

A Filter-based Approach to Robot Learning Concepts from Images

Nicolas Bredeche², Jean-Daniel Zucker¹, and Yann Chevaleyre¹

¹ LIP6-CNRS, Pole IA, University Pierre and Marie Curie
Paris, France

{[nicolas.bredeche](mailto:nicolas.bredeche@lip6.fr), [jean-daniel.zucker](mailto:jean-daniel.zucker@lip6.fr), [yann.chevaleyre](mailto:yann.chevaleyre@lip6.fr)}@lip6.fr
<http://www-poleia.lip6.fr/>

² LIMSI-CNRS, AMI, University Paris XI
Orsay, France
<http://www.limsi.fr/>

Abstract. To efficiently identify properties from its environment (be it the presence of a human, or a fire extinguisher or a paper on the floor) is an essential ability of a mobile robot who needs to interact with humans. Successful approaches to provide robots with such ability are based on machine learning that heavily rely on ad-hoc perceptual representation provided by AI designers. Our goal is to endow Pioneer 2DX autonomous mobile robots with a perceptual system that can efficiently adapt itself to the context so as to enable the learning task required to anchor symbols. Our approach is in the line of meta-learning algorithms, that iteratively change representations so as to discover one that is well fitted for the task. This architecture is based on a widely used approach in meta-learning: the Filter-model. Experiments show the interest of such an approach that dynamically abstracts a well fitted image description depending on the concept to learn.

1 Introduction: anchoring symbols, detecting and identifying objects

Recent works in both Robotics and Artificial Intelligence have shown a growing interest in providing mobile robots with the ability to interact and communicate with humans. One of the main challenges in designing such robots is to give them the ability to perceive the world in a way that is useful or understandable to us. One approach is to give the robot the ability to identify physical entities and relate them to perceptual symbols that are used by humans (to refer to these same physical entities). To perform this task, the robot has to ground these symbols to its percepts (i.e. its sensor data). Recently, the term "Anchoring" [1] has emerged to describe the *building and maintenance of the connection between sensor data and the symbols used by a robot for abstract cognition*. As a matter of fact, anchoring is an important issue for any situated robot performing abstract reasoning based on physically grounded symbols. Amongst other, anchoring plays an important role to communicate or relate to either other robots or humans.

There are tasks such as object manipulation or functional imitation where anchoring requires explicitly recognizing objects and localizing them in a three-dimensional space. Fortunately, such an *object recognition* is not always necessarily required to achieve a good anchoring. In applications such as human/object tracking, face and object identification or grounded robot-human communication, *object identification* is enough. Informally, to recognize an object often requires from the robot to both identify from its percepts what is an object and use a model of the object to localize it. This task has been studied for a few decades now and is known to be difficult in unknown environments [2]. On the contrary, identifying the presence of an object is clearly simpler. Moreover, there exists many easy to use and reliable descriptions for characterizing the presence of an object. To identify the presence of a fire in a room, one does not have necessarily to recognize it. Smelling smoke, hearing cracks, feeling heat, seeing dancing shapes on a wall are different ways of identifying the presence of a fire. For an autonomous robot, the ability to identify objects is a first step towards more complex tasks. *Object detection* (detecting a fire) may be built by regularly checking whether the object is identified. Identifying objects is therefore a simple form of anchoring symbols (such as "fire") to its percepts.

In this paper, we are concerned with a practical task where a Pioneer 2DX mobile robot has to rely on its limited vision sensors to anchor symbols such as "human being", "mobile robot" or "fire extinguisher" that it encounters while navigating in our laboratory. Anchoring is then used to support human/robot or robot/robot communication. For instance, an interaction may be engaged if a "human being" is identified or a rescue operation may be initialized if a non-responding "Pioneer 2DX" is identified. Identifying a "fire extinguisher" may allow the robot to respond to a query formulated by a human. To design an autonomous robot living in a changing environment such as our laboratory with identification ability described above is a difficult task to program. As such it is a good candidate for a Machine Learning approach which may be easily recasted as a classical *concept learning task*. To teach the robot to anchor symbols using Machine Learning approaches have proven successful [3]. To use machine learning techniques, the designer has to both define learning examples and a representation language based on the robot percepts to describe them.

It is clear that a great part of the success of the learning task per se depends on the representation chosen [4]. Having an AI designer providing the robots with an adequate representation has a major drawback that has been pinpointed by famous detractors of AI [5]: it is a fixed adhoc representation. Any change of setting (a museum instead of an AI lab) will require a new perceptual description.

