# Efficient Codebooks for Visual Concept Recognition by ASSOMs Activation

**Grégoire Lefebvre and Christophe Garcia**
France Telecom R&D - Orange - TECH/IRIS/CIM
4, Rue du Clos Courtel
35512 Cesson S´evign´e Cedex – France
{gregoire.lefebvre,christophe.garcia}@orange-ft.com

*(Paper received on August 10, 2006, accepted on September 25, 2006)*

**Abstract.** In this paper, we propose a novel method for robustly characterizing and classifying visual concepts, and more precisely for detecting several categories of complex objects in images. Toward this aim, we propose a scheme that relies on Adaptive-Subspace Self-Organizing Maps (ASSOMs). Robust local signatures are first extracted from training object images and projected into specialized ASSOM networks. The extracted local signatures activate several neural maps producing activation energies. These activation energies are then fused into global feature vectors representing the object images. Object recognition is then performed via a supervised SVM (Support Vector Machine) classification. A multiscale search approach completes the system in order to obtain the object localization and identification in complex scenes. The proposed method allows a good detection rate of 85.08% for the PASCAL 2005 challenge1, composed of 689 complex real world images, containing four different objects undergoing strong variations in shapes, sizes, poses and illumination conditions.

## 1. Introduction

According to several psycho-visual experiments [1], the human vision system performs saccadic eye movements between salient locations to capture image content. Many systems in computer vision are inspired by this observation, in order to describe visual information for image classification or retrieval. In opposition to global approaches, for which a signature is computed by considering all pixels in the image, local approaches represent image content via a set of local signatures centered on interest points (IP) [2–4], which are extracted on perceptually important areas.

Tversky studies [5] showed that when we compare two images, we detect common and distinct concepts between the regions around the IPs. Our method aims at reproducing these concepts with a codebook learning strategy based on ASSOM activation maps for each category. Visual similarity is then estimated by the distance between different activation histograms.

Our method has been experimented in the context of an object detection task where the good detection rate reaches 85.08% for 689 complex real world images, from the Pascal 2005 Challenge[1], containing four different object categories.

[1] http://www.pascal-network.org/challenges/VOC/voc2005

© H. Sossa and R. Barrón (Eds.)
Special Issue in Neural Networks and Associative Memories
Research in Computing Science 21, 2006, pp. 61-71

This paper is organized as follows. In Section 2, we first present our object detection scheme based on ASSOM activation energies. Then, Section 3 demonstrates our system performances with some experimental results. Finally, conclusions are drawn.

## 2. Object Detection Based on ASSOM Energies

### 2.1. Proposed Scheme Overview

As outlined by R.O. Duda [6], a classification scheme is generally composed of three main steps: pre-processing, feature extraction and feature classification. In the proposed study, we mainly focus on the two first steps, the last step being performed by a SVM classifier.

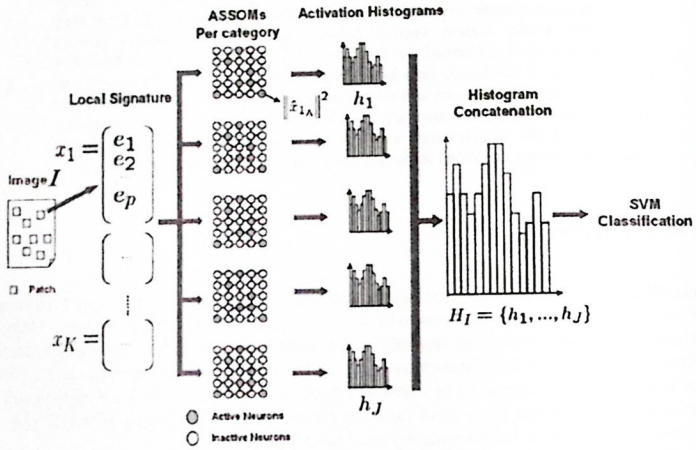


**Fig. 1.** The Proposed System Architecture.

Our system architecture consists of six steps in the learning phase (see Figure 1):

- We first locate the salient zones with an IP detector [2] mainly on sharp region boundaries.
- Local visual features are then extracted in order to describe the orientation and the regularity of the singularities contained in the different patches around each detected IP.

