ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 281-287

Multi-resolution codes for scene categorization

Nathalie Denquive[1], Philippe Tarroux[1,2]
[1]LIMSI-CNRS BP 133 - 91403 Orsay cedex France
[2]ENS 45 rue d'Ulm 75230 Paris cedex 05

Abstract – The development of fast and reliable image classification algorithms is mandatory for modern image applications involving large databases. Biological systems seem to have the ability to categorize complex scenes in an accurate and very fast way. Our aim is to develop an architecture that leads to similar performances in computer vision. In this work, we present a coding method based on some principles inspired from biology that achieves a fast classification of complex visual scenes. A signature vector is extracted from the visual scene by a multi-scale filtering obtained through a bank of Gabor filters. These vectors constitute the inputs of a radial basis function network. The first connection layer implements a recoding of the filter outputs. The second one achieves a linear separation of the classes in the space of coding. We showed that an incremental approach in which each class is learned separately outperforms a more global one in which we tried to learn all classes together. According to the considered image category, the subset of features leading to the best result could be different, suggesting the use of feature vectors adapted to each image category. However, one of the major results of our study is that the signature vector we used, albeit very simple to compute, contains enough information to allow a correct image classification.

1. Introduction

The development of image applications (video, Internet) increases the need for processing methods allowing recognition and fast classification of complex scenes. Several recent works [1, 2] seem to show that humans and animals are able to reliably categorize complex scenes by using their low frequency content. We try to analyze here how such a coding can be used in the design of an automatic classification system for visual scenes.

By introducing into artificial systems some principles at work in biological visual systems, it appears possible to confer them some of the properties of the latter. The first processing step in the mammalian visual system extracts a set of low-level orientation features from the visual scene. This step is similar to the one achieved through a multi-scale wavelet filter. Field [3] has shown that such transforms lead to minimize the statistical dependencies between the components of the feature vector. Bell and Sejnowski [4], Olshausen and Field [5] showed that this mechanism is similar to an independent components analysis (ICA).

In vivo, this low level representation is built in the cortical layers V1 and V2. Although object recognition is achieved in high-level cortical areas (Infero-temporal cortex IT), fast scene recognition could be based on a bottom-up mechanism mediated by top-down attentional modulation of low-level filters. Thus an adequate combination of low-level features could support complex categorization abilities. This is this approach we wanted to illustrate in this work.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 281-287

2. Material and methods

The images used in this study belong to the Corel image database. Gray-level images were extracted from this database and sorted into 14 categories. From this selection, various training and generalization sets have been built by random sampling. Images initially standardized to a size of 374x251 pixels were scaled down to 256x128 pixels by Gaussian filtering followed by an appropriate size reduction.

2.1. Initial filtering

To bypass the requirement for a complex process of object identification before the identification of the visual scene, we searched for a low-dimensional coding space able to capture the semantic information contained in the visual scene. Thus, a bank of Gabor wavelet filters adjusted to four space orientations (horizontal, vertical, diagonals) and five spatial frequencies covering four octaves was first used to extract the frequency characteristics of the visual scene. Each initial image results in 20 filtered images. As stated above and according to Field [3], this filtering tends to maximize the statistical independence between the output images. This multi-scale processing was carried out using a Burt pyramid in a way similar to the method proposed by Guerin-Dugué [6] as described elsewhere [7].
The average energies of these 20 filtered images were then computed to constitute the components of a signature vector. This step allows to reduce the representation size of each image and to obtain a translation-invariant representation. It must be pointed out that the components of this vector are not independent anymore.

2.2. Training method

To perform the classification of the images coded in the preceding input space, we used a radial basis function network (RBF). The first stage of this network was used to adequately describe the projection of the data into the coding space. The output layer was then used to perform a linear classification of the scenes.
The input layer units were initialized to the components of the signature vectors (the low-level coding vectors). The weight vectors of the radial basis units were selected at random among the low-level input vectors. The thresholds were fixed at values of the same order of magnitude as the class variances.
An explicit linear classification of the images in the high level coding space was obtained by a supervised classification criterion: the connection matrix of this layer is the matrix which optimizes the desired outputs in a least-square sense. The method is based on the computation of the More-Penrose pseudo-inverse as described in Haykin [8].

