A Cascade of Unsupervised and Supervised Neural Networks for Natural Image Classification

Julien Ros, Christophe Laurent, and Grégoire Lefebvre

France Télécom R&D - TECH/IRIS/CIM
4, rue du Clos Courtel
35512 Cesson Sévigné Cedex - France
{julien.ros, christophe2.laurent, gregoire.lefebvre}@francetelecom.com

Abstract. This paper presents an architecture well suited for natural image classification or visual object recognition applications. The image content is described by a distribution of local prototype features obtained by projecting local signatures on a self-organizing map. The local signatures describe singularities around interest points detected by a wavelet-based salient points detector. Finally, images are classified by using a multilayer perceptron receiving local prototypes distribution as input. This architecture obtains good results both in terms of global classification rates and computing times on different well known datasets.

1 Introduction

With the dramatic increase of available digital contents, advanced content management solutions become essential. If we focus on the particular situation of digital images (Infotrends[1] expects that the number of images captured on camera phones will reach 227 billion by 2009), efficient images management solutions such that supervised image classification have to be found.
The goal of a supervised image classification system is to group images into semantic categories giving thus the opportunity of fast and accurate image search. To achieve this goal, these applications should be able to group a wide variety of unlabelled images by using both the information provided by unlabelled query image as well as the learning databases containing different kind of images labelled by human observers.
In practice, a supervised image classification solution requires three main steps [1]: pre-processing, feature extraction and classification. Based on this architecture, many image classification systems have been proposed, each one distinguished from others by the method used to compute the image signature and/or the decision method used in the classification step. Regarding the signature computation, the most efficient methods are probably the local aproaches firstly introduced in [2]. In this case, local signatures are computed around some interest points and their values are chosen in a dictionnary obtained from the training

[1] http://www.infotrends-rgi.com/home/Press/itPress/2005/1.11.05.html

H. Sundaram et al. (Eds.): CIVR 2006, LNCS 4071, pp. 92–101, 2006.

database. Local signatures are used to represent the image by a distribution of local image features easily classifiable as in [3,4] or are directly used to learn a model used for the next recognition step [5,6,7].
In the state of the art, the dictionnary is classically computed thanks to a K-means algorithm [5] or by a bottom-up clustering procedure[8]. We propose here to use a self organizing map [9] to generate the visual dictionnary. Furthermore, in our approach, a Multilayer Perceptron classifier is built with the training dataset and is used for the last classification step.
The paper is organized as follows. Section 2 describes the method which was introduced earlier in [10] to detect interest points and extract local image features. Section 3 presents the self-organizing map algorithm, the construction of the vocabulary of local descriptors and the construction of the image feature vector. The design of the Multilayer Perceptron classifier and the decision rule are explained in detail in section 4. Experiments are presented in section 5 and finally, section 6 concludes the paper.

2 Local Features Extraction

The goal of feature extraction is to reduce the amount of data contained in an image by extracting relevant and discriminating features. In local approaches, this extraction phase results in feature vectors computed around interest points and an image I_j is thus represented by a set of local signatures $S(I_j) = \{s_{1j}, \ldots, s_{nj}\}$. It is important to mention here that local approaches result in a lack of ordering between signatures.

2.1 Interest Points Detection

The goal of interest point detectors is to find image locations that are perceptually relevant for the next recognition step. Many detectors have been proposed in the literature, each one focusing on a particular local property of the image content such as contrast [11], corners [12,13], edges [10,14], etc.

The salient points detector presented in [10] uses a wavelet analysis in order to find relevant pixels located on sharp region boundaries. The use of wavelet analysis is motivated by observing that multi-resolution, orientation and frequency analysis are of prime importance for the human visual system during the recognition step. This detector has proven its efficiency in many vision applications[10] and thus will be used in the present work.