- Visual feature vectors are fed into specialized ASSOM networks to characterize the main visual prototypes via neural activation maps. These maps synthesize the activation energies for each category.
- Activation energies are represented by activation histograms for each class.
- These histograms are concatenated to build the image global feature vector.
- Finally, a SVM classifier is trained in a supervised way from these discriminative global feature vectors.

### 2.2 Regularity Foveal Descriptor

Most local descriptors represent the neighborhood around salient points by characterizing edges in this area [7]. To describe edges, gradient orientations and magnitudes are generally used. In a recent study [8], it has been shown that an edge or more generally a singularity can also be efficiently characterized by considering its Hölder exponents.

**Definition 1.** *$f : [a,b] \rightarrow \mathbb{R}$ 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 $\alpha$, $h_f(x_0) = sup\{\alpha, f \text{ is Hölder } \alpha \text{ at } x_0\}$.*

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 $h_f(x_0)$ is, the more singular the signal is. For example, the Hölder exponent is 1 for a triangle function, 0 for a step function and −1 for a Dirac impulse.

To describe an ROI associated to an interest point in an image $I_f$, both orientation and Hölder regularity of singularities are characterized. The Hölder exponent is estimated in the gradient direction. For this purpose, orientation $\theta(x,y)$ and gradient magnitude $m(x,y)$ are computed at each pixel $(x,y)$ :

$$\begin{cases} m(x,y)^2 = (I_f(x+1,y) - I_f(x-1,y))^2 \\ \qquad\qquad +(I_f(x,y+1) - I_f(x,y-1))^2 \\ \theta(x,y) = \tan^{-1}\left(\frac{I_f(x,y+1) - I_f(x,y-1)}{I_f(x+1,y) - I_f(x-1,y)}\right) \end{cases} \tag{1}$$

Then, for each singularity, the Hölder exponent $\alpha$ is estimated with foveal wavelets as presented in [9]. Orientations and Hölder exponents maps are then conjointly used and we approach their distribution with 3D histograms. To build such histograms, we consider a 32x32 ROI around each IP that we split into 16 8x8 subregions and we quantify the number of times each pair $(\alpha,\theta)$ appears in each subregion (see Figure 2). We use three Hölder exponents bins into the range $[-1.5, 1.5]$ and eight orientation bins into $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. All bins of each subregion 3D histogram are concatenated to form the final signature : the Regularity Foveal Descriptor (RFD).

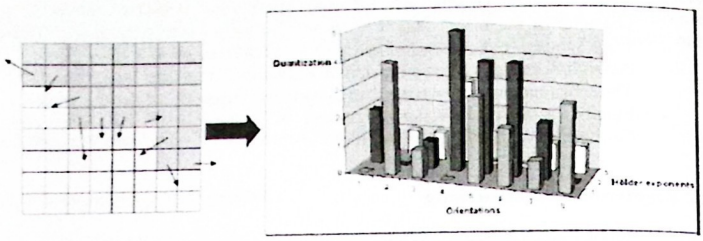


**Fig. 2.** Orientations and Hölder exponents for a subregion, resulting in a 3D histogram.

### 2.3. ASSOM Learning Process

ASSOM is basically a combination of a subspace method and a competitive selection and cooperative learning as in the traditional SOM, introduced by Kohonen [10]. ASSOM differs from other subspace methods by permitting to generate a set of topologically-ordered subspaces. Two units that are close in the map will represent two feature subspaces closed in the global feature space. In ASSOM, the unit is composed of several basic vectors that expand together a linear subspace. This unit is called "module" in an ASSOM neural network. This method aims at learning data features, without assuming any prior mathematical forms of their representation, such as Gabor or wavelet transforms, which are frequently encountered in the traditional image analysis and pattern recognition techniques. In other words, the forms of the filter functions are learned directly from the data. The input to ASSOM is a group of vectors, called "episode". The vectors in each episode are supposed to be close according to some affine transformation variations.

There are mainly two phases in the learning process with ASSOM:

1. For an input episode, locate the winning subspace from ASSOM modules.
2. Adjust the winning subspace and its neighbor modules in order to better represent the input episode.