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 281-287

3. Results

3.1. Global approach

In this first approach, 12 classes were considered. 537 images formed the training set and 545 the test set. We tried first to categorize these scenes using a single network with as many output units as the class number. With 20 RBF centers (data no reprinted), the results were not satisfactory. The training score was low and consequently a correct generalization could not be achieved. Several examples were not classified at all yielding an overall rejection rate higher than 68%. The use of 200 RBF centers improved the training score (Values ranging from 38% to 69 %).

%	1	2	3	4	5	6	7	8	9	10	11	12	Reject
1	**50,7**	-	3,3	-	-	-	-	-	-	-	1,3	-	**44,7**
2	-	-	-	6,3	0,8	-	-	-	-	-	-	1,6	**91,3**
3	1,1	-	**23,3**	-	3,3	1,1	-	-	2,2	-	-	-	**68,9**
4	0,8	-	0,8	**28,8**	-	2,3	-	0,8	-	-	-	3,0	**63,6**
5	-	-	4,0	0,8	**11,1**	2,4	-	0,8	-	-	3,2	0,8	**77,0**
6	-	-	3,7	2,2	-	**0,7**	-	-	-	-	0,7	0,7	**91,9**
7	-	-	2,9	-	0,7	-	**9,4**	-	20,3	15,9	-	-	**50,7**
8	-	-	-	7,3	-	1,3	-	**13,3**	1,3	-	-	-	**76,7**
9	2,0	-	0,7	-	-	-	0,7	-	**45,3**	4,7	-	-	**46,7**
10	-	-	-	-	-	-	18,7	-	6,7	**42,0**	0,7	-	**32,0**
11	4,1	-	-	-	2,0	-	-	-	-	2,0	**34,0**	-	**57,8**
12	-	-	-	2,1	-	0,7	-	1,4	-	-	0,7	**24,8**	**70,2**

Table 1 Generalization scores for the 12 classes with 200 RBF centers. Class names :planes (1), dishes (2), Utah (3), minerals (4), dogs (5), fishes (6), glasses (7), butterflies (8), porcelains(9), figurines (10), cars (11), flowers(12).

Generalization (Table 1) was improved accordingly but remained heterogeneous across the different classes. When the number of RBF was low, there was a strong sampling effect leading to a great variability of the results. In some cases, the confidence interval remained significantly high even with 200 RBF centers. This was the case for the two classes “Glasses” and “Figurines”. In these cases, the variability seemed to be inherent to the class.

3.2. Incremental approach

3.2.1. Preliminary experiments

In this second protocol, each class was learned separately against an equal number of counter-examples (called “Others”) which are taken at random in the other classes. The set size for examples and counterexamples was about 50 as well for training as for generalization. These sizes being relatively low, the experiments were repeated three times with three independent training and test sets for each case. Three classes were studied: “Planes”, “Porcelains” (which had good performances in the preceding approach) and “Dogs” (which had worse performances) (Figure 1). The used networks were initialized with 20 RBF centers.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 281-287

We observed first that, contrarily to the preceding results, none of the examples were rejected. We obtained satisfactory training scores. The variability of the recognition scores depended on the class. “Planes” and “Porcelains” reached the highest scores. “Dogs” still gave the worst results. In order to test whether these score resulted from the small size of the example set or from a greater heterogeneity of the class, the number of RBF centers was increased by a factor of two. The results were appreciably better (Figure 1 bottom right). Thus, we can conclude that the relative weakness of these last results was due to the heterogeneity of the “Dogs” class and not to a lack of information in the input coding-vector.

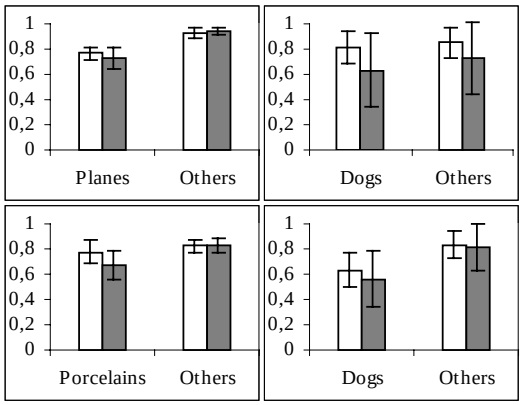

Figure 1 : Training (white) and generalization (gray) scores (one class vs counter-examples). 20 RBF centers: top left “Planes”; top right “Dogs”, bottom left “Porcelains”. 40 RBF centers: bottom right“Dogs”, “Dogs”.

