Keywords: Recognition, appearance-based, active strategy, robotics.

Abstract: This paper concerns a topics studied by many researchers around the world: 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 Mitsubishi PA10-6C: 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 (2) and (Denzler et al., 2001): color histogram, shape context, shape signature, interest points harris et sift 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. But, endowing a robot of this capability is not an easy task. Many researchers in Computer Vision have worked in this topic, providing a very large quantity of publications written about this topics.

Let us cite overall the recent appearance-based methods, e.g. works of D. Lowe (Lowe, 1999) (Lowe, 2001) who proposes 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. 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 (Johnson and Hebert, 1996) (Zhang and Hebert, 1996). The Fergus works (Fergus et al., 2003) and those of Ke et al (Ke and Sukthankar, 2004) have proposed independantly PCA methods in order to improve the original Lowe approach based on SIFT descriptors. We can refer also our own work (Trujillo-Romero et al., 2004), where an approach of active recognition has been proposed by exploiting the color attributes of the analyzed objects. Finally let us cite also an older work in our lab (Jonquires, 2000) in which we proposed an active function, dedicated for the recognition and the localization of polyhedral objects from a camera mounted on the wrist of a manipulator: Bayesian Belief Networks (BBN) have been 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: in this paper only the recognition task is discussed. Many factors makes difficult such a task: illumination conditions, relative camera-object positions, occlusions, etc. Object recognition is made by the system shown in fig 1. Reference view points from which objects could be recognized, were learnt in a first step. Using
robots to move sensors, allows to propose an active object recognition, 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 on the environment from previous images.

Our robot will have to grasp any object of common use: a telephone, a mug, a cup, a bottle... Figure 3 presents different objects that allow us to validate our object recognition algorithm. In the preliminary step not described in this paper, modelling and planning functions have determined grasping positions on every learnt objects. A large majority of these current objects are not polyhedral, so that the task is made more complex.

Moreover, learning and recognition functions must be performed on line, in a human environment, typically at home, and not in a lab where controlled experiments could be executed. For example, the variability of the illumination conditions makes non efficient approaches of object recognition based on the color. It is the reason we must incorporate more attributes in order to make more robust our object recognition system.

Our experimental setup and recognition scenarios will be described first in the 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

2.1 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 is used at home by a user, typically an elder or disabled person, in order to execute services, like Search Object, Grasp Object, Place Object, Make interactions between Objects e.g. manipulate a bottle to pour water in a glass, Give Object to User e.g. the robot holds the object in the user hand... The execution of such services involve the integration of many functions in the onboard 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 User. In this paper, we are only concerned by the object perception. So let us describe a partial scenario centered on this topics: 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 cameras mounted on the wrist in a good position to perceive the table and the object, i.e on a semisphere centered on Something.

---

1http://www.cogniron.org
Then in a first step the system executes the Recognition function, with two possible results: (1) if it fails after several motions to evaluate different viewpoints, the system interprets that it is a new object that must be learnt on line. (2) if it succeeds in recognizing this object, at least an object-based environment model is updated, or the robot task continues depending on the global scenario.

For the learning step, the camera is moved on the semisphere in order to extract the object appearance from different viewpoints. This trajectory has been 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. The new object characteristics are recorded in the robot database, so that this object will be recognized in the future. With the same images, the 3D model of this object is also learnt, from registration and fusion of stereo images, and from bundle adjustment executed on interest points.

During learning or recognition steps, it is mandatory to distinguish the object from its background in every image; the table has a uniform texture and color, in order to make simpler the image segmentation. In our example hereafter, a table without texture is used. The object can be placed anywhere on the table, but by now, it is supposed that all the learning or recognition task could be executed from the same docking position near the table: once docked, the mobile platform stays in the same position until the task end. It means that many positions on the semisphere are in the reachable space.

2.2 The Experimental System

Figure 1 presents the mobile manipulator. This robot is used to place the sensor in the different positions selected on the semisphere centered on the object for learning acquisition.

This work exploits the robotic arm and the stereo rig mounted on the wrist of the arm. By now stereovision is used only to acquire a dense geometrical model of an object to be learnt; such a model is required mainly for grasping planning. For recognition only one camera of the stereo rig is used to build an appearance-based representation of every object to be recognized. It is assumed here that the robot is docked along the table, that it see the table and the object of interest and that the object is isolated during the learning step. By now the presented method is based on global attributes, so that it requires also that the object is isolated when it must be recognized.

