Contextual priming for artificial visual perception

Hervé Guillaume[1,] Nathalie Denquive[1], Philippe Tarroux[1,2]

[1]LIMSI-CNRS BP 133 – F-91403 Orsay cedex France
[2]ENS 45 rue d'Ulm F-75230 Paris cedex 05 France

Abstract – The construction of robotics autonomous systems able to identify objects in their environments requires the elaboration of efficient visual object recognition algorithms. Our knowledge of the mechanisms of natural perception suggests that, when the recognition process fails due to the degradation of the observation conditions and to the blurring of the intrinsic attributes of the objects, the information concerning the context is used by human for object recognition priming. In this case the indices used for object identification can be greatly simplified. We present in this paper an attempt to precise how such a principle can be applied to autonomous robotics. We show that using a compact frequency coding of the scene together with an unsupervised SOM learning we obtain syntactic categories that exhibit specific relationships with object categories. Thus, the construction of these syntactic categories should be useful for estimating the occurrence probability of object categories during the exploration of the perceptual space of a robotic system.

1. Introduction

The development of autonomous robotics systems requires the elaboration of efficient visual object recognition algorithms. However, recognition often fails when poor observation conditions alter the identification of the intrinsic attributes of objects. Some works from natural visual perception [1] show an early use of contextual information, thus allowing contextual priming of objects. In this case, an object is not defined by a collection of intrinsic attributes but rather by a limited set of deictic attributes [2] related to the considered context. Our middle term aim is to precise how this principle can be applied to autonomous robotics visual systems.

A first step toward an implementation of contextual priming is the definition of a context representation independent from the objects present in the scene. Torralba [3] has recently proposed to use the spatial frequency characteristics of the visual scene for estimating the probability of occurrence of various object categories. In this approach a context vector the components of which is the mean energy of the scene in a range of spatial frequencies and spatial orientation is computed and used to estimate the probability density function (PDF) $p(o|V_C)$ describing the relationship between the context and the presence of a given object in this context.

However, the proposed method needs to compute a mixture of Gaussians PDF and is based on a Bayesian approach lying on supervised procedures. We propose here a method based on a first step of unsupervised clustering in order to exploit the structure of the data input space.

Until now, the probabilities of occurrence of objects were obtained in a supervised way from the direct statistical linkage between the frequential properties of the scene and the objects composing it. In a robotics framework this approach must be reconsidered for two reasons. First, like natural systems which are the product of an evolution, the sensors of a robot must be adapted to their purpose. With supervised learning, the actual relevance of the coding is not really considered. One just tries to adapt the parameters of the system to obtain the right classification from an a priori given coding. Second, one can think that a way to lighten the computational load is to assign the visual scenes to a small number of contextual categories from their holistic characterization. The so defined contexts must be significant, in the sense that priming requires that they exhibit specific relationships with different object categories.

To define these contextual categories we propose a factorization of the conditional probability of object occurrence according to the contextual vector through the introduction of an intermediate level of clustering. We thus obtain the relation:

$$p(o|V_C) = \sum_i p(o|K_i,V_C)p(K_i|V_C)$$

An unsupervised learning step with a self-organizing map [4] allows an implicit computation of the probability $p(K_i|V_c)$ that an example belongs to a given cluster. Clusters can be viewed as intermediate categories based on the syntactic properties of the image.

In a second step, one computes the probability $p(o|K_i)$ of occurrence of objects associated with the different clusters according to the approximation $p(o|K_i,V_c) \approx p(o|K_i)$.

This approach is an answer to three questions: how to determine the kind of clusters that emerges spontaneously from the used coding, what are the contextual semantic categories related to these clusters, how to encode the specific relationships between these clusters and the object categories?

3. Experiments and results

3.1. Coding choice

Context can be represented by some holistic visual properties of the scene. The statistics of the structural elements of the scene resulting from different orientation configurations and different textures seems to allow us to discriminate contexts [5, 6].

A multi-scale visual filtering of the images was thus carried out [7] : i) the size of images was standardized by convolving them with a Gaussian filter ; ii) A Gaussian pyramid [8] was then used: an appropriate Gaussian filter was applied followed by a reduction by a factor of 2 of the image size at each step. We thus obtained a representation of the scene at 5 different scales, corresponding to 5 spatial frequencies (1/2, 1/4, 1/8, 1/16, 1/32 cyc/pixel) iii) then a bank of Gabor filters according to 4 orientations (0; Π/4; Π/2; 3Π/4) was applied to the resulting images iv) eventually, a 20 dimension signature vector (V_c) of the average energies according to the 5 scales and 4 orientations was computed.

