Supervised Image Classification by SOM Activity Map Comparison

Grégoire Lefebvre, Christophe Laurent, Julien Ros, Christophe Garcia

Supervised Image Classification by SOM Activity Map Comparison

Grégoire Lefebvre, Christophe Laurent, Julien Ros, Christophe Garcia
France Telecom R&D – TECH/IRIS/CIM
4, Rue du Clos Courtel
35512 Cesson Sévigné Cedex, France
{gregoire.lefebvre|christophe2.laurent|julien.ros|christophe.garcia}@francetelecom.com

Abstract

This article presents a method aiming at quantifying the visual similarity between two images. This kind of problem is recurrent in many applications such as object recognition, image classification, etc. In this paper, we propose to use self-organizing feature maps (SOM) to measure image similarity. To reach this goal, we feed local signatures associated to salient patches into the neural network. At the end of the learning step, each neural unit is tuned to a particular local signature prototype. During the recognition step, each image presented to the network generates a neural map that can be represented by an activity histogram. Image similarity is then computed by a quadratic distance between histograms. This scheme offers very promising results for image classification with a percentage of 84.47% *of correct classification rates.*

1 Introduction

In many computer vision applications such as multimedia data mining, pattern recognition, etc., evaluating the inter-image similarity is fundamental. Whereas human beings are able to compare immediately two images, by automatically extracting discriminative image features, this project in computer vision stays unsolved.

Measuring the similarity between two images is a very challenging problem. In fact, the image similarity measurement is tightly linked with the image content representation. Three approaches appear in the literature. First, the image description can be global with one or more representations describing the whole content, in a compact structure. Classical color histograms [9] are an example of such an approach. In this case, image similarity is generally computed by classical Minkowski metrics. The second issue is based on a preliminary segmentation step. A low level descriptor is affected to each region. Image similarity corresponds thus to an attributed graph matching problem [1]. The third way to represent image content is to extract several interest points (IP) and to consider related patches, called regions of interest (ROI) [10]. These salient points are considered as perceptually important and their neighborhood can be described by local descriptors. In this way, the whole image content is represented by a set of local regions. Due to the lack of order between detected salient areas, the similarity is determined by registration-based methods.

Based on some psycho-visual experiments [6], our approach focuses on the last technique. Indeed, human vision system executes saccadic eye movements between salient locations to capture image content. Likewise, Tversky studies [12] showed that when we compare two images, we detect common and distinct concepts between these regions. Our method tries to reproduce this extraction and distinction concept with a codebook learning strategy based on SOM algorithm [7] and an activity histogram distance. We firstly search salient locations in the images to be compared. Local visual features are then extracted from salient regions and projected onto a set of SOM-based learned visual prototypes, resulting in a visual activity map. This activity map is represented by a neural activity histogram coding the frequency of prototype appearance. By construction, the inter-bins similarity is embedded in this vector. Finally, we measure the image similarity by a quadratic distance [4] between neural activity histograms. This distance allows to quantify common and distinct concepts between the compared images.

This method has been experimented for two kinds of applications: a supervised image classification problem where the system detects 75.60% of correct classification rules on a database containing 4200 photos for 5 categories; an adult content filtering method where the correct classification rate reaches 84.47%.

This paper is organized as follows: In Section 2, we first present our image classification scheme based on SOM learning from ROI descriptions. Then, Section 3 contributes to some experimental results. And finally, conclusions are discussed in Section 4.

2 Supervised Image Classification Scheme

2.1 System Architecture

As outlined in [3], a classification scheme is generally composed of three main steps: pre-processing, feature extraction and classification. In this paper, we mainly focus our attention on the two first items, the last being performed by a k-nn algorithm.

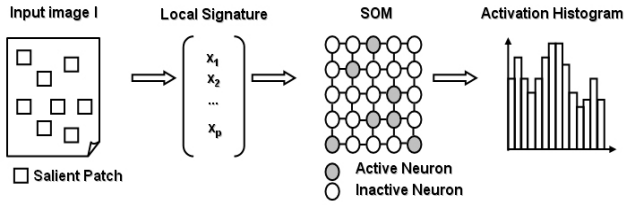

Figure 1. System architecture.

