A Heterogeneous Descriptor Fusion Process For Visual Concept Identification

Grégoire Lefebvre, Christophe Garcia

A Heterogeneous Descriptor Fusion Process For Visual Concept Identification

Grégoire Lefebvre
Orange Labs - R&D Division
4 Rue du Clos Courtel
35512 Cesson Sévigné - France
Email: gregoire.lefebvre@orange-ftgroup.com

Christophe Garcia
Orange Labs - R&D Division
4 Rue du Clos Courtel
35512 Cesson Sévigné - France
Email: christophe.garcia@orange-ftgroup.com

***Abstract*—In this paper, we propose a novel method for robustly classifying visual concepts. In order to achieve this aim, we propose a scheme that relies on Self Organizing Maps (SOM [6]). Heterogeneous local signatures are first extracted from training images and projected into specialized SOM networks. The extracted signatures activate several neural maps producing activation histograms. These activation histograms are then combined into a global fusion process in order to build our final image representation. This fusion scheme is generic and shows promising results for automatic image classification and objectionable image filtering.**

Keywords: SOM, bag of keypoints, image classification.

I. Introduction

In many applications, such as data mining, object recognition, image classification, etc., the data input dimension may be very large, and these features may be provided by different captors or descriptors. Thus, it is a very challenging problem for an analyst to extract the pertinent features or to detect useful structure inside this volume of information. The main idea of this study is to automatically learn visual concepts from learning image series described with several heterogeneous features. Indeed, nowadays, it appears fundamental to take into account a multi-modal strategy to deal with face identification, video content-based retrieval or illicit content filtering. The combination of color, texture, and shape descriptors seems to be very helpful in performing classification rates. Classically, no knowledge about data distribution is *a priori* available. Consequently, the system should be able to extract the most important information in order to cluster and select these features before fusing the heterogeneous descriptions.

In the state-of-the-art methods, these three operations are realized sequentially and independently. The common techniques join the features in an *a priori* static fusion process. The feature extraction considers some similarity criterion (mutual information, feature correlation, etc.), and then operates information selection thanks to PCA (Principal Component Analysis) or LDA (Linear Discriminant Analysis) algorithms [18], and finally agglomerates the multi-modal descriptors with a simple concatenation [15]. Another issue to join heterogeneous features is called "*a posteriori* dynamic fusion process", proposing to weight different classifier results with a boosting approach (*e.g.* Adaboost [4]). The main drawback of these strategies is the lack of communication between the heterogeneous feature fusion and discriminative information learning. That is why, here, our approach tries to keep crucial features from heterogeneous descriptors by clustering, selecting and fusing them in one learning process.

This paper is organized as follows. In Section II, we present our image categorization scheme based on SOM activation histograms. Then, Sections III and IV expose how to extract relevant image information before fusing them in a final neural model. In Section V, we demonstrate our system's performance with several experimental results and finally conclusions are drawn.

II. Proposed Scheme Overview

This article presents a new strategy to automatically classify image contents from heterogeneous information. Classically, image classification is realized in a competitive way between several descriptors, and the best configuration results give the best descriptor choice. Here, we propose to use all descriptors to better represent visual concepts. Nevertheless, this huge volume of information should be structured by a selection and fusion strategy, without human intervention. That is why we investigate the use of SOM to select and structure the information. First, for each category, intermediate models are built, one for each descriptor. Then, a final model combines the intermediate representations to give a new image feature vector. The fusion learning process then takes place during the class model constitution, not in the feature extraction or in the final classification, as shown in Figure 1. This proposed architecture is composed of three steps: the first step describes image content with color, texture and shape information; the second step proposes our new image representation; and finally a classifier takes the decision about the image class. Thus, for the first and the last steps, we refer to well known state-of-the-art results. Describing an image, we extract some points of interest thanks to a salient point detector [1], [5], [7], and for each region of interest we compute a local signature focusing on color [16], texture [11] and shape [12] information. In the same way, the third step is performed via a classifier: KNN (K-Nearest Neighbor), MLP (Multi-Layer Perceptron [13]), RBF (Radial Basis Function [2]) or SVM (Support Vector Machine [17]).