These vectors are used to train self-organizing maps of various size according to the conventional Kohonen method [4] with a Mexican hat lateral inhibition and a progressive decay algorithm. The activations of the units were computed according to the Euclidian distance between the input vector and the unit vectors.

3.2. Results.

3.2.1. Relationships between clusters and contextual categories

The obtained clusters exhibited properties reflecting their spatial frequency characteristics. The main criteria that characterize the location of the images on the map were linked to the degree of openness/closing, to the complexity, and to the preferred orientations of the image (Figure 1).

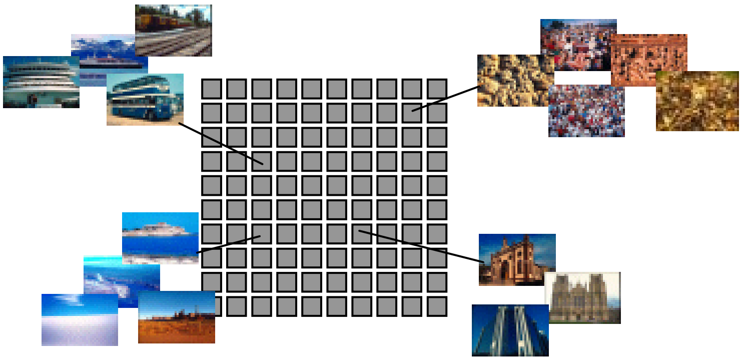


Fig. 1. A simplified representation of a map of 10x10 units showing examples of characteristic associated images. Clustering is carried out using the 20-dimension signature vector based on spatial frequency coding, with a whole of 2496 images drawn from the Corel database.

We analyzed the relationships between the obtained clusters and two contextual categories, Nature and Buildings. Several self-organizing maps, ranging from 4 to 25 units, were used for this purpose. The results obtained with the different maps being similar, we only show here those obtained with a map of 9 units.

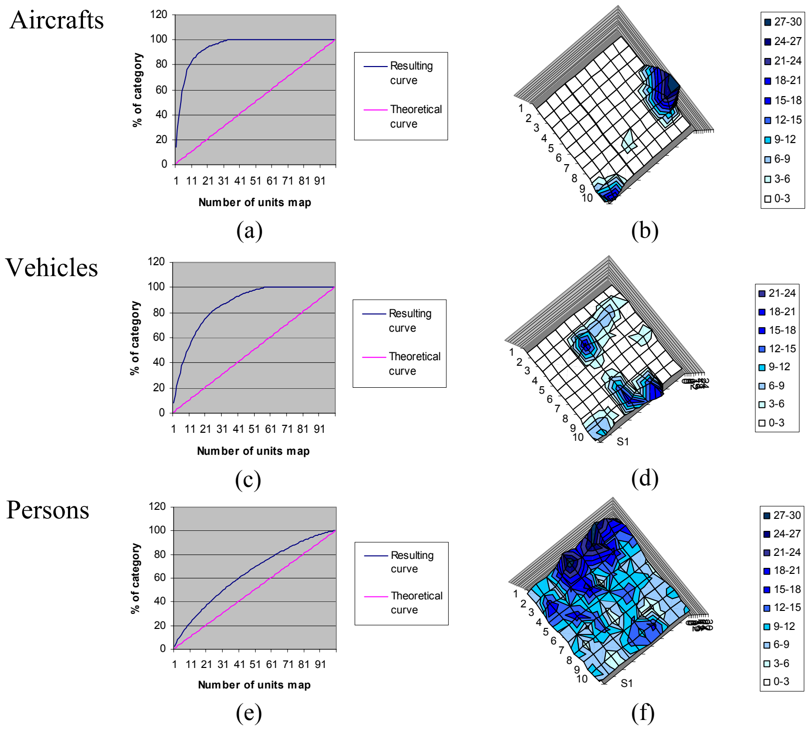
	Cluster1	Cluster 2
Construction	0.79	0.28
Nature	0.21	0.72

Table 1. Percentage of pictures from a category by cluster. The learning database consists in pictures of context (634 pictures : 320 from Nature and 314 from Construction). Note that the used contextual categories are exclusive: a visual scene cannot belong both to Nature and Construction.

We observed the formation of two clusters on the map and a strong overlap between clusters and semantic categories (Table 1). However, the sub-clusters corresponding to the units seem not match with any semantic sub-categories.

3.2.3. Relationships between Clusters and object categories

A second category of linkage is represented by the relationship between the observed clusters and the objects that can be identified in the scene. In order to study this point we indexed our database for the presence of three object classes: aircraft (211 pictures), vehicles (278 pictures) and persons (1081 pictures), over 2496 pictures from the Corel database.