In our approach, the pre-processing step consists of detecting some salient points in the image to be compared, reducing thus the zones of interest to a limited number of regions considered as been perceptually relevant. From each detected salient point, a salient patch is extracted (Figure 1) and a local feature vector is calculated. Each local feature vector is then fed into a SOM network resulting a neural activity map composed of all winning cells. Finally, to complete the feature extraction step, the obtained activity map is converted into a neural activity histogram measuring the frequency of SOM prototype appearance. Assessing the similarity between two images is then reduced to compute the quadratic distance [4] between the associated neural activity histograms. Using this distance is offered by the topology preservation property of SOM [7]. Indeed, as neighboring cells in the SOM network are tuned to similar visual features, an inter-bin similarity matrix can be easily constructed, collecting the similarity between SOM cells.

The different computational steps used in this method are detailed in the next sections.

2.2 Regions of Interest Detector

According to the active vision mechanisms, the goal of salient point detectors is to find perceptually relevant image locations. Many detectors have been proposed in the literature [5] [2] [8]. The salient locations selected by human visual system contain generally high contrast, lines and edges [6]. Following this observation, we focus our interest on the Harris detector [5], the contrast detector [2] and a wavelet salient point detector [8].

The Harris's detector [5] aims at locating salient zones on corner by searching for the maxima of a function based on the local autocorrelation matrix of the signal. The second descriptor [2] proposes to locate salient points in high contrasted area. For this purpose, a multi-resolution contrast pyramid is built and can be viewed as a saliency map. The third salient point detector [8] uses a wavelet analysis to find pixels on sharp region boundaries.

2.3 Self Organized Map Learning

The Kohonen model [7] is based on the construction of a neuron layer in which neural units are arranged in a lattice L. Usually, the lattice is two dimensional (rectangular or hexagonal). The neural layer is innervated by d input fibers, called *axons* , which carry the input signals and excite or inhibit the cells via synaptic connections. As underlined in the previous section, the Kohonen network aims at preserving the topology of the input space and at tuning each cell to a particular set of stimuli. To reach these goals, the excitation of neurons has to be restricted to a spatially localized region in L and the location of this region has to be determined by those neurons that respond most intensively to a given stimulus. Moreover, L acts as a topographic feature map if the location of the most strongly excited neurons is correlated in a regular and continuous fashion with a restricted number of signal features of interest [11]. Neighboring locations in L correspond thus to stimuli with similar features. To satisfy these properties, a neighboring function between cells must be added in the network model. For this purpose, each cell $i \in L$ is connected to a set $N_L(i)$ of neighboring cells, defining thus a topological ordering. The goal of the Kohonen learning algorithm is then to adapt the shape of L to the distribution of the input vectors. As shown in Figure 2, the 2D lattice shape changes during the learning process to capture the input information and the topology existing in the input space. Those two properties can be considered as a competitive learning and a topological ordering.

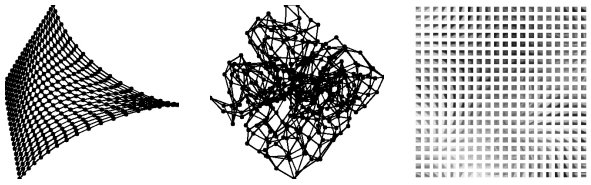

Figure 2. (a)Linear initialization. (b)Final topological evolution. (c)Patches projection on SOM lattice.

Let us now describe the SOM algorithm by assuming a SOM lattice structure composed of $N \times N$ neural units. Let M be the input space and $X = x(t)$ be a set of observable samples with $x(t) \in M \subset \Re^d$, $t \in \{1, 2, \ldots\}$ being

the time index. Supposing $M = m_i(t)$ is a set of reference vectors with $m_i(t) \in \Re^d$, $i \in \{1, 2, \ldots, N \times N\}$. For a linear initialization, we compute the two eigenvectors of the autocorrelation matrix of x that have the largest eigenvalues. The rectangular lattice is then defined along a 2D linear subspace spanned by this two eigenvectors.

If $x(t)$ can be compared simultaneously to all $m_i(t)$ by using a distance measure $d(x(t), m_i(t))$ in the input space, then the best matching unit (BMU) $m_c(t)$ is defined by :

$$m_c(t) = \arg\min_i d(x(t), m_i(t)), \forall i = 1, 2, \ldots, N \times N. \quad (1)$$

A kernel-based rule is used to reflect the topological ordering observed in the human visual cortex. The updating scheme aims at performing a stronger weight adaptation at the BMU location than in its neighborhood. This kernel-based rule is defined by :

$$m_i(t+1) = mi(t) + \alpha(t) h_{ci}(t)[x(t) - m_i(t)], \quad (2)$$

where $\alpha(t)$ designates the learning rate i.e. a monotonically decreasing sequence of scalar values with $0 < \alpha(t) < 1$.