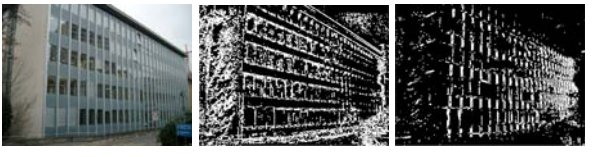
2.2 Description of Local Singularities

Most local descriptors describe the local neighborhood of salient points by characterizing edges in this area. Edge information thus appears fundamental in the process of local neighborhood description. To describe edges, gradient orientation and magnitude are generally used. Nevertheless, from a mathematical point of view, an edge or more generally a singularity can also be efficiently characterized by considering its Hölder exponents. We propose to use this mathematical notion to design our local descriptor.

Definition 1. *$f : [a, b] \to \mathbb{R}$ is Hölder $\alpha \geq 0$ at $x_0 \in \mathbb{R}$ if $\exists K > 0$, $\delta > 0$ and a polynom P of degree $m = \lfloor \alpha \rfloor$: $\forall x$, $x_0 - \delta \leq x \leq x_0 + \delta$, $|f(x) - P(x - x_0)| \leq K|x - x_0|^\alpha$.*

Definition 2. *The Hölder exponent $h_f(x_0)$ of f at x_0 is the superior bound value of all α. $h_f(x_0) = sup\{\alpha$, f is Hölder α at $x_0\}$.*

The local regularity of a function at a point x_0 is thus measured by the value $h_f(x_0)$. It is worth noting that the smaller $h_f(x_0)$, the more singular is the signal at the point considered. For example, the Hölder exponent of a Dirac impulse is -1 and 0 for a step function. For an image, the Hölder exponent is measured in the direction of the minimal regularity of the singularity (in the gradient direction). The different singularities met in an image are shown on figure 1. To



Original Image $-0.2 \leq h \leq 0.2$ $-1.2 \leq h \leq -0.8$

Fig. 1. Different Type of Singularities

describe an ROI associated to an interest point in an image I_j, both orientation and Hölder regularity of singularities contained in that ROI are characterized. For this purpose, orientation $\theta(x, y)$ and gradient magnitude $m(x, y)$ at each pixel location (x, y) of the ROI are first computed:

$$m(x,y)^2 = (I_j(x+1,y) - I_j(x-1,y))^2 + (I_j(x,y+1) - I_j(x,y-1))^2 \quad (1)$$

$$\theta(x,y) = tan^{-1}\left(\frac{I_j(x,y+1) - I_j(x,y-1)}{I_j(x+1,y) - I_j(x-1,y)}\right). \quad (2)$$

Then, for each singularity, the Hölder exponent h is estimated with foveal wavelets as presented in [15]. Orientations and Hölder exponents maps are then conjointly used to construct different 3D histograms. To build such histograms, each ROI is first partitionned into 4×4 blocks and each histogram is computed in a particular block before being normalized by the block size (See figure 2). This last step of the signature design is realized in the same spirit as the construction of the SIFT descriptor presented in [16]. Finally, the signature is obtained by concatenating the different 3D histograms and thus has a size of $n \times r \times o$ where n is the number of subregions (i.e. the number of interest points), r is the number of Hölder exponents bins into the range $[-1.5, 1.5]$ and o is the number of orientations bins into $[-\frac{\pi}{2}, \frac{\pi}{2}]$. We typically use 4 orientations, 16 subregions and 3 Hölder exponents bins resulting in a signature size of 192.

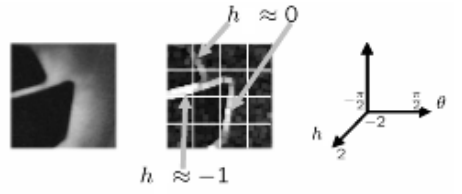


Fig. 2. Principle of the Singularity Descriptor

3 Image Representation

To use a classical machine learning methods for the last classification step, images must be represented by a vector of equal size. When local image descriptors have been extracted, a powerful and recent method is to represent the image by an histogram of local descriptors, this is the "bag of keypoints" representation introduced in [3]. Nevertheless, it supposes that local descriptors are quantized into a visual dictionnary of fixed size. We propose to build such a dictionnary by using a self-organizing-map.