For a linear subspace $\mathcal{L}$ of dimensionality $H$, one can find a set of basis vectors $\{b_1, b_2, \ldots, b_H\}$, such that every vector in $\mathcal{L}$ can be represented by a linear combination of these basis vectors. Such sets of basis vectors are not unique. However they are equivalent in the sense that they expand exactly the same subspace. For computational convenience, the basis vectors are orthonormalized by the Gram-Schmidt process.

The orthogonal projection of an arbitrary vector $x$ onto the subspace $\mathcal{L}$, written as $x_{\mathcal{L}}$, is a linear combination of its orthogonal projections on the individual basis vectors, and can be computed by:

$$\hat{\mathbf{x}}_{\mathcal{L}} = \sum_{h=1}^{H} (\mathbf{x}^{\mathrm{T}} \mathbf{b}_h) \mathbf{b}_h. \tag{2}$$

If $\hat{\mathbf{x}}_{\mathcal{L}} = \mathbf{x}$, then x belongs to $\mathcal{L}$, otherwise we can define the distance from x to $\mathcal{L}$ as $\|\mathbf{x}_{\mathcal{L}_1}\| = \|\hat{\mathbf{x}}_{\mathcal{L}_2}\|$, by using the Euclidean norm. When several subspaces exist, the original space is partitioned into pattern zones and the decision surface between two subspaces, for example $\mathcal{L}_1$ and $\mathcal{L}_2$, is determined by those vectors x such that $\|\hat{\mathbf{x}}_{\mathcal{L}_1}\| = \|\hat{\mathbf{x}}_{\mathcal{L}_2}\|$. By comparing the distances of a vector to all the subspaces, we can assign this vector to the nearest subspace.

In Kohonen's realization of ASSOM, the subspace is represented by a twolayered neural architecture, as depicted in Figure 3. The neurons in the first layer compute the orthogonal projections $\mathbf{x}^{\mathrm{T}} \mathbf{b}_h$ of the input vector x on the individual basis vectors $\mathbf{b}_h$. The second layer is composed of a single quadratic neuron and computes the squared sum from the outputs of the first layer neurons.

The output of the whole neural module is then $\|\hat{\mathbf{x}}_{\mathcal{L}}\|^2$, the square of the norm of the projection. It can be regarded as a measure of the degree of matching of the input vector x with the subspace $\mathcal{L}$ represented by the neural module. In the case of an episode, the distances should be calculated from the subspace of the vectors in the episode and the subspace of the module, which are generally difficult to compute. Kohonen proposed another much easier but robust definition of subspace matching: the *energy*: the sum of squared projections over the episode on a module subspace. This is the energy that we use to build our activation histogram.

The classical Kohonen's ASSOM learning algorithm proceeds as follows.
For the learning step $t$,

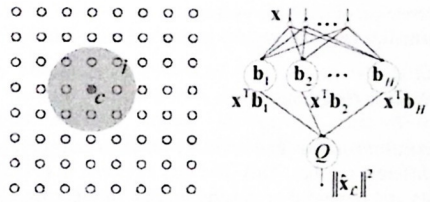


**Fig. 3.** Left: a rectangular ASSOM topology, a winning module $c$ and its neighborhood. Right: the projection of x on $\mathcal{L}$ by a module.

1. Feed the input episode **x**, composed of $S$ vectors **x**(s), $s \in S$. Locate the winning module indexed by $c$:

$$c = \arg\max_{i \in I} \sum_{s \in S} \|\mathbf{x}_{\mathcal{L}_i}(s)\|^2. \tag{3}$$

where $I$ is the set of indices of the neural modules in the ASSOM.

