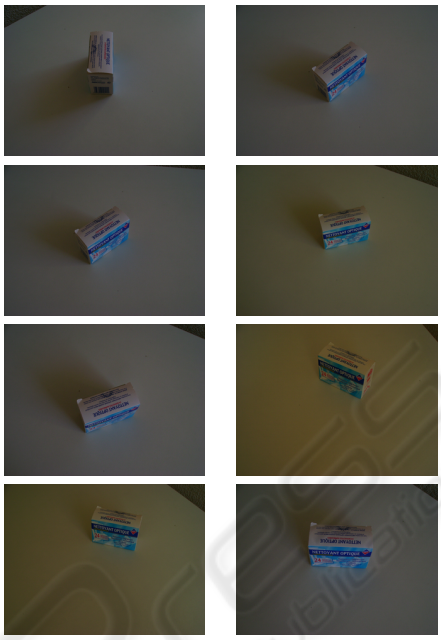


Figure 6: Several images of the objet 1 in learning phase.

Then some more images are acquired to evaluate our recognition approach. Let us consider four situations for testing our algorithm:

1. Only one known object in the scene.
2. The same situation, but with occlusions.
3. An unknown object is presented to system.
4. several known objects are put on the table, with mutual occlusions.

4.1 Case 1: An Isolated Object

Images on figure 7 have been acquired from different sensor positions during the recognition step. In these images we can see that the object position is distinct from that one in which the system learnt the object features.

The graphic shown on figure 8 is the result of the recognition step. We can observe the successive probabilities of having an object of a given class on the image. After the first image, the system generates three hypothesis, on classes 1, 4 and 5, but after some images, new observations reenforce only the hood hypothesis, and the object is classified in the class 1.

We have repeated this experiment about thirty times for every object class presented on figure 3. We obtained the matrix of confusion presented on table 1. One can see very good recognition rates.

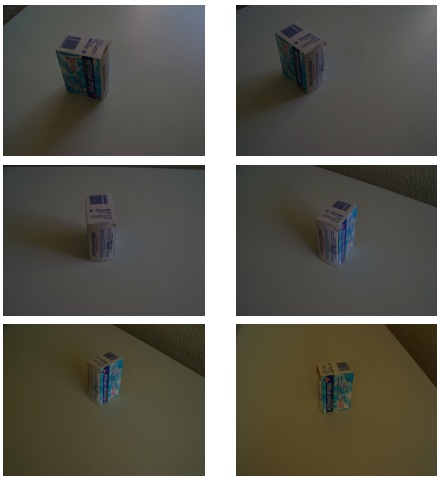


Figure 7: Recognition of a known isolated object.

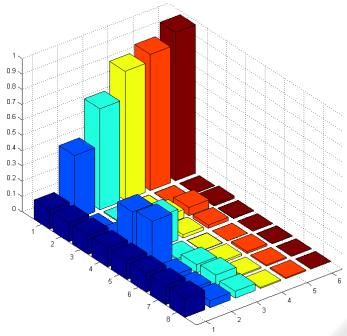


Figure 8: Recognition result for a known isolated object.

Table 1: Matrix of confusion.

Object	1	2	3	4	5	6	7	8
1	27	0	1	0	1	1	0	0
2	1	25	0	0	2	0	1	1
3	0	2	20	2	1	0	5	0
4	0	2	1	26	0	0	1	0
5	0	1	0	0	28	0	1	0
6	0	0	0	0	0	30	0	0
7	2	3	2	0	1	0	22	0
8	0	0	0	0	0	0	0	30

4.2 Case 2: A Partially Occluded Object

Now the system must recognize an learnt object but with occlusions. The initial image is shown on figure 9. In these images an object from class 1 is partially hidden by a big letter A.

The result of the recognition process is shown on figure10. It is possible to observe the doubt of the system since the noise introduced by the letter A, makes the system believe that another object is on the table. But when advancing in the recognition process the system reenforces the hypothesis about the presence of an object of class 1 in the scene, and the system finally converges on the good solution.

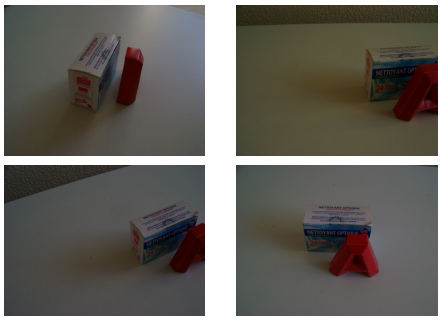


Figure 9: Recognition of a partially occluded object.

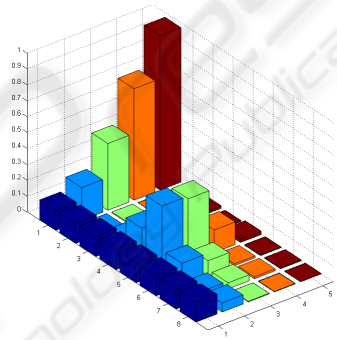


Figure 10: Recognition result with a partially occluded object.

4.3 Case 3: An Unknown Object

Now, the region *Something* extracted on the first image, is the image of an unknown object. In order to avoid errors on such a situation, we need to incorpore a new class: the objet class NULL. If not, if the perceived object has not been previously learnt, our system could return an unpredictable answer according to the features which have been extracted from the image and the ones that have been learnt.

Thanks to the object class NULL, the system is able to make the difference between an object learnt in its database and another which looks like to this one.