3.1 Self Organizing Map Learning

The self-organizing map (SOM) is an unsupervised classification algorithm based on competitive learning[9]. It is a variant of the k-means algorithm that has the advantage of preserving the topology of input datas $X = \{x(t), t = 1, 2, \ldots\}$ with $x(t) \in D \subset \mathbb{R}^n$ and providing thus a better description of them.
The SOM aims at projecting the input data space D into a lower dimensionnal space (1D, 2D,...) defined by a regular discrete lattice L composed of N nodes. Therefore, it is a vector quantization algorithm which preserves the topology of the input space because each node c of the lattice is a neuron with a codebook vector $w_c \in \mathbb{R}^n$ such that if c_1 and c_2 are close then w_{c_1} and w_{c_2} are close in $\mathbb{R}^n$. For this purpose, the SOM is trained thanks to a competitive learning algorithm

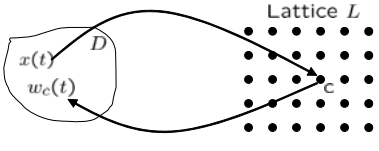


Fig. 3. General Principle of a Self Organizing Map

which supposes first that the SOM has been correctly initialized. For example, $w_i(0)$ could have been initialized randomly $\forall i = 1, 2, \ldots, N$. At epoch t of the learning step, $x(t)$ is compared simultaneously to all $w_i(t)$ by using a distance measure $d(x(t), w_i(t))$ on the input space D and the best candidate vector $w_c(t)$ associated to the node c (the best matching unit or BMU) is chosen such that:

$$w_c(t) = arg \min_i d(x(t),\ w_i(t)) i = 1, 2, \ldots, N. \quad (3)$$

The learning scheme uses then a kernel based rule to update the weights:

$$w_i(t+1) = w_i(t) + \alpha(t)h_{ci}(t)[x(t) - w_i(t)] \quad (4)$$

where $0 < \alpha(t) < 1$ is the monotically decreasing learning rate. Furthermore, h_{ci} denotes a neighborhood function that governs the strength of weight adaptation as well as the number of reference vectors to be updated (generally, a gaussian function is used). It is worth noting that a good choice for the number of iterations during the learning is 500 times the number of cells in the SOM.

3.2 Bag of Local Descriptors Representation

As previously emphasized, at this stage of the algorithm, an image I_j is described by a set of local signatures $S(I_j) = \{s_{1j}, \ldots, s_{nj}\}$ representing $3D$ histograms around interest points presented in section 2.2. Thus, this kind of representation could not be directly interpreted by a classifier because of the lack of ordering between signatures. Moreover, the number of signatures could be different for two images (due to different number of interest points detected). Thus, to build and use an image classifier, the image I_j should be represented by a feature vector $H(I_j) = [h_{1j}, \ldots, h_{Nj}]$. For this purpose, an indexing step should be used to transform the set of local signatures into a precise and compact representation of the image content.

This is a classical problem met in text categorization where a document composed of a set of words has to be characterized by a vector describing its content. For this purpose, a text is often represented by a vector of term weights, where the terms are chosen in the codebook (a set of meaningful words for the understanding of texts); this is the well known "Bag of Words" representation. This approach has influenced the work presented in [3] and denoted "Bag of Keypoints" which proposed to adapt text categorization methods to the computer vision problems.