2. For each module $i$ in the neighborhood of $c$, including $c$ itself, and for each input vector $\mathbf{x}(s)$, $s \in S$, adjust the subspace $\mathcal{L}_i$ by updating the basis vectors $\mathbf{b}_h^{(t)}$, according to the following procedure:
   (a) Rotate each basis vector according to:

$$\mathbf{b}_h^{(t)} = \mathbf{P}_c^{(t)}(\mathbf{x}, t)\mathbf{b}_h^{'(t)} \quad . \tag{4}$$

   In this updating rule, $\mathbf{b}_h^{(t)}$ is the new basis vector after rotation and $\mathbf{b}_h^{'(t)}$ the old one. $\mathbf{P}_c^{(t)}(\mathbf{x}, t)$ is the rotation operator matrix, defined as:

$$\mathbf{P}_c^{(t)}(\mathbf{x}, t) = \mathbf{I} + \lambda(t) h_c^{(t)}(t) \frac{\mathbf{x}(s)\mathbf{x}^{\mathrm{T}}(s)}{\|\dot{\mathbf{x}}_{\mathcal{L}_i}(s)\| \|\mathbf{x}(s)\|} \quad . \tag{5}$$

   where $\mathbf{I}$ is the identity matrix, $\lambda(t)$ a learning-rate factor that decreases with the learning step $t$. $h_h^{(t)}(t)$ is the neighborhood function defined on the ASSOM lattice with a support area shrinking with $t$.
   (b) Dissipate the components $b_{hj}^{(t)}$ of the basis vectors $\mathbf{b}_h^{(t)}$ to improve the stability of the results [10]:

$$\tilde{b}_{hj}^{(t)} = \mathrm{sgn}(b_{hj}^{(t)}) \max(0, |b_{hj}^{(t)}| - \varepsilon), \tag{6}$$

   where $\varepsilon$ is the amount of dissipation, chosen proportional to the magnitude of the correction of the basis vectors.
   (c) Orthonormalize the basis vectors in module $i$.

### 2.4. Final Feature Vector Construction

The proposed architecture contains one ASSOM for each category, producing specific ASSOM units for different patches. This idea was explored in [11] for the recognition of handwritten digits and produced promising results. In this case, the image size was small (25 × 20 pixels) allowing a straightforward learning of all pixels through ASSOM. 10 ASSOMs were used, one trained for each category of handwritten digits. For digit classification, a test digit is sent simultaneously to all the 10 ASSOMs, which output 10 reconstruction error values. The ASSOM with the smallest reconstruction error determines the digit category. An obvious limitation here is that there is no interaction between the different ASSOMs during the learning phase. An ASSOM learns the features of its own category, however it does not learn how to separate them from the other categories. The optimum decision surface is thus not guaranteed.

In our context, the images to analyze are much larger and complex. Therefore, we decide to use a local approach by extracting image patches at salient locations. Our

strategy is thus to build a visual dictionary for each class from the activation of different ASSOMS. Other studies [12–16] are interesting regarding the creation of codebooks or bags of keypoints. Here, our object approach focuses on the pertinent image areas.

To construct the feature vector $\mathbf{H}_{\mathcal{I}}$ from the object $\mathcal{I}$ we proceed as follows (see Figure 1 for notations).

- We select the interest points with the strongest salience from the image $\mathcal{I}$
- For each patch:
  - We compute the local signature with the RFD descriptor (see 2.2).
  - We build an episode for the local signature by applying some artificial affine transformations.
  - For each episode vector:
    * $J$ specialized ASSOM networks receive a signature and compute an energy $\|\hat{\mathbf{x}}_{kj}\|^2$ defined by:

$$\|\hat{\mathbf{x}}_{kj}\|^2 = \max_{i \in I_j} \|\hat{\mathbf{x}}_{k,i}\|^2 \quad . \tag{7}$$

where $I_j$ is the module index set of the $j^{th}$ ASSOM, $J$ is the number of category. $\|\hat{\mathbf{x}}_{kj}\|^2$ is the maximal value of the square of the norm of the projection of $\mathbf{x}_k$ on the linear subspaces of the $j^{th}$ ASSOM.

    * Each activation histogram $h_j$ corresponding to each network is then updated. The maximal output energy increments the corresponding histogram bin, as follows.

$$h_j[i^*](t+1) = h_j[i^*](t) + \|\hat{\mathbf{x}}_{kj}\|^2 \tag{8}$$

with $i^* = \arg\max_{i \in I_j} \|\hat{\mathbf{x}}_{k,i}\|^2$ and $t$ is the time.