Figure 3: Images of objects to be recognized in our database.

3 System Description

The learning and the recognition steps require the extraction of characteristics from every image, to be recorded in a database for the learning, and to be matched with the recorded ones for the recognition. Figure 2 shows the flow diagram for the object recognition system. When extracting global characteristics from a view of an object, the first problem concerns segmentation: how to isolate the object 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 few edge points only are extracted on the table. Color histograms are only evaluated on pixels inside this object silhouette. With respect to (Denzler et al., 2001) and (Trujillo-Romero et al., 2004), based only on color histograms, we propose to represent also the object boundaries, using two descriptors: shape context and shape signature and the use of interest points: Harris and Sift.

The shape signature gives a representation of a boundary with respect to its barycenter; the object boundary is given by its silhouette. 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 example 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 example, three bins for the radius and eight bins for the orientations, generating a shape context of 24 elements.

The figure 4 we can see the image segmentation. This segmentation will be used as the area work. It to say, we going to analys only this section of whole image. The images presented in the figure 5 show the different features that we get from the object and that going to be useful for modelling our object and save it into our database.

An active sensor is very useful when a robotic system tries to recognize an object. J. Denzler and C. Brown (Denzler et al., 2001) proposed to select successive sensor configurations in order to discriminate the object in a learned 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 estimated state of recognition of $\Omega_k$ classes for $k \in 1,n$. We want to compute the true state given an observation $o_t$ at each acquisition. Optimal estimation is given by an action $a_t$ that optimizes mutual information. Mutual information can be defined as:

$$ I(x_t; a_t | o_t) = H(x_t) - H(x_t | o_t, a_t) $$

where $H(\cdot)$ denotes the entropy of a probability distribution. Considering

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

and

$$ I(x_t; a_t | o_t) = \int_{x_t} \int_{a_t} p(x_t) p(o_t | x_t, a_t) \log \frac{p(o_t | x_t, a_t)}{p(o_t | a_t)} dx_t dx_a $$

an optimal action $a^*_t$ that maximizes mutual information is given by

$$ a^*_t = \max_{a_t} I(x_t; a_t | o_t) $$

### 3.1 Learning

The learning step creates a data base, in which every object learnt through several views (typically, more than 50), is represented by a set of conditional probability density functions (PDFs).

The recognition process can activate at each step one over six actions corresponding to the six dofs of the arm; an action is parameterized by the angle value of the corresponding joint. The camera is mounted close to the end effector reference frame, with a known hand-eye transform (especially, a tilt angle of 22.5deg. As joint angles are discretized into a set of discrete values, each action $a_t$ is defined as:

$$ a_t = (q_{0t}, q_{1t}, q_{2t}, q_{3t}, q_{4t}, q_{5t}) $$

where $q_i$ and $n_i$ are respectively the angle value for the joint $i$, and the number of its discrete values. In others words it means a posible position of sensor at the sphere of Gauss.

Several features are learnt from every view of a learnt object. These features are: the chromatic intensity extracted from the normalized $(r, g, b)$ components, the edges, represented by the shape context and signature, and the interest points represented by the Harris points et the Sift.

One feature (color histograms, shape context, shape signature, Harris or sifts) could be sufficient for the object recognition, but the fusion of several appearance proprieties makes the system more robust.
So all features extracted from all views of all objects, constitutes the memory of the recognition system. Every feature is recorded as a gaussian PDF, so that a view is represented by seven PDFs; every PDF is characterized by a mean and a variance.

\[
I_p = \left( \frac{I_r(p)}{I_g(p)} \frac{I_b(p)}{I_g(p)} \right) = \frac{1}{n} \sum_{i=0}^{n-1} \left( \frac{I_r(i)}{I_g(i)} \frac{I_b(i)}{I_g(i)} \right)
\]

In Equation 5, \(I_p\) 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 some characteristic of the object when the camera has been set to a given configuration.

### 3.2 Recognition phase

This section presents the main objective of this work: the active object recognition.

An unknown object, supposed to be in a class \(k\) is presented to the system, which executes a sequential process to recognize the presented object. At the beginning, we assume equal probabilities \(P_k\) for all learnt class object.