Figure 11 shows images acquired successively on an unknown objectduring the recognition process. Figure 12 shows how the system updates the probabilities of the learnt classes along iterations. Initially the system doubts between the object NULL and the objects belonging to classes 3,5 and 6 for finally converging to the obviousness that the object it sees, is not similar to any one that have been learnt.

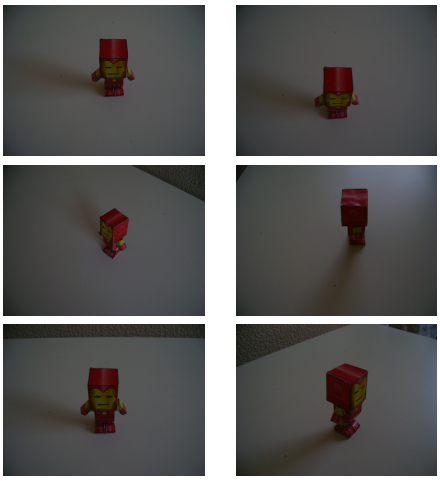


Figure 11: Recognition of an unknown object.

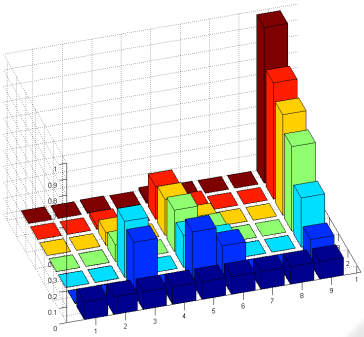


Figure 12: Recognition result for an unknown object.

4.4 Case 4: Several Known Objects

Finally, let us consider a situation with several objects put down on the table, with mutual occlusions depending on the camera position with respect to the scene. Figure 13 shows the twelve successive images (from left to right, and top to bottom) acquired on this scene, until three objects have been recognized. On the first line of figure 14, it supposes equal probabilities 1/9 to find on the table either one object from the eight classes or an unknown object. On the first image, only two objects are seen: after the analysis of this view on the second line of figure 14, the probability to have objects from classes 1 (*Box*), 5 *Bottle* and NULL are higher: five images are sufficient in order to confirm that an object of the class *Box* is in the scene (the higher probability on the first column, line six on figure 14).

Then, this objet class is inhibited: it means that the system does not consider this class in the following steps. It is the reason why on figure 14, the probability to have an objet from class 1 becomes 0 after column 6. The system selects camera positions in order to confirm that an object of class 5, the bottle, is on the table (three images third line on figure 13. After 9 images, the probability to have such an object on the scene, is over a threshold and finally the three last images allows to confirm that an object of class 7, initially occluded, is also on the table.

Figure 13: Recognition with several objects.

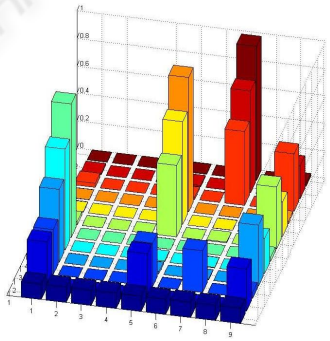


Figure 14: Recognition result with several objects.

5 CONCLUSIONS AND PERSPECTIVES

In this paper, an object recognition system has been presented, providing good performances with a recognition rate of 98%. This recognition rate is achieved thanks to the active strategy to generate and verify hypothesis moving the camera; failures could occur because our system is sensible to illumination changes. More generally, our system fails when the object appearances change between images acquired during the learning step and the recognition one. In such

a situation, our system will generate oscillations between the actual object in front of our robot, and other learnt object, close in the feature space. These oscillations or the lack of convergence could be detected at an higher level, so that a recovery action could be performed, like for exemple, ask the user to remove wrong hypothesis.

In a future work, several extensions of the presented approach are foreseen. Then stereovision will be considered to add 3D characteristics in the object descriptions.

REFERENCES

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522.

Denzler, J. and Brown, C. (2000). Optimal selection of camera parameters for state estimation of static systems: An information theoretic approach. Technical Report 732, The University of Rochester, New York.

Denzler, J., Brown, C., and Niemann, H. (2001). *Pattern Recognition – 23rd DAGM Symposium*, chapter Optimal Camera Parameters for State Estimation with Applications in Object Recognition, Springer, Berlin.

Fergus, R., Perona, P., and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, vol.2

Johnson, A. and Hebert, M. (1996). Recognizing objects by matching oriented points. Technical Report CMU-RI-TR-96-04, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Jonquires, S. (2000). *Application des reseaux baysiens la reconnaissance active d'objets 3D: Contribution la saisie*. PhD thesis, LAAS, 7, Av Colonel Roche Toulouse France.

Ke, Y. and Sukthankar, R. (2004). Pca-sift: A more distinctive representation for local image descriptors. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proc. Int. Conf. on Computer Vision*, Corfu, Greece.

Lowe, D. G. (2001). Local feature view clustering for 3d object recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, Hawaii.

Trujillo-Romero, F., Ayala-Ramírez, V., Marín-Hernández, A., and Devy, M. (2004). Active object recognition using mutual information. In *MICAI 2004: Third Mexican Int. Conf. on Artificial Intelligence*, vol. 2972 of *Lecture Notes in Computer Science*, Springer.

Zhang, D. and Hebert, M. (1996). Multi-scale classification of 3-d objects. Technical Report CMU-RI-TR-96-39, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.