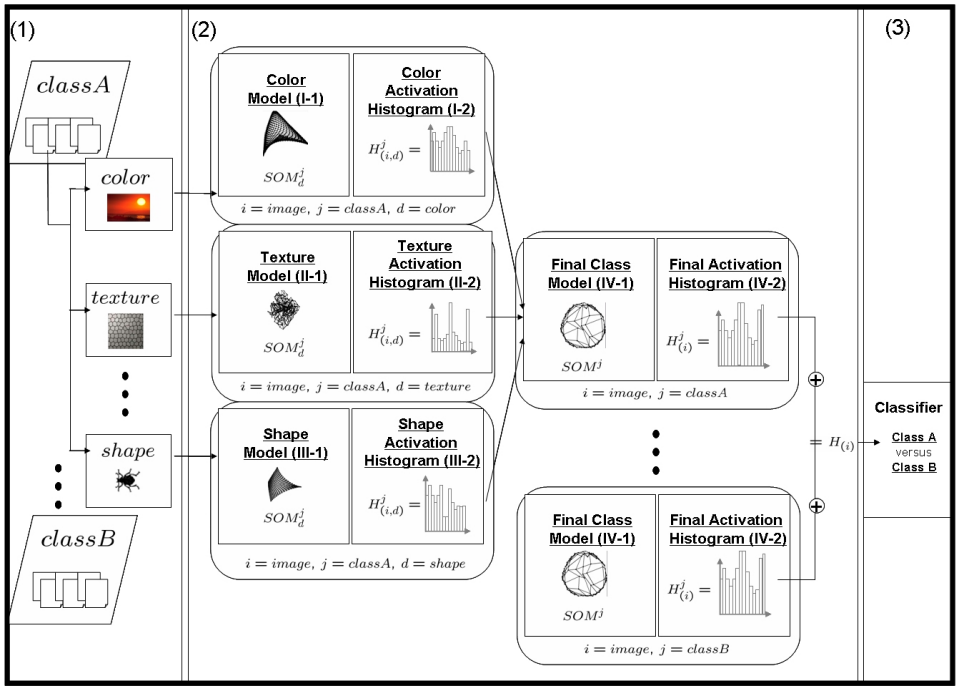

Figure 1. The Proposed System Architecture

A. Notation

In the following, let us use the following notations to build our fusion strategy (cf. Figure 1):

- $H^j_{(i,d)}$ is the activation histogram of the image i for the category j, obtained with the intermediate model SOM^j_d from its signatures d;
- $H^j_{(i)}$ is the activation histogram of the image i for the category j, obtained with the final model SOM^j;
- $H_{(i)}$ is the final feature vector of the image i for all categories, used as an input to a classifier.

B. Fusion Strategy

Our main contribution corresponds to the second scheme step where we propose a topological map as an image class model (cf. Figure 1 IV-1). This final single model per image category identifies its visual concept by a neural activation histogram (cf. Figure 1 IV-2). Indeed, for a test image, the model activation strength designs the belonging class, as shown in our previous studies [8], [9], [19]. Thus, the final classifier learns the adequacy between a class model and its learning samples in order to deduce which class model is the most representative for a test image. Our final class model construction is built as follows:

- Here, we use a SOM for each image descriptor. More precisely, the first layer of the second step is composed of 3 SOMs: one for color, one for texture and the third one for shape information. The SOMs are known to automatically emerge descriptors in a competitive and adaptive process. The unsupervised SOM learning process offers three main properties: dimension reduction, topology preservation and data accommodation. Thus, each map synthesizes a particular type of feature for the current category, performing a process of selecting and clustering. Indeed, the learning map operation brings to light inside the 2D lattice the best matching units (BMUs) which better represent the input data in terms of visual similarity. Finally, we focus on BMU neural activity during the learning signature stimulation. This activation is calculated with the quantization error eq_c of the BMU c with the signature x_d of the image i for the descriptor d (see Algorithm 1 in Section IV).

- The activation histogram $H^j_{(i,d)}$ of the image i with its descriptor d of the class j agglomerates the neural activation of the intermediate SOM model. This histogram becomes an intermediate signature, representing the image content for the considered descriptor d. Thus, we use several intermediate image representations (See Figure 1 I-2, II-2 and III-2) which have the same dimension (the SOM size) for different descriptors. These activation histograms $H^j_{(i,d)}$ can be compared and learned, corresponding to the adequacy between the color, texture or shape models with its learning examples.

- When an image is proposed to the system, we focus on its three intermediate representations. The three activation histograms are then learned by a final SOM^j per category (cf. Figure 1 IV-1). We randomly inject the

intermediate activation histograms in this new SOM and the fusion is performed during the SOM learning process to build the final model. This fusion is made possible because the intermediate representations are comparable, using the neural map activations belonging to the same data space.