Our main objective is to endow a robot with the ability to dynamically abstract from its percepts the representation that are the most suited for the learning task required to learn concepts (an identification or anchor symbols). The idea is that the robot explores the space of possible example descriptions (with various colors, resolution, representation formalisms etc.) so as to discover for each concept a well-fitted representation. The intuition being that for an-

choring the symbol "human being" a robot does not need a representation with a resolution that is necessary for identifying a "power-plug" on a wall.

In this paper, we describe the PLIC architecture that enables a robot to achieve efficient object detection by discovering for each object to identify a good trade-off between *expressiveness* and *complexity* of its perceptual system. This architecture may be seen as a combination of the two widely used approach in meta-learning: the Wrapper-model and the Filter-model [6].

2 Problem settings

The practical task we are concerned with takes place in a wider project called Microbes [7] whose goal is to have a colony of eight robots co-habit with AI researchers. We aim at providing each PIONEER2DX autonomous mobile robot with the ability to identify -but not recognize- objects encountered in its environment. Each robot navigates during the day and when resource are available it takes snapshots of its field of vision with its video camera. The snapshots are taken either randomly from time to time or upon a specific human request³. At the end of each day, the robot may report to a supervisor and "ask" her/him what objects (whose symbols belong to a pre-defined lexicon or not) are to be identified on a subset of taken pictures⁴. It then performs a learning task in order to create or update the connection between sensory data and symbols which is referred to as the anchoring process. Figure 1 describes this whole process. The learning task associated to the anchoring is therefore characterized by a set of images description and attached labels. It corresponds to a multi-class concept learning task.

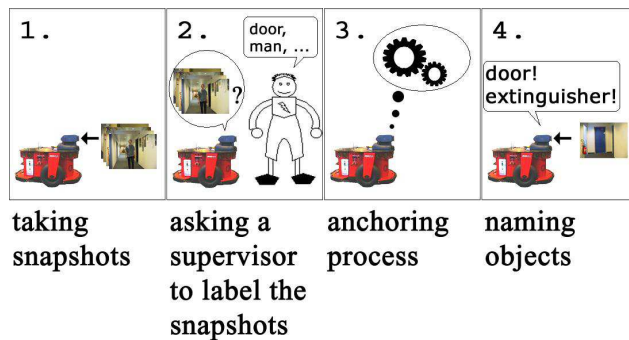


Fig. 1. The 4 steps toward lexicon anchoring

³ Thanks to *active learning* techniques, the robot may also take snapshots of scenes that appear to be interesting w.r.t. enhancing the detection accuracy of a known object (e.g. ambiguous images).

⁴ Again *active learning* techniques may be used by the robot to select the most *informative* images.

A key aspect of the problem lies in the definition of the learning examples (the images) used by the robot during the anchoring process. In effect, a first step in any anchoring process is to identify (preferably relevant) information out of raw sensory data in order to reduce the complexity of the learning task.

The PIONEER2DX mobile robot provides images thanks to its LCD video camera while navigating in the corridors of our lab. The images are 160*120 wide with a 24 bits color information per pixel. Humans, robots, doors, extinguishers, ashtrays and other possible targets can be seen among the images as shown in figure 2. All these possible targets as they appear in the images are in different shape, size, orientation and are sometimes partially occluded. Finally, each image is labeled with the names of the occurring targets.



Fig. 2. Two snapshots taken by the robot. Left: image labeled with a fire extinguisher and a door. Right: image labeled with a human

In this paper, our goal is to give the robot a perceptual system that can dynamically extract a relevant representation out of the camera snapshots in order to achieve the learning task. We define the role of the robot's perceptual system as to extract *abstract percepts* out of *low-level percepts* such as a set of pixels from the video camera or sonar values. These abstract percepts provide a representation of the world perceived on which will be based further computation and can be anything from sets of clustered colored regions to a matrix resulting from a Hough transform. The choice of a representation is motivated by finding a good trade-off that reduces the size of the search space and the expressiveness of the abstract percepts.

From the robot's point of view, each pixel from the camera is converted into a *low-level percept*. Hence, images from the camera are described as a set of low-level percepts. Each low-level percept is defined by 5 property attributes which are: x and y localization in the visual field, *hue*, *saturation* and *value*. In order to ease object detection, the learning task is limited to finding the most relevant set of attributes for low-level percepts (i.e. values for pixels) in order to be able to detect an object in an image by using only a single low-level percepts. This bias is very strong but can be considered enough in many case. Moreover, the matching complexity between detection rules and the image content is very low. This bias will be explained in section 4, along with the algorithm and learning task.