(a) (b)

(c) (d)

(e) (f)

Fig. 2. (b), (d) et (f) : Self-organizing map of 10x10 units showing the dispersion of the object categories over the units. (a), (c) et (e) : Cumulated histograms of the number of examples of each category associated with the map units. The theoretical curve stands for the case of a uniform distribution of the examples over the map (1% by units)

Clustering was done with a map of 10x10 units. Figure 2 shows the dispersion of the object class examples on the map. We used entropy as a measurement of the non uniformity of the distribution of the examples. More the examples are grouped on few units, lower is the entropy of the distribution. Aircrafts are associated with units mainly located in two regions of the map (entropy : 4,05 ; more than 75 % of the pictures are grouped on 8 units). Vehicles are slightly more dispersed (entropy : 5,21 ; 75 % of the pictures are grouped on 21 units). Dispersion of the person class is much more significant (entropy : 6,48 ; 75 % of the pictures are grouped on 58 units). All units have at least one associated picture including a person.

4. Discussion and Conclusion

The joint use of a compact coding and a self-organization algorithm allows us to gather the visual scenes according to intrinsic characteristics mainly based on their statistical properties. Oliva and Torralba have observed a similar result [9] when human subjects are asked to group visual scenes according to the scene intrinsic properties no matter their semantic contents. This result is in support of the adequacy of the coding, since we show here that an unsupervised approach enables to reach the same kinds of clusters without calling upon subjective categories.

In addition, one observes an overlap between the obtained clusters and the Nature and Construction semantic categories. That reveals, on one hand, that these categories, of semantic nature, have different syntactic properties at low level and, on the other hand, that the used frequency coding captures the information which makes it possible to discriminate them at least to some extent.

The studied categories of objects are more or less scattered on the map. This result can be explained by the "degree of freedom" inherent to each category of objects. The aircrafts and the vehicles appear either on the road, either in the sky or on an airfield. People have a degree of freedom much larger and thus appear in a larger range of contexts. This assumption is also supported by the relative over-representation of this category compared to the others (aircraft: 211 pictures; vehicle: 278 pictures; person: 1081 pictures).

Objects belonging to different context are grouped on different units of the map. Therefore, priming should be done according to the clusters[1] and uses the probability matrices computed for each object category.

Practically, for an unknown example, the winning unit is first determined. Then, for a given object category the occurrence probability is obtained from the suitable matrix. However, this probability reflects the properties of the training set. In order to improve generalization, a PDF must be estimated from the learning data.

This PDF can be obtained through the association with each unit of a normal density distribution. This PDF is thus viewed as a mixture of Gaussians introduced after the training phase of the SOM using the RKDE [10] method or during learning with a probabilistic version of the Kohonen algorithm [11]

We see here the advantage of determining intermediate contextual categories. The direct computation of $p(o|V_c)$ requires the introduction of a mixture of Gaussian for each object category. Here the PDF is computed once on the training data. For each object category, one has to compute a simple matrix the size of which is the size of the SOM.

[1] These clusters are syntactic contextual categories whose definition is extensional (determined by the examples associated with the units of the map).

5. Bibliography

[1] Chun, M.M. and Y. Jiang, Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, **36** (1): 28-71, 1998.

[2] Ballard, D.H., et al., Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, **20** (4): 723, 1997.

[3] Torralba, A., Contextual priming for object detection. *International Journal of Computer Vision*,(53): 2, 2003.

[4] Kohonen, T., *Self-organization and associative memory*. 3 edition ed, Berlin: Springer-Verlag, 1989.

[5] Guérin-Dugué, A. and A. Oliva, Classification of scene photographs from local orientations features. *Pattern Recognition Letters*, **21**: 1135-1140, 2000.

[6] Vailaya, A., et al., Bayesian framework for hierarchical semantic classification of vacation images. *IEEE Transactions on Image Processing*, **10** (1): 117-130, 2001.

[7] Denquive, N. and P. Tarroux. Multi-resolution codes for scene categorization. in M. Verleysen, Ed. *European Symposium on Artificial Neural Networks ESANN 2002*, pp 281-287, April 23-26, Bruges, Be, 2002.

[8] Burt, P.J., Fast filter transform for image processing. *Computer Graphics and Image Processing*, **16**: 20-51, 1981.

[9] Oliva, A. and A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, **42** (3): 145-175, 2001.

[10] Hämäläinen, A., *Self-Organizing Map and Reduced Kernel Density Estimation*, University of Helsinki 1995.

[11] Gaul, W., O. Opitz, and M. Schader, *Data Analysis Scientific Modeling and Practical Application*: Springer, Berlin, 2000.