- The final classification competes on final image representations (cf. Figure 1 IV-2). The final image representation of the image i is an activation histogram $H_{(i)}$, the result of a concatenation of all final class stimulation $H^j_{(i)}$.

III. Feature Extraction and Structuring

A. Salient Point Detection

The first step of an image categorization scheme is the feature extraction. Extracting local signatures is realized here in two steps: the salient point detection and the region of interest description. The goal of salient point detectors is to find perceptually relevant image locations. Many detectors have been proposed in the literature. In this paper, we investigate three types of interest point (IP) detectors, each one focusing on a particular local property of the image content such as corners [5], contrast [1], edges [7] detailed in the following:

(a) (b) (c)

Figure 2. IP detection with Harris (a), contrast (b) and wavelet (c).

- The Harris detector [5] aims to locate salient zones on corners by searching for the maxima of a function based on the local autocorrelation matrix of the signal. The authors observed that a corner is a place where the signal varies strongly in two directions. Based on this fact, Harris and Stephens proposed to compute a corner response function at each pixel to indicate the presence of a corner.

- Motivated by the sensitivity of the human visual system to the multi-resolution and the contrast, the detector presented in [1] proposes to locate salient points in high contrast areas. Bres and Jolion define a multi-resolution pyramid of contrast by computing the local and background luminance on different scales. Each point in the pyramid can thus be characterized by a significant local amount of the normalized contrast measure. This normalized contrast measure allows the extraction of an interest point at a different resolution level by searching for the local maxima above a predefined threshold.

- The salient point detector in [7] uses a Haar wavelet analysis to find pixels on sharp region boundaries. Working with wavelets is justified by the consideration of the human visual system for which multi-resolution, orientation and frequency analysis are of prime importance. In order to extract the salient points, a wavelet transform is firstly performed on the grayscale image. The obtained wavelet coefficients are represented as introduced by Shapiro [14]. This tree is then scanned for the first time from leaves to the root to compute the saliency value at each node. A second scanning occurs in order to determine the salient path from the root to the locations on the original image, where the raw salient points are located.

B. Local Patch Description

In our study, we are interested in three different types of descriptors to perform a feature fusion. Thus, we present two MPEG-7 descriptors [11] to describe color and texture and a local singularity descriptor proposed in a recent study [12]:

- The Histogram Color Descriptor (HCD [16]) is a color histogram in the RGB color space that represents color distribution in the region of interest. Here, we use 32 bins per color channel, resulting in a signature dimension of 32^3.

- The Homogeneous Texture Descriptor (HTD [11]) is designed to characterize the properties of texture, based on the assumption that the visual properties of the texture are relatively constant over the region. The descriptive features are extracted from a bank of orientation and scale Gabor filters. Its size is 62.

- The Regularity Foveal Descriptor (RFD [12]) efficiently characterizes an edge by considering its Hölder exponents that evaluate the signal regularity. Then, for each singularity point, the Hölder exponent is estimated with foveal wavelets as presented in [10]. Gradient orientations θ and Hölder exponents α are then jointly used in 3D histograms. To build such histograms, a 32×32 patch around each interest point is split into 16 sub-regions and the number of times each pair (α, θ) appears in each sub-region is quantified. We use three Hölder exponent bins in the range $[-1.5, 1.5]$ and eight orientation bins in $[-\frac{\pi}{2}, \frac{\pi}{2}]$. All 3D histograms are concatenated to form the final signature. The dimension is then: $8 \times 3 \times 16 = 384$.

C. Spatial Information Structuring

In order to structure the local feature vectors into a "bag of features", we propose a system based on Kohonen topological maps. The SOM [6] is based on the construction of a neuron layer in which neural units are arranged in a lattice L as shown in Figure 3. The neural layer is innervated by some input fibers, called *axons*, which carry the input signals and excite or inhibit the cells via synaptic connections. The 2D lattice shape changes during the learning process to capture the input information and the topology existing in the input space. These two properties can be considered as competitive learning and topological ordering. At the end of the learning process, the patches are clustered in terms of common visual similarity and each SOM unit synthesizes the most recurrent local signature for each visual concept, composing our "bag of features".