Mutual information is computed as:

\[
I_0(\Omega, c | a) = \sum_{k=1}^{K} e_k(a) P_k
\]

where the entropy \(e_k\) is defined as follows:

\[
e_k(a) = \sum_{c_i} P(c_i | \Omega_k, a) \log \frac{P(c_i | \Omega_k, a)}{P(c_i | a)}
\]

In order to compute the mutual information we obtain the best matching between current estimate state and the observation made in this step. We look then the action value \(a_0^*\) that maximizes mutual information: \(a_0^* = \max_a I_0(\Omega, c | a)\)

The system executes this action and probabilities \(P_k\) are updated for every possible class, reenforcing probability of possibles ambiguous classes and in the other side, decreasing probabilities for non similar classes.

\[
P_k = \frac{P(c_0 | \Omega_k, a_0)}{P(c_0 | a_0)}
\]

This procedure iterate in a sequential way until that the probability of the most probable class exceeds a given certainty threshold.

### 4 Experiments and results

This section presents the evaluation of our method. We used 8 different objects, shown in Figure 3. For learning, a set of images for each object is acquired by moving the camera on the semisphere centered on every object, put down successively isolated on the table. Then some more images are acquired to evaluate our recognition approach. In Figure 6 we can see several images of the object that we use for make this test. These images are acquired from different view points with respect to the learning phase.

![Figure 6: Several images of the objet 1 in learning phase.](image)

Let us consider four cases for testing our algorithm. These cases are:

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

#### 4.1 Case 1: an isolated object.

In this case the system must recognize an object that it is on a table. The object has been learnt previously during the learning phase. 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 of the known object seen on images
presented in figure 7. We can observe on figure 7, the successive probabilities of having an object of a given class on the image. The system begins to think that he has three possible objects, but after some images, he decides to give more probability to object class 1. Finally he gives the response that an object of the class 1 is on the table.

Table 1: Matrix of confusion.

<table>
<thead>
<tr>
<th>Object</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>2</td>
<td>20</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
</tbody>
</table>

We have repeated this experiment about thirty times for every object class presented on figure 3 and we obtained the matrix of confusion presented on table 1. One can see very good recognition rates, especially if one takes into account the fact that we work with real objects.

4.2 Case 2: a partially occluded object.

In order to evaluate the behavior of our system face to a situation in which the system must recognize an object known but with occlusions, we have prepared the situation shown on figure 9. In these images an object from class 1 is partially hidden by a big letter A. Although only one part of the object is hidden, it could cause a failure for the system.

The result of the execution of recognition task is shown with the graphic in figure 10. It is possible to observe the doubt of the system since the noise introduced by the letter A, makes the system believe that he has another object different to object class 1 in the scene. But when advancing in the recognition process the system fortifies the hypothesis about the presence of an object of class 1 in the scene. At the end the system gives the good response that the object indeed belongs to the class 1.
4.3 Case 3: an unknown object.

When we present an unknown object to the system and that we execute the process of recognition, the system starts to extract features from the region something extracted on the image; he seeks in his database for an object that has characteristics similar than the ones extracted from the image. But if the object has not been previously learnt, he will return an unpredictable answer according to the features which have been extracted from the image and the ones that have been learnt on the known object classes. Then if we want that the system avoids errors on such a situation, we need to incorpore a new class: the objet class NULL.

Figure 11: Images for recognition of an unknown object.

It is thanks to the use of the object class NULL that the system is able to make the difference between an object located in its database and another which looks like to this one. In this case the system will treat the request of recognition like a research of an object of the class NULL and will tell us that there is an unknown object in the scene.

Figure 12: Evolution of the probabilities of a scene with 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 occlusions between objects depending on the camera position with respect of 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. Initially the system has learnt 8 object classes. 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 are devoted in order to confirm that an object of class 7, initially occluded, is also on the table.

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 on the current scene interpretation; failures could occur because our system is sensible to illumination changes. More generally, our system fails when the object appearances change in 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 in the decisional layer of the robot, so that a recovery action could be performed, like for example, 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. It will make our method more robust and efficient. These 3D features will allow to solve the problem of confusion between similar objects, because two objects can be similar by their appearance, but could be perceived as different considering more discriminant 3D characteristics.

REFERENCES