Table 1 shows the results from a learning task on 120 images at three different scales, where the goal is to identify a human by using only one pixel. The exper-

iment with images subsampled to 1x1 (i.e. a global color histogram) achieves an accuracy close to random choice (approx. 50%), which can be explained because the background is a strong distractor for a representation based on global image information. As expected, the experiment with 16x12 images performs better, but enhancing granularity do not lead to better results since the search space becomes quickly too big. As a matter of fact, the learning task with 160x120 images could not be achieved due to memory limitation. However, since identification is based on only one pixel, it is likely that the accuracy will decrease if the resolution is too high because pixels won't endow enough information anymore.

The previous results showed the importance of representation for detecting a target object, even with a strong bias. Moreover, the most fitted representation may vary from one target object to another, which raises our concern about how to choose an efficient representation for anchoring symbols.

3 Related works

In robotics, reliable information is most of the time provided by simple stimuli since vision processing is difficult to handle and costs much of the processing time. Works on anchoring a lexicon are usually focused on higher level cognition and adaptation, such as language games [8], and thus do not deal with complex scenes. However, when it comes to applications with complex vision processing needs, works in robotics share much with works from other communities such as object detection and tracking or image indexing where there is an impressive body of literature.

On the one side, tracking applications are mostly embedded in systems using a video camera, with or without stereo vision. A known approach to tracking is multiple hypothesis tracking, where temporal information is used to choose between hypothesis. A more recent approach relies on a detection method where tracking is the process that checks the spatial coherence through time of a single hypothesis given by a detection algorithm [9]. Both approaches usually relies on model-based detection and use image transforms to detect shapes along with correlation to cope with the occlusion problem. Some tracking applications in robotics also implement such approaches and make use of the robot capability to follow the target.

On the other side, image indexing is about classifying an image based on its content, without precisely identifying the location of the target concept in

scale	accuracy
1x1	52.42%
16x12	62.88%
160x120	<i>n/a</i>

Table 1. Identifying a human at 3 different scaling.

the image. Approaches to image indexing include model-based classification, images description using Fourier transforms, etc. A popular approach is based on matching sets of connected color regions between images [10]. The goal is to find instance of a given spatial configuration between regions extracted from the images (i.e. using a region growing algorithm). The main drawbacks of this approach are the imprecision of region growing techniques and the cost of the matching phase between undirected planar graphs representing sets of connected regions. However, good classification results were also achieved by simply comparing the global color histogram of each image [11]. In mobile robots, image classification into categories is used for creating landmarks or for navigating. In this case, image indexing using global color histograms is particularly well fitted because it classifies quickly the whole image.

In this paper, the object detection task shares with image retrieval the problem that anchoring takes place while it is difficult to identify the target due to the richness of the environment. In this case, anchoring relies on identifying hints in the images that prove the occurrence of an object instead of extrapolating and matching a pattern of the whole target object. Thus, object detection can be referred to as *weak anchoring*.

In the previously described domains, few works are concerned with adaptive issues, which is a main concern in robotics since [5]. Extending Marr's visual object recognition paradigm [12] where recognition is a bottom-up process, perceptual adaptation is considered as an important component for any perception-based cognitive activities and initiated many works concerning perceptual learning [13], active perception, perceptual attention, etc. However these concerns are of primary importance in the field of psychology of perception and could provide an interesting framework for artificial perception, we do not know of any application in anchoring for a mobile robot.

4 A meta-learning filter-based approach

As mentioned in section 2, in our context, object recognition is not required (as it would be needed for an application such as tracking) and object identification is sufficient for achieving the anchoring we are concerned with. The problem is to find a good representation in order to perform an efficient identification. As a matter of fact, representing the snapshots taken by the robot video camera as a set of pixels is far too complex for this task. Thus, the goal is to define a representation such as visual information is described in a way that reveal the object to be detected. As a consequence, the robot divides images into parts where relevant data about the target object are likely to appear. To sum it up, our approach is based on a simple property: to detect an *object* O within an image I , it is enough to detect O in some *part* of I . Here, the word *object* refers either to an inanimate object (door, window, ashtray...) or a dynamic one (another robot, a human...). Thus, this definition of detection can be termed *occurrence detection* or *verification*. One of its main interest is to limit the complexity of matching between images to a single part-to-part match as seen in section 2.

This definition relies on two important aspects for dividing the images into parts : The choice of **granularity**, which represents a trade-off between complexity and expressiveness and the choice of a part’s **structural configuration** in order to grab specific structural properties of the target objects.

First, this definition raises the question of how to define the granularity of the parts to be extracted from the video camera. The set of parts that describes a given image is considered to be the abstract representation on which will be based further processing. As a matter of fact, this new representation is abstracted from the initial representation provided by the video camera, that is a matrix of colored pixels. Pixels can be referred to as *low-level percepts* as seen previously, and parts can be referred to as *abstract percepts*, since they will be used as percepts for further processing.