Let us now describe the SOM algorithm by assuming an SOM lattice structure composed of U neural units. Let $X = x(t)$ be the sample set with $x(t) \in \Re^n$, $t \in \{1, 2, \ldots\}$ being the time index. Supposing $M = m_k(t)$ a set of reference

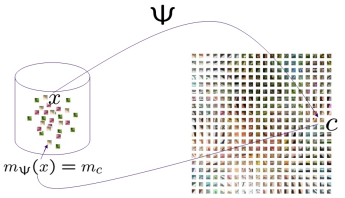

Figure 3. Patch projection on SOM lattice.

vectors (the weights) with $m_k(t) \in \Re^n$, $k \in \{1, 2, \ldots, U\}$, we define the BMU with (See Figure 3):

$$c = \arg\min_k ||x(t) - m_k(t)||, \forall k = 1, 2, \ldots, U. \quad (1)$$

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

$$m_k(t+1) = m_k(t) + \lambda(t)\phi_{ck}(t)[x(t) - m_k(t)], \quad (2)$$

where $\lambda(t)$ designates the learning rate i.e. a monotonically decreasing sequence of scalar values with $0 < \lambda(t) < 1$. $\phi_{ck}(t)$ represents the neighborhood function. Classically, a Gaussian function is used, leading to:

$$\phi_{ck} = \exp - \frac{||r_c - r_k||^2}{2\delta(t)^2}. \quad (3)$$

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

IV. Final Model Building

In the proposed method, the new final image representation is performed thanks to two following algorithms. The first one is the creation of three intermediate histograms per image, one for each input descriptor; and the second one builds the final image signature, proposing an image feature vector for the final classification.

For the first algorithm (cf. Algorithm 1), we focus on image information class by class. Firstly, we locate the points of interest for each image. Secondly, local signatures are computed describing color, texture and shape for each region of interest. In the third step, color signatures are used to structure a specialized SOM, representative of the main color information of the current category. Two others SOMs are learned in order to give a texture and a shape model of the current class. The first algorithm presents how intermediate histograms are built from the neural activation of the previously learned SOMs. The SOM neural activation is evaluated by the quantization error $eq_c(t)$ between the local signature $x_d(t)$ and its BMU $m_c(t)$ (cf. Equation 4). This value $eq_c(t)$ corresponds to the stimulation of the neuron c at the time t and we build a histogram $H^j_{(i,d)}(t)$ resuming all neuronal unit activation of one image i of the class j (cf. Equation 5).

For the second algorithm (cf. Algorithm 2), a similar process creates a final SOM from the three previously computed intermediate histograms $H^j_{(i,d)}$, class by class. Thus, each class information is finally structured by one SOM. Here, the histograms $H^j_{(i,d)}$ are used as input vectors for a new SOM learning process. The second algorithm presents how final feature vectors are built for each image. The SOM neural activation is then evaluated by the error quantization eq_γ between the local signature $H^j_{(i,d)}$ and its BMU $m_\gamma(t)$ (cf. Equation 6). Similar to the first algorithm, the histogram $H^j_{(i)}[\gamma]$ resumes all neuronal unit activations of one image i of the class j (cf. Equation 7).

Algorithm 1 Intermediate Model and Activation Histograms

for each category $j = \{1, \ldots, J\}$ **do**
  for each image $i = \{1, \ldots, I\}$ **do**
    Detect the points of interest of the image i.
    for each point $t = \{1, \ldots, P\}$ **do**
      for each descriptor $d = \{1, \ldots, D\}$ **do**
        Compute the local signature $x_d(t)$ with the descriptor d of the region of interest detected by t.
        The SOM^j_d, corresponding to the descriptor d and the class j, locates the BMU vector $m_c(t)$ with the signature $x_d(t)$.
        We compute the quantization error $eq_c(t)$ by:

$$eq_c(t) = ||m_c(t) - x_d(t)||, \quad (4)$$

        The neural activation histogram $H^j_{(i,d)}(t)$ of the map SOM^j_d is then updated by:

$$H^j_{(i,d)}(t)[c] = H^j_{(i,d)}(t-1)[c] + eq_c(t), \quad (5)$$

        where c is the BMU index of the map SOM^j_d.
      end for
    end for
  end for
end for