The different samplings led to variable results. This variability could be due to the random initialization of the weight matrix of the RBF unit or to the inherent variability of the sampling method used with a too small number of examples. In order to test this hypothesis a final experiment with an increased number of examples in each class was carried out.

3.2.2. Final results

To check the robustness of the method in more realistic conditions (many more examples and counter-examples), we designed a new test set with 400 “Planes” and 800 counter-examples. The learning sets were the same as previously described. The obtained results were satisfactory (learning rate: “Planes” 0.81±0.14, counter-examples 0.79±0.26; recognition rate: “Planes” 0.84±0.16, counter-examples 0.62±0.07).

To check whether all the information contained in the coding vector was required to obtain correct classification results, we performed experiments with a reduced set of components extracted from this coding vector. In a first experiment, we considered each frequency separately with all the orientations (Table 2). Training results were satisfactory. However, the best ones were obtained with the high-frequency channels (f1 through f3). Test results were correct for f2 and f3.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 281-287

		f1	f2	f3	f4	f5
Training	**Planes**	0,94±0,06	0,90±0,02	0,90±0,06	0,86±0,02	0,88±0,06
	Others	0,81±0,08	0,84±0,12	0,86±0,05	0,78±0,05	0,79±0,05
Recognition	**Planes**	0,83±0,02	0,86±0,08	0,86±0,02	0,86±0,03	0,85±0,05
	Others	0,59±0,12	0,75±0,08	0,69±0,03	0,60±0,10	0,52±0,12

Table 2 Training and generalization results: one frequency, all orientations. "Planes" vs counter-examples with 20 RBF centers.

There was a significant degradation for f1 and f5. In a second experiment (Table 3), we considered each orientation separately with all the frequency channels. "Planes" were correctly learned, but the counter-examples learning rates were worse than in the previous experiments. An orientation component alone did not yield a correct recognition rate, except for the vertical one (Table 3 Column O2).

		O1	O2	O3	O4
Training	**Planes**	0,85±0,09	0,85±0,07	0,84±0,04	0,77±0,04
	Others	0,71±0,03	0,75±0,12	0,62±0,05	0,62±0,05
Recognition	**Planes**	0,69±0,04	0,81±0,04	0,86±0,04	0,76±0,07
	Others	0,41±0,03	0,71±0,04	0,45±0,10	0,60±0,08

Table 3 Training and generalization results: one orientation, all frequencies. "Planes" vs counter-examples with 20 RBF centers.

Up to this stage, all the images used consisted in focused objects (planes, animals…) and the features suitable for their recognition could be very specific of this kind of images. In order to study the importance of frequency or orientation components in the coding vector, we performed the same set of experiments with scenes rather than object images. We compared "Nature" scenes and "Building" scenes. The results showed that the full coding vector did not allow reaching the same training and test rates as those previously obtained (learning rate: "Nature" 0.68±0.28, "Buildings" 0.68±0.14; recognition rate: "Nature" 0.58±0.08, "Buildings" 0.58±0.20).
There were no significant differences between the two sets. However, when we use only a part of the coding vectors, there were significant differences with the results obtained in the "Planes" experiment (Tables 4,5). The frequency effect (Table 4, e.g. f1 versus f4) was much more pronounced for "Nature" than for "Building". There seemed to be no orientation effect (Table 5).

		f1	f2	f3	f4	f5
Training	**Nature**	0,74±0,00	0,77±0,02	0,74±0,12	0,71±0,13	0,62±0,14
	Building	0,74±0,08	0,73±0,06	0,67±0,16	0,64±0,07	0,63±0,16
Recognition	**Nature**	0,73±0,03	0,74±0,09	0,73±0,12	0,67±0,09	0,62±0,10
	Building	0,63±0,04	0,56±0,06	0,43±0,09	0,44±0,05	0,53±0,12

Table 4 Training and generalization results: one frequency, all orientations. "Nature" vs "Building" with 20 RBF centers.

		O1	O2	O3	O4
Training	**Nature**	0,68±0,04	0,72±0,07	0,73±0,02	0,65±0,05
	Building	0,75±0,06	0,71±0,08	0,78±0,04	0,83±0,06
Recognition	**Nature**	0,62±0,01	0,62±0,01	0,57±0,03	0,53±0,05
	Building	0,67±0,03	0,56±0,13	0,54±0,06	0,61±0,08

Table 5 Training and generalization results: one orientation, all frequencies. "Nature" vs "Building" with 20 RBF centers.