Similarly, we propose to represent the image content by the probabilistic distribution H over local images features. This distribution is in fact the activation histogram of the SOM previously learned. For this purpose, each local signature of the image activates a particular cell (The BMU) and participates to an update of the histogram $H(I_j)$. The bins h_{lj} are defined as follow:

$$h_{lj} = card\{s_k \in I_j, ||s_k - w_l|| < ||s_k - w_i|| \forall l \neq j, k \in \{1, \ldots, n\}\}. \quad (5)$$

4 Neural Network Classification

At this stage of the algorithm, each image is represented by a unique feature vector denoted H. The natural image classification problem is thus reduced to a multi-class supervised classification problem. For this purpose, we have tested a multilayer perceptron (MLP), a Radial Basis Function network classifier (RBF) and a Support Vector Machine classifier (SVM). Experimentally MLP exhibited better results than the RBF and equivalent results than the SVM. Thus we restrict our discussion on this classifier.

MLPs can be used for classification problems and are multi layers feedforward neural networks fully connected. Thanks to their fundamental property of parcimonious approximation, they are well suited to modelize any continuous function $g : \mathbb{R}^N \rightarrow \mathbb{R}^p$, where N is the dimension of the input space and p is the number of classes. However, it supposes that sufficient neurons are chosen during the definition of the network architecture. A three layer perceptron architecture with N intputs, n_h hidden neurons and p output neurons is presented on figure 4.

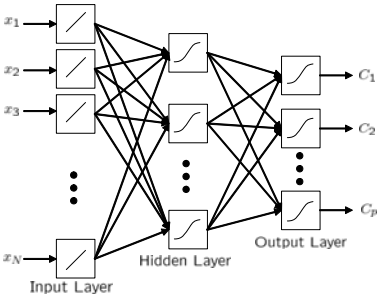


Fig. 4. General Principle of a Multilayer Perceptron

For an input data $X = [x_1, \ldots, x_N] \in \mathbb{R}^N$, the output of the k^{th} output neuron of this MLP is given by the discriminant function:

$$g_k(X, W) = \varphi\left(\sum_{j=1}^{n_h} w_{kj}\varphi\left(\sum_{i=1}^{N} w_{ji}x_i + w_{j0}\right) + w_{k0}\right) \forall k \in \{1, \ldots, p\} \qquad (6)$$

where $W = \{w_{ij}\}$ is the set of weights of the neural network considered and φ is the activation of the neurons of both the hidden and ouput layer. This one should be non-linear allowing the MLP networks to model nonlinear mappings well and is the standart sigmoidal function in the following.
MLPs are trained with the backpropagation algorithm which adapts the weights of the network in W to their optimal values for the given pairs $(X_l, t(X_l))\forall l \in \{1, \ldots, N_L\}$ in the training dataset. If the target vector $t(X_l)$ gives the correct class of X_l and is such that $t_k(X_l) = 1$ if $X_l \in C_k$ and 0 otherwise then the trained network approximates the correct a posteriori probabilities concerning the classification problem: $g_k(X, W) \approx P(C_k|X)$. An input data X can thus be easily classified by regarding the maximal value of the output neurons.

5 Experimental Results

In this section, the system scheme is tested on different well known datasets. The experiments particularly emphasize on the influence of the vocabulary size

(i.e. the size of the SOM) on the classification results. These results are analyzed by evaluating global classification rates. Furthermore, the analysis of ROC curves by computing the area under curve (AUC) for the best parameters will be presented providing a direct comparison to other algorithms. It is worth noting that the MLPs used in the experiments have a number of nodes in the hidden layer which is the mean of the number of nodes in the input and ouput layer $n_h = \frac{N+p}{2}$. It permits to achieve good classification results in reasonable computing times.

5.1 Presentation of the Datasets

The first dataset is extracted from the SIMPLICITY database[2]. It contains 500 images of size 384×256 and is divided into a learning dataset of 250 images and a test dataset of 250 images. There are five clusters: beaches, buildings, buses, elephants and flowers as shown on figure 5.

Fig. 5. Images from the SIMPLICITY Database

The Pascal database[3] is used to compare results with other major approaches of the state of the art presented during the PASCAL visual object classes challenge 2005. It is composed of four clusters (bikes, bicycles, persons and cars) as shown on figure 6. The test set contains 688 images whereas the training set contains 684 images. This base is representative of what a person has on his computer, they are of various sizes and have been shot from various viewpoints.