$h_{ci}(t)$ represents the neighborhood function that governs the strength of weight adaptation as well as the number of reference vectors to be updated. Classically, a Gaussian function is used, leading to :

$$h_{ci} = \exp -\frac{||r_c - r_i||^2}{2\delta(t)^2}. \quad (3)$$

Here, the Euclidian norm is chosen and r_i is the 2D location for the i^{th} neuron in the network. $\delta(t)$ specifies the width of the neighborhood during time t.

2.4 Similarity Evaluation

When the SOM learning is over, the last step consists in quantifying the visual similarity between two images I_1 and I_2. The extracted patches from I_1 and I_2 are presented to the learned SOM. Each patch actives a particular cell (i.e., the BMU) and increments an activity histogram corresponding to answer responses for each neuron : $\forall i \in \{1, 2, \ldots, N \times N\}$,

$$H_1(i) = card\{p \in I_1, ||p - m_i|| < ||p - m_j||, \forall j \neq i\} \quad (4)$$

To obtain a probability distribution, the histogram is then normalized. The same strategy is applied to image I_2 to build the H_2 histogram. To quantify the image similarity we compare these two histograms. This is made possible by the preserving topology property of SOM : two input vectors closed in the observation space are now close in the SOM lattice and activate two cells in a small neighborhood. We use here an efficient quadratic histogram distance from Hafner and al. studies [4]. In particular, the quadratic distance D_Q uses an inter-bins similarity matrix A. This quadratic distance is defined by :

$$D_Q(H_1, H_2) = (H_1 - H_2)^t A (H_1 - H_2), A = |a_{ij}| \in \Re^{N \times N}, \quad (5)$$

where a_{ij} represents the similarity between the i^{th} element of H_1 and the j^{th} element of H_2. We have to keep in mind that indices i and j are the activated SOM neurons from I_1 and I_2 stimuli. The weight normalization denotes a value closed to 1 for similarity and closed to 0 for dissimilarity.

$$a_{ij} = 1 - \frac{||r_i - rj||}{\max_{i,j} ||r_i - r_j||}. \quad (6)$$

Finally, for determining the test image category, we compute the k-nearest learning images based on the previous criterion (5) and a k-nn classification is performed.

3 Experimental Results

For all the experiments, we configure our SOM network with the following rules to reach good learning results in terms of accurate input data representation [7] :

- the learning steps are 500 times the number of cells ;
- the available samples are applied cyclically ;
- $\alpha(t) = \frac{T}{T+99t}$ forms a monotonically decreasing sequence, with the number T of learning steps ;
- the width of neighborhood function $\delta(t)$ decreases linearly from $\frac{\sqrt{2N^2}}{2}$ to 0.5 .

The first experiment is to test how our system behaves with large databases (1200 learning images, 3000 test images). These images are divided into 5 categories: buildings, flowers, motorbikes, mountains and planes. We decide to evaluate the IP detector influence on the learning step. The results are exposed in Figure 3. This experimentation shows that the best configuration for clustering these 5 categories is a wavelet based detector [8] with 7x7 patches associated to a 20x20 SOM. The patch dimension and the SOM lattice size are chosen by experimental results. Indeed, the classifier offers a correct detection rate averaged across categories of 75.60%. Some correctly classified examples are shown in Figure 4.

The confusion matrix of the image classification system with the best configuration is shown in Table 1. The best classified category is motorbikes: 83.17% are correctly detected; The worst case happened to the flower category. This can be explained by the large variety of color and shape in this cluster.

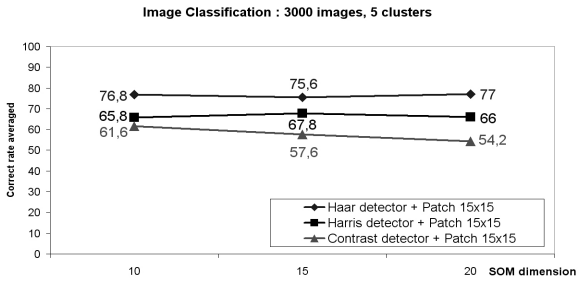

Figure 3. SOM with different IP detectors.