4. Discussion and conclusion

In spite of good training results, the global approach investigated first in this paper did not allow correct generalization scores. This could be due to the large number of RBF centers (200) required to reach a correct learning score. This increase in RBF number would probably lead to overfitting. However, the fact that such a complex set of very different images could be satisfactory learned is an indication that the coding vector introduced in this study reflects at least partially the information content suitable to characterize each image category.

The first experiments with the incremental approach suggest that the complexity of the recognition problem differs from class to class. For examples, "Planes" and "Porcelains" are correctly learned (more than 0.8) and recognized (about 0.7). On the contrary "Dogs" seem to be much more difficult (Training : 0.65; Test : 0.5). This could be due to a much greater heterogeneity of the data, which is obvious at visual inspection. This last category has a more pronounced semantic meaning that could be extremely difficult to relate to low-level frequency composition. The main result of this section is the demonstration that correct classification scores can be obtained even with a very small number of training examples (about 50). Further experiments should be done to generalize the approach to a larger number of object classes. With a large set of classes, it will be possible to use committee machines to determine which class a given image belongs to.

In a second step, we tried to analyze more precisely what kind of information is useful for classification. To investigate this question, we performed some experiments with a reduced set of features among the components of the coding vector. We observed that correct results could be achieved with only orientation information if we consider the high and mid-level frequency channels. The use of the full frequency bank according to one orientation usually gave poor results. This could be due to a higher correlation across the frequency channels of the same orientation. The results obtained with "Nature" and "Buildings" sets suggest that the optimal features could differ according to the scene semantics and that in the case of scenes characterized as a whole all the features contained in the coding vector are useful. These results suggest that feature selection techniques, which cannot be simply achieved with RBF networks, may be more appropriate. In this respect, the use of more sophisticated covariance indexes might improve the results. The use of auto-organization algorithms to project the RBF centers at appropriate locations within the data could also improve the results. However, more adaptive methods could be introduced. Our results suggest that each scene category could have its own optimal descriptors. Recent studies lead to the same conclusion [9] and propose to use ICA to

identify which sub-set of low-level filters is appropriate for the description of a given class of images.
Another important question raised by these studies is the fact that a given scene does not belong to a unique category. A scene showing a dog running on a park with is master can be categorized as “Animal”, “Dog”, “Person”, “Sport”, “Nature”, … depending on the context. To achieve such multi-categorical classification requires more complex descriptors that can be searched among complex combination of the low-level features described in the present study.

5. Acknowledgements

The authors would like to thank Antoine Cornuéjols, Jean-Sylvain Liénard and Michèle Sebag for their helpful comments during the course of this work.

6. Bibliography

1. Thorpe, S., D. Fize, and C. Marlot: Speed of processing in the human visual system. Nature, 381(6), 520-522 (1996).
2. Hérault, J., A. Oliva, and A. Guérin-Dugué. Scene categorization by curvilinear component analysis of low frequency spectra. in ESANN'97, Bruges (1997).
3. Field, D.J.: What is the goal of sensory coding? Neural Computation, 6, 559-601 (1994).
4. Bell, A.J. and T.J. Sejnowski: The "independent components" of natural scenes are edge filters. Vision Research, 37(23), 3327-3338 (1997).
5. Olshausen, B.A. and D.J. Field: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381(6583), 607-609 (1996).
6. Guérin-Dugué, A. and P.M. Palagi: Texture segmentation using pyramidal Gabor functions and self-organising feature maps. Neural Processing Letters, 1(1), 25-29 (1994).
7. Machrouh, Y., J.S. Lienard, and P. Tarroux. Multiscale feature extraction from visual environment in an active vision system. in International Workshop on Visual Form 4, Capri, It: Springer Verlag, Berlin (2001).
8. Haykin, S., Neural Networks. A comprehensive Foundation. Second Edition ed, Upper Saddle River, NJ: Prentice Hall (1999).
9. Le Borgne, H. and A. Guérin-Dugué. Propriétés des détecteurs corticaux extraits de scènes naturelles par analyse en composantes indépendantes. in Les Journées Valgo 2001, Larnas, France (2001).