Algorithm 2 Final Model and Activation Histograms

for each category $j = \{1, \ldots, J\}$ **do**
  for each image $i = \{1, \ldots, I\}$ **do**
    for each histogramm $H^j_{(i,d)}$, $d = \{1, \ldots, D\}$ **do**
      The final map SOM^j, corresponding to the class j, locate the BMU vector $m_\gamma(t)$ with the signature $H^j_{(i,d)}$. (N.B. Here, $t = d$.)
      We compute the quantization error $eq_\gamma(t)$ by:

$$eq_\gamma(t) = ||m_\gamma(t) - H^j_{(i,d)}||, \quad (6)$$

      The neural activation histogram $H^j_{(i)}(t)$ of the map SOM^j is then updated by:

$$H^j_{(i)}(t)[\gamma] = H^j_{(i)}(t-1)[\gamma] + eq_\gamma(t), \quad (7)$$

      where γ is the BMU index of the map SOM^j.
    end for
    $H^j_{(i)}$ is concatenated with $H^{(j-1)}_{(i)}$ to form the image feature vector $H_{(i)}$.
  end for
end for

V. EXPERIMENTS

A. Natural Image Databases

We perform experiments on two different image databases in order to deal with different image classification issues.

- The first experiment tests an objectionable database. There are respectively 733 adult and 733 benign images in the training set. The test database is composed of 377 adult images and 467 benign images downloaded from the Internet. The benign class creation is a challenging issue. We decide to include: landscapes, flowers, animals, buildings, portraits and planes. We use this database to evaluate our strategy versus other fusion methods.
- The second dataset is the PASCAL 2005 recognition challenge [3]. The training set is composed of 684 images by merging the initial training and validation sets. The test set 1 proposes 689 images for four categories: motorbikes, bicycles, cars and people. This database was constructed with different well-known datasets except test set 2 (1072 images) which was obtained from the Internet. This database is a reference for image classification algorithms and the test set 2 is known to be more difficult, having no relationship with the learning images.

B. System Configuration

In all experiments, the SOM networks are configured using the following rules to ensure optimal performance in terms of accurate data representation: the learning steps T are 500 times the cell number U; the learning rate forms a monotone decreasing sequence: $\lambda(t) = \frac{T}{T+99t}$; the neighborhood width $\delta(t)$ decreases linearly with t from $\frac{\sqrt{2}}{2}U$ to 0.5.

C. Experimental Results

1) Objectionable Database: The first experiment examines the assumption of increasing classification rates with descriptor fusion. First of all, we classify the adult image database with each descriptor separately. The color, texture and shape descriptor are computed respectively thanks to the HCD, HTD and RFD descriptor. Here, we study the influence of the salient point detector by locating $2,000$ points with three extractors: Harris, contrast-based and wavelet-based. Moreover, we test the intermediate representation based classification with a SOM of size 20x20. $H_{(I,C)}$, $H_{(I,T)}$, $H_{(I,S)}$ are the activation histograms of the intermediate models of color, texture and shape, built from the HCD, HTD and RFD descriptors respectively. The final classification is performed by an SVM.

Approach	Harris	Contrast	Wavelet
$H_{(I,C)} \rightarrow$ SVM	91.11%	87.08%	**93.36%**
$H_{(I,T)} \rightarrow$ SVM	80.69%	79.50%	**82.11%**
$H_{(I,S)} \rightarrow$ SVM	91.82%	93.24%	**95.02%**

Table I
SVM CLASSIFICATION RATES FOR THE ADULT IMAGE DATABASE.

The results, shown in Table I, demonstrate slightly better performance when we use the wavelet-based detector. Nevertheless, the gain is minor compared to the other detectors, but it can be explained by the salient point location on sharp region boundaries that offers a good criteria for local image signature computation. Moreover, our intermediate image representation allows us to keep the discriminative and redundant information for each initial descriptor thanks to the activation histograms. The SVM classification then gives the following classification rates, respectively 93.36%, 82.11%, and 95.02%. It appears here that color and shape information is fundamental to filter these objectionable images.

Descriptor	KNN	MLP	RBF	SVM
Our approach 1	87.68%	91.23%	89.45%	**92.18%**
Our approach 2	94.91%	98.10%	95.50%	**98.46%**

Table II
CLASSIFICATION RATES FOR THE ADULT IMAGE DATABASE.

In the second part of this experiment, we test our fusion architecture using two strategies:

1) Our approach 1 uses one final SOM for all categories to join the three single activation histograms ($H_{(I,C)}$, $H_{(I,T)}$ and $H_{(I,S)}$); the final image feature vector is then the activation histogram of this SOM;
2) Our approach 2 combines several final SOMs, one for each category; the final image feature vector is then the concatenation of the activation histograms of each SOM.