Fig. 6. Images from the PASCAL Database

5.2 Influence of the Self Organizing Map Size

The influence of the vocabulary size on the classification results is conjointly tested with the number of interest points extracted from the image. The SOM is rectangular and its size varies from 5×5 to 15×15 which constitutes a vocabulary from 25 to 225 prototypes. On figure 7, the global classification rates are shown.

[2] http://wang.ist.psu.edu/ jwang/test1.tar
[3] http://www.pascal-network.org/challenges/VOC/

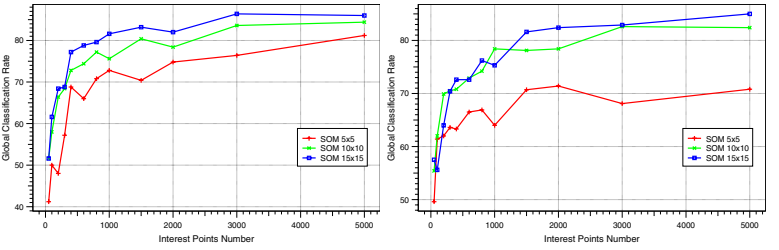


Fig. 7. Global Classification Rates for SIMPLICITY (a) and Pascal (b) datasets

For the two dataset used, the optimal SOM size is 15×15 which constitutes a small dictionnary of 250 prototypes. Nevertheless, the results between 15×15 and 10×10 are not so different compared from those obtained with a 5×5 lattice. Moreover, a study has shown that the quantization error does not decay very fast if the SOM is larger. It is thus not necessary to improve the dictionnary size.

For the two dataset, 3000 interest points are a good compromise in term of computing times and global classification rates. In this case, the method obtains a global classification rate of 86.4% for the SIMPLICITY dataset (86.4% for the SVM and 82% for the RBF) and 82.9% for the PASCAL database (81.7% for the SVM and 75% for the RBF).

5.3 ROC Curve Analysis and Confusion Matrix

On figure 8, AUC and confusion matrix are displayed for the parameters exhibited in the previous section. For the PASCAL dataset, the results are good and comparable to those obtained during the PASCAL 2005 recognition challenge.

Class	AUC
Beaches	0.9125
Buildings	0.9330
Buses	0.9959
Elephants	0.9844
Flowers	0.9971

Beach	Buildings	Buses	Elephants	Flowers	classified as
37	6	2	3	2	Beach
8	37	0	4	1	Buildings
2	1	47	0	0	Buses
4	1	0	45	0	Elephants
0	0	0	0	50	Flowers

Class	AUC
Bicycles	0.926
Cars	0.9622
Motorbikes	0.9793
People	0.8862

Bicycles	Cars	Motorbikes	People	classified as
71	23	14	6	Bicycles
7	252	5	11	Cars
7	3	202	4	Motorbikes
13	18	7	46	People

Fig. 8. AUC and Confusion Matrix for the SIMPLICITY and Pascal Dataset

5.4 Computing Times

The system must be efficient both in classification results and in computing times in order to be attractive for a human. Whereas features extraction, SOM and MLP learning are realized offline and could thus be long, the classification of a query image must be fast. The computing times obtained on a Pentium IV with 3Ghz are shown on figure 9. Moreover the features extraction for the entire training set takes $290s$ for the SIMPLICITY database and $1301s$ for the PASCAL database. The training steps (SOM and MLP learning) are not too long and so totally realistic in an offline mode. It is worth noting that SOM learning does not depend on the database size and the interest point number but only on the dimension of the SOM as emphasized in section 3.1 because the number of iterations is 500 times the number of cells in the SOM. The classification of a new instance is very fast and could thus be realized online in a professional application.