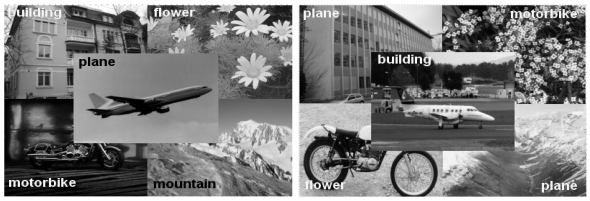

Figure 4. Correct and misclassified images.

Classified as →	A (%)	B (%)	C (%)	D (%)	E (%)
A=building	72.00	6.17	9.67	5.17	7.00
B=flower	2.17	67.83	23.67	4.67	1.67
C=motorbike	2.67	12.50	83.17	0.67	1.00
D=mountain	4.17	6.33	5.67	78.17	5.67
E=plane	8.50	4.33	4.17	6.17	76.83

Table 1. Confusion matrix for 5 clusters

In a context of web content filtering, we try to detect harmful content to censure pornographic images. The database is downloaded from Internet and is composed of 1110 adult images and 1200 benign images. The second category, known to be the rest of the world, is mainly constituted of landscapes, portraits and life scenes. 733 images of each category is preserved for the learning database. This study illustrates, with the salient point detector previously determined, interesting results with 7x7 patches. Indeed, 78.59% of adult images are correctly classified and 79.31% of benign images are recognized (Table 2). By adding the mean color descriptor (MCD)[9], we increase performances comparatively to a simple RGB patch. We can note that our approach gives better results than a classical MCD strategy.

Approach	Descriptor	Adult	Benign	Mean
Patch3x3	Haar + RGB + SOM 20x20	72.94	75.38	74.16
Patch7x7	Haar + RGB + SOM 20x20	78.59	79.31	78.35
Patch7x7	Haar + MCD	77.98	75.68	76.83
Patch7x7	Haar + MCD + SOM 20x20	88.86	80.09	84.47

Table 2. Adult and benign detection rates

4 Conclusion

In this paper, we proposed an original classification system using directly patches information. Based on the two main properties of SOM - which are dimension reduction and topology preservation - this architecture features image categories with activity histograms. In order to quantify the visual similarity between two images, we only need to compare their individual histogram. This solution implemented for image classification gives us very promising results. However, a growing and pruning strategy or a hierarchical SOM could be useful for learning large databases. Further improvements may be applied by coupling SOM algorithm and vector quantization methods as LVQ. Another possible issue is to learn activity maps built by SOM with a supervised neural network to get more robustness.

References

[1] Berretti S., Del Bimbo A., and Vicario E. Efficient matching and indexing of graph models in content-based retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1089–1105, october 2001.

[2] Bres S. and Jolion J.M. Detection of interest points for image indexation. *In 3rd Int. Conf. on Visual Information Systems*, pages 427–434, June 1999.

[3] Duda R.O., Stork D.G., and Hart P.E. *Pattern Classification*. Wiley Interscience, 2000.

[4] Hafner J., Sawnhey H., Equitz W., Flickner M., and Niblack W. Efficient color histogram indexing for quadratic from distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 17(7), july 1995.

[5] Harris C. and Stephens M. A combined corner and edge detector. *Proc. Fourth Alvey Vision Conf.*, pages 147–151, 1988.

[6] Hoffman J. E. and Subramanium B. The role of visual attention in saccadic eye movements. *Perception and Psychophysics*, 57:787–795, 1995.

[7] Kohonen T. *Self-Organizing Maps*. Springer-Verlag, Berlin, Heidelberg, New York, 2001.

[8] Laurent C., Laurent N., Maurizot M., and Dorval T. In depth analysis and evaluation of saliency-based color image indexing methods using wavelet salient features. *Multimedia Tools and Application*, 2004.

[9] Manjunath B. S., Ohm J-R., Vasudevan V., and Yamada A. Color and texture descriptors. *IEEE Trans. Circuits Syst. Video Techn.*, 11(6):703–715, 2001.

[10] Mikolajczyk K. and Schmid C. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[11] Ritter H., Martinetz T., and Schulten K. *Neural Computation and self-Organizing Maps : an introduction*. Addison-Wesley, New York, 1992.

[12] Tversky A. Features of similarity. *Psychological Review*, 4(84):327–352, 1977.