- Each energy histogram $h_j$, computed from all patches, is fused into a global activation histogram $\mathbf{H}_{\mathcal{I}}$. This final feature vector is then the concatenation of the individual ASSOM energy histograms. This discriminative object information is finally introduced to a SVM classifier for supervised training.

### 2.5. Object Detection by multi-resolution

The visual concept learning process is tuned on normalized object samples. Thus, the classification procedure consists of an object search. This search is made by a fixed size sliding window in a multi-scaled image pyramid.

Here, the object detection is realized on three pyramid levels and the window moves with a step of the half of its width (see Figure 4).

For each image extracted inside the sliding window, we observe the SVM outputs. When the classifier recognize a learned object, the corresponding area is marked in a voting map. The vote weight is proportional to the SVM output given that the output $i$

of the SVM classifier represents the *a posteriori* probability of an object of belonging to class *Ci*.

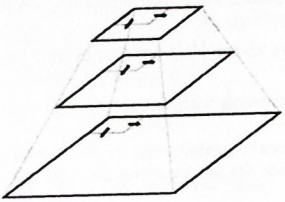

**Fig. 4.** Scaling pyramid and its moving window.

A last step fuses the different vote intensities in order to locate the object. This procedure clusters the multi-scale vote intensities in the original dimension (see Figure 5). Thus, we locate some candidate areas where the sum of multiscale votes is upper than a decision rule threshold. Typically, when one of the SVM outputs is greater than 0.9, we consider the corresponding area as relevant. Then, a final classification refines the results in the merged areas.

## 3. Experimental results

We tested the proposed scheme on the challenge PASCAL 20051 database. The goal is to classify 689 images using 684 training labeled objects of four different classes: bicycle, car, motorbike and people.

The local signatures are extracted around the IPs within 32x32 rectangular patches. The RFD descriptor is computed within 16 subregions of these patches, 8 orientation bins and 3 Hölder exponent bins for all patches. Therefore, the RFD dimension is 16x8x3=384.

For all experiments, we configure our ASSOM networks with the following rules for optimal performance in terms of accurate data representation:

– the number of training epochs is : $T = 500 \times N$;

– the learning rate forms a monotonically decreasing sequence: $\lambda(t) = \dfrac{T}{T + 99t}\lambda(t)$ :

– the neighborhood function is defined by: $h(i) = \begin{cases} 1, r_c - r_i < \mu(t) \\ 0, otherwise. \end{cases}$

Here, we choose the euclidean norm and $r_i$ is the 2D position of the $i^{th}$ ASSOM module. $\mu(t)$ specifies the neighborhood width which decreases linearly with $t$ from $\dfrac{\sqrt{2}}{2}N$ to 0.5.

The *Area Under Curve* (AUC) and the confusion matrix of the image classification system with the best configuration are shown in Table 1. The best-classified category is "cars": 92.39% are correctly detected. The worst case happened to be the people category. This can be explained by the large variety of textures, colors and shapes in this cluster.

**Table 1.** Confusion matrix and AUCs. (B=Bicycle, C=Car, M=Motorbike, P=People)

| Classified as → | B | C | M | P | Classes | AUC |
|---|---|---|---|---|---|---|
| B | 82 | 12 | 14 | 6 | B | 0.863 |
| C | 2 | 243 | 5 | 13 | C | 0.938 |
| M | 3 | 11 | 199 | 3 | M | 0.944 |
| P | 9 | 17 | 6 | 52 | P | 0.879 |


**Fig. 5.** Good and false classifications. Voting maps for a bicycle image.

Some correctly classified examples are shown in Figure 5. We can observe that a bicycle object stimulates strongly the bicycle-voting map. Moreover, the motorbike-voting map is lightly activate, which demonstrates the generalization power of the proposed system.