Figure 3 synthesizes our approach. First of all, a low-level perception is given by the video camera as a matrix of pixels. Then, an abstraction operator called *association* [4] related to granularity is applied to cluster sets of contiguous pixel, in order to get *abstract percepts*. The robot uses sub-sampling as the association operator, but a region growing algorithm has also been tested. At last, an abstraction operator called *aggregation* [4] related to structural configuration creates a set of *structured abstract percepts*, each embedding one to many contiguous abstract percepts according to a given structural configuration. For clarity, structured abstract percepts will be referred to as *s-percepts*⁵ in the following. The previously mentioned figure shows an example where an abstract percept is defined by its color histogram and localization. In this example, each s-percept uses a L-like structural configuration to embed three abstract percepts. In the scope of this article, we limit our investigation to granularity, that is the association operator.

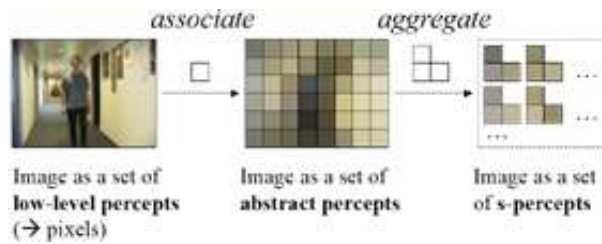


Fig. 3. Abstraction operators used to provide a description as a set of s-percepts

⁵ s-percept, as in *structural percept*

5 Experiments On Anchoring

5.1 Experimental Setup

In order to validate our approach and to train the perceptual system, the PIONEER2DX mobile robot acquired a learning set of 470 images thanks to its LCD video camera in the corridors of our lab. All these possible targets as they appear in the images are in different shape, size, orientation and are sometimes partially occluded. Finally, each image is labeled with the names of the occurring targets by a supervisor as explained in section 2.

The multiple instances rule learner RIPPERMI[14] developed in our lab was used on the descriptions obtained from these images with a ten-fold cross validation. Moreover, each experiment is repeated 10 times in order to get a good approximation of the results. RIPPERMI returns a set of rules that covers the positive examples.

Each abstract percept is defined by 8 property attributes which are: x and y localization in the visual field, *hue*, *saturation*, *value*, and the *standard deviation* for the three previous properties. Standard deviations are computed from the original $160 * 120$. In order to reformulate the abstract percepts, we developed the PLIC system (PLIC stands for Perceptual Learning by Iterative Construction), which is both a structural reformulation tool as well as a wrapper that explores the possible definition for abstract percepts according to given search hypotheses. Given a set of building rules, PLIC provides a description of the video camera information into abstract percepts or s-percepts. In this case, PLIC acts according to the filter model and was used to generate all the descriptions for each granularity level.

5.2 Experiments on Granularity

Many good results in object recognition or image classification have been achieved by using low resolution images. In order to perform experiments with varying granularity, PLIC generates 6 selected sets with fixed number of abstract percepts ranging from a single percept per image (i.e. $1 * 1$) up to a 3072 abstract percepts per image (i.e. $64 * 48$). An abstract percept correspond to a rectangle region of a given size. This size is fixed for each experiment and abstract percepts do not overlap : this can be seen as some kind of sub-sampling, where an abstract percepts correspond to a pixel of the sub-sampled image. The goal is to learn both the extinguisher and the human class.

Figure 4 shows the results from the experiments. The human and extinguisher detection accuracies from the global histogram description (i.e. $1 * 1$ grid) respectively achieves 58.2%⁶ and 55.04%. This shows that the global histogram-based approach used in indexation is somewhat fitted for a detection task, while quite

⁶ this accuracy differs from the one given in section 2, because abstract percepts includes new attributes about standard deviation of hue, saturation and value as explained earlier