Dataset	SIMPLICITY			PASCAL		
Size of the SOM	5×5	10×10	15×15	5×5	10×10	15×15
SOM learning	4s	46s	185s	5s	46s	195s
MLP learning	12s	127s	665s	30s	413s	1706s
MLP classification	0.00012s	0.00064s	0.00244s	0.00012s	0.00073s	0.00232s

Fig. 9. Computing Times

6 Conclusion

This paper presents a neural network architecture for natural image classification using local images features. It has been shown that a self organizing map could learn a small visual dictionnary subsequently used to represent the image content by a distribution over the prototypes. Moreover, a classification step based on a multilayer perceptron has shown to be efficient. The approach exhibits high classification rates and small computing times. Its implementation in a professional application is thus possible. The perspectives are to use the Growing Hierarchical Self Organizing Map to generate the codebook [17] and to represent the image as a "bags of graphs" generated by grouping interest points which could be learned thanks to a SOM for structured datas [18].

References

1. Duda R.O, Hart P.E., Stork D.G.: Pattern Classification. 2^{nd} edition edn. John Wiley & Sons (2001)
2. Schmid C., Mohr R.: Local grayvalue invariants for image retrieval. IEEE Transaction on Pattern Analysis and Machine Intelligence **19**(5) (1997) 530–535
3. Csurka G., Bray C., Dance C., Fan L.: Visual categorization with bags of keypoints. In: The 8th European Conference on Computer Vision, Prague, Czech Republic (2004) 327–334

4. Jurie F., Triggs B.: Creating efficient codebooks for visual recognition. In: International Conference on Computer Vision, Beijing, China (2005) 604–610
5. Weber M., Welling M., Perona P.: Unsupervised learning of models for recognition. In: The 6th European Conference on Computer Vision, London, UK, Springer-Verlag (2000) 18–32
6. Fei-Fei L., Perona P.: A hierarchical bayesian model for learning natural scene categories. In: International Conference on Computer Vision and Pattern Recognition. Volume 2., San Diego, CA, USA (2005) 524–531
7. Marée R., Geurts P., Piater J., Wehenkel L.: Random subwindows for robust image classification. In: International Conference on Computer Vision and Pattern Recognition. Volume 1. (2005) 34–40
8. Agarwal S., Awan A., Roth D.: Learning to detect objects in images via a sparse, part-based representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(11) (2004) 1475–1490
9. Kohonen T.: Self-Organizing Maps. Springer-Verlag, Berlin, Heidelberg, New York (2001)
10. Laurent C., Laurent N., Maurizot M., Dorval T.: In depth analysis and evaluation of saliency-based color image indexing methods using wavelet salient features. Multimedia Tools and Application (2004)
11. Bres S., Jolion J.M.: Detection of interest points for image indexation. In: 3^{rd} International Conference on Visual Information Systems, Amsterdam, The Netherlands (1999) 427–434
12. Harris C., Stephens M.: A combined corner and edge detector. In: 4^{th} Alvey Vision Conference. (1988) 147–151
13. K. Mikolajczyk, Schmid C.: Scale and affine invariant interest point detectors. International Journal of Computer Vision **60**(1) (2004) 63–86
14. Loupias E., Sebe N., Bres S., Jolion J.M.: Wavelet-based salient points for image retrieval. In: IEEE International Conference on Image Processing, Vancouver, Canada (2000) 518–521
15. Mallat S.: Foveal Approximations for Singularities. Applied and Computational Harmonic Analysis **14**(2) (2003) 133–180
16. Lowe D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
17. Rauber A., Merkl D., Dittenbach M.: The growing hierarchical self-organizing maps: Exploratory analysis of high-dimensional data. IEEE Transactions on Neural Networks **13**(6) (2002) 1331–1341
18. Hagenbuchner M., Sperduti A.: A self-organizing map for adaptive processing of structured data. IEEE Transactions on Neural Networks **14**(3) (2003) 491–505