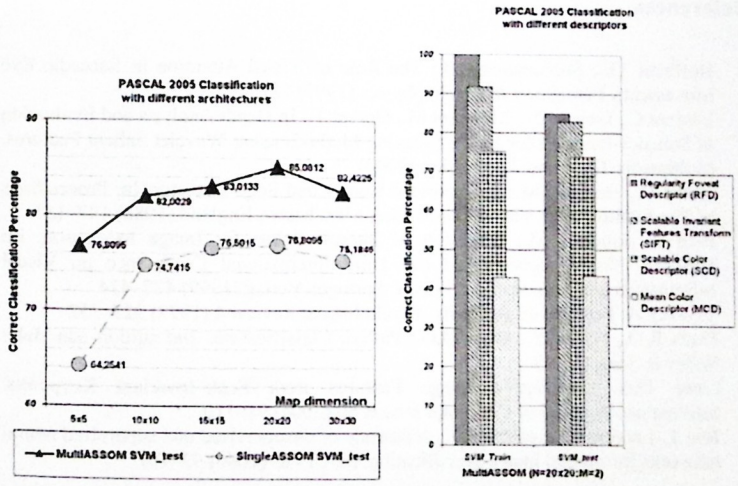


**Fig. 6.** Results on PASCAL 2005 database with architecture and descriptor variations.

This multi-ASSOM architecture offers a better global classification rate (85.08%) than a single ASSOM (76. 81%) for all test classes. The best configuration for an ASSOM network is : N=20x20. M=2 (see Figure 6). It is worth noting that the global classification rate for the training database reaches 100% for our multi-ASSOM scheme. and only 89.96% for the single ASSOM scheme.

It is also interesting to compare the performances when using different descriptors. With the same configuration. the RFD descriptor provides better results than the SIFT descriptor or some MPEG-7 descriptors. Consequently, we can see that the ASSOM competition with the RFD descriptor allows us to construct more discriminative feature vectors for the SVM classification[2].

## 4. Conclusion

This article describes a new system to detect visual concepts. using singularity information contained in the salient regions of interest. Based on the three main properties of ASSOM - which are dimension reduction, topology preservation and invariant feature emergence - our scheme give very promising results to detect objects with a SVM classifier.

We plan to study the fusion of heterogeneous descriptors with feature selection to learn useful object information. and to develop a growing strategy to find the optimal ASSOM parameters.

## References

1. Hoffman J.E., Subramaniam B.: The Role of Visual Attention in Saccadic Eye Movements. Perception & Psychophysics (1995) 787–795
2. 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)
3. Harris C., Stephens M.: A Combined Corner and Edge Detector. In: Proceedings of The Fourth Alvey Vision Conference. Manchester, England (1988) 147–151
4. Bres S., Jolion J.-M.: Detection of Interest Points for Image Indexation. In: VISUAL '99: Proceedings of the Third International Conference on Visual Information and Information Systems. Springer-Verlag (1999) 427–434
5. Tversky A: Features of similarity. Psychological Review 4 (1977) 327–352
6. Duda R.O., Hart P.E., Stork D.G.: Pattern Classification. 2*nd* edition edn. John Wiley & Sons (2001)
7. Lowe D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60 (2004) 91–110
8. Ros J., Laurent C., Lefebvre G.: A cascade of unsupervised and supervised neural networks for natural image classification. In: CIVR. (2006) 92–101

[2] WEKA SVM classification (http://www.cs.waikato.ac.nz)

9. Mallat S.: A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (1989) 674–693
10. Kohonen T.: Self-Organizing Maps. Springer (2001)
11. Zhang B., Fu M., Yan H., Jabri M.A: Handwritten digit recognition by adaptivesubspace self-organizing map (ASSOM). IEEE Transactions on Neural Networks 4 (1999) 939–945
12. 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
13. Quelhas P., Monay F., Odobez J.-M., Gatica-Perez D., Tuytelaars T., Van Gool L.: Modeling scenes with local descriptors and latent aspects. In: IEEE Int. Conf. on Computer Vision. (2005) IDIAP-RR 04-79.
14. Fei-Fei L., Fergus R., Perona P.: One-shot learning of object categories. IEEE Transactions on PAMI 4 (2006) 594–611
15. Lazebnik S., Schmid C., Ponce J.: Spatial pyramid matching for recognizing natural scene categories. IEEE CVPR 2 (2006) 2169–2178
16. Lefebvre G., Laurent C., Ros J., Garcia C.: Supervised image classification by som activity map comparison. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Volume 2. (2006) 728– 731.