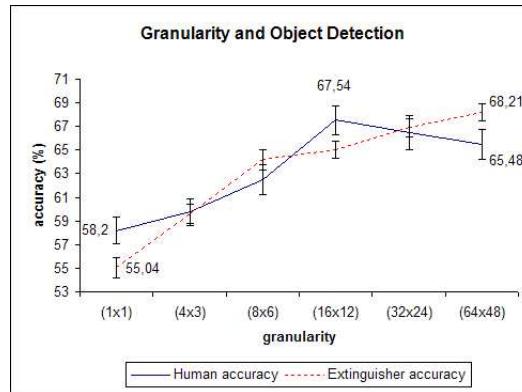


Fig. 4. Experiments with 6 different granularities for the 2 targets

weak in this case. Hence, a global histogram is able to capture some of the relevant information for detection (for example, a high amount of red is sometimes enough to detect an extinguisher). Moreover, the human detection accuracy increase up to 67.54% if the granularity level is enhanced to $16 * 12$. This is due to a better approximation of relevant information by the grid elements (i.e. the abstract pixels). Moreover, the extinguisher detection accuracy keeps on rising. However, enhancing granularity should not improve the accuracy forever, as shown for the $32 * 24$ and $64 * 48$ grids where the human detection accuracy drops successively to 66.45% and 65.48%. As a matter of fact, since detection is based on a single abstract percept, accuracy will deteriorate once the instance covers too few information (e.g. the robot won't be able to detect anything from a single pixel) or once the search space is too big (i.e. overfitting). Experiments with higher level of granularity are not shown here because of the memory size needed to perform such a learning task. Note however that at some point, accuracy should tend to drop to 50%, that is random prediction, when the granularity is too high.

6 Conclusion

In this paper, we address the problem of anchoring a lexicon of object names in a real-world autonomous mobile robot using a video camera. To achieve this task, the robot is meant to choose a relevant representation out of low-level percepts in the context of multi-class learning.

The goal is to create high-level abstract percepts in order to make object detection faster and more reliable for the robot. In order to do this, we analyzed the impact of choosing a specific granularity for dividing the images provided by the video camera into parts. A filter-based approach is used to apply and evaluate several possible *association* operators [4] to the snapshots.

We have shown that our approach to build vision-based percepts improves the accuracy of anchoring object names into a robot's perceptions. We presented and used the PLIC system that acts as a filter and reformulates raw data into high-level percepts in order to evaluate descriptions at several levels of granularity. Experiments done with the generated percepts yielded good results and evaluation issues were considered.

The search for an appropriate resolution level may indeed be described as a kind of filter-like component. At a given resolution, accuracy could also be enhanced with a search for the adequate part's structural configuration.

This work finds its application in a real-world environment within the MICROBES multi-robots project [7] where it provides a basis for the use of shared symbols between the robot and its human interlocutors.

References

1. Coradeschi, S., Saffiotti, A.: Anchoring symbols to sensor data: preliminary report. In: Proceedings of AAAI-2000, Austin, Texas (July 2000)
2. Stone, J.: Computer vision: What is the object? In: Prospects for AI, Proc. Artificial Intelligence and Simulation of Behaviour, Birmingham, England., IOS Press, Amsterdam. pages 199–208 (1993)
3. Klingspor, V., Morik, K., Rieger, A.D.: Learning concepts from sensor data of a mobile robot. *Machine Learning* **23** (1996) 305–332
4. Saitta, L., Zucker, J.D.: A model of abstraction in visual perception. In: *Applied Artificial Intelligence*. 15(8): 761-776. (2001)
5. Brooks, R.: Intelligence without representation. *Artificial Intelligence* **47** (1991) 139–159
6. Kohavi, R., John, G.: The wrapper approach. In: *Feature Selection for Knowledge Discovery and Data Mining*, H. Liu and H. Motoda (eds.), Kluwer Academic Publishers, pp33-50. (1998)
7. Picault, S., Drogoul, A.: The microbes project, an experimental approach towards open collective robotics. In: *Proc. of the 5th International Symposium on Distributed Autonomous Robotic Systems*, Springer-Verlag Tokyo Inc. (2000)
8. Steels, L.: The origin of syntax in visually grounded robotic agents. In: *Proceedings of IJCAI97*, Morgan Kaufman Pub. Los Angeles. (1997)
9. Beymer, D., Konolige, K.: Real-time tracking of multiple people using continuous detection. In: *Proceedings of the International Conference on Computer Vision (ICCV'99)*. (1999)
10. Hsieh, I., Fan, K.: Color image retrieval using shape and spatial properties. In: *ICPR00, Vol.I*: pp 1023-1026. (2000)
11. Stricker, M., Swain, M.: The capacity and the sensitivity of color histogram indexing. In: *Technical Report 94-05*, Communications Technology Lab, ETH-Zentrum. (1994)
12. Marr, D.: *Vision*. Freeman and Co., Oxford (1982)
13. Goldstone, R.L.: Perceptual learning. In: *Annual Reviews of Psychology*. 49:585-612. (1998)
14. Chevaleyre, Y., Zucker, J.D.: A framework for learning rules from multiple instance data. In: *Proc. European Conference on Machine Learning (ECML2001)*. (ECML2001)