Table II shows the gain obtained from our second approach, compared to the previous classifications. Indeed, the classification rate grows to 98.46% with this approach. Moreover, we can see that our fusion strategy performs better (98.46%) with a multi-model scheme than with a single SOM for all categories (92.18%). We can remark that one single SOM is not efficient enough to gather the full information, and that is why a multi-scheme strategy is preferable. Moreover, the second approach permits a competition between the different class models that increases the performances. In this experimental part, we similarly show that several final classifiers are applicable, but here the SVM seems to better deal with our final activation histograms better than a KNN, a MLP or an RBF classification. The performance decreasing of the last three classifiers are due to the curse of dimensionality. Moreover, the SVM classifier with a one-against-all strategy improves the good identification rates.

Approach	Method	CR
a priori 1	(HCD, HTD, RFD)	63.98%
a priori 2	$(H_{(I,C)}, H_{(I,T)}, H_{(I,S)})$	84.12%
a posteriori	Adaboost	96.21%
Our approach 2	Multi-SOM	**98.46%**

Table III
SVM CLASSIFICATION RATES (CR) WITH DIFFERENT FUSIONS.

In the third part of this experiment, we evaluate our fusion strategy versus other fusion methods. First, we compare our second approach to two *a priori* fusion methods before the final SVM classification:

- the *a priori 1* method corresponds to the input local descriptor concatenation: HCD, HTD and RFD;

- the *a priori 2* method is the concatenation of the intermediate representations: $H_{(I,C)}$, $H_{(I,T)}$ and $H_{(I,S)}$.

Secondly, we test the Adaboost method. This *a posteriori* method trains a series of classifiers and iteratively focuses on the hard training examples. The algorithm relies on continually changing the weights of its training examples so that those that are frequently misclassified get higher and higher weights. AdaBoost predicts one of the classes based on the sign of a linear combination of the weak classifiers trained at each step. Here, the weak classifiers are obtained with the $H_{(I,C)}$, $H_{(I,T)}$ and $H_{(I,S)}$ representation of Table I.

As shown in Table III, our second approach performs better than these *a priori* and *a posteriori* methods. These results highlight the importance of fusing different descriptors during the class model learning process.

2) PASCAL 2005 Database: The color, texture and shape descriptors are again computed thanks to the HCD, HTD and RFD descriptors respectively. The wavelet-based salient point detector is used to locate $2,000$ points and each SOM has a size of 20x20. All classification rates are obtained by a SVM.

The second experiment shows that our fusion strategy gives promising results on a visual object classification challenge. Indeed, the final SVM classification achieves 90.56%, overrating each intermediate classification, respectively 77.93%, 83.74% et 85.63%, for color, texture and shape description (cf. Table IV). This gain is performed on the four categories as shown on Table V and Figure 4, where the confusion matrix and the ROC (Receiver Operating Characteristic) curves are displayed for the test 1 database.

For the test 2 dataset, the methods are evaluated by measuring the Area Under the ROC Curve denoted AUC in the following. The comparison of classification results presented during the PASCAL 2005 recognition challenge with our results are presented in table VI. Finally, our method performs very well in comparison with other PASCAL 2005 methods. On this database, our AUC for people is the first one, for bicycles we are in the top two and the AUCs for motorbikes and cars are in fourth place. Thus, the proposed method increases the K-means based method presented by the University of Edinburgh [3], and the SOM-like representation of Helsinki University of Technology [3]. The method called "INRIA-Zhang" has got the best results for this database. This approach represents images as distributions of features extracted from a sparse set of keypoint locations and learns a SVM classifier with a kernel based on an effective measure for comparing distributions. Their vocabulary histograms are built from a direct clustering in opposition to our neural learning process which synthesizes the data in an unsupervised mode.

Database	Intermediate Model	CR	Final Model	CR
PASCAL 2005 Test1	$H_{(I,C)}$	77.93%	$H_{(I)}$	$\mathbf{90.56\%}$
	$H_{(I,T)}$	83.74%		
	$H_{(I,S)}$	85.63%		

Table IV
SVM CLASSIFICATION RATES (CR): TEST 1 PASCAL 2005

$\rightarrow$	A	B	C	D	AUC (Test 1)
A = bike	93	11	5	5	$0.9551(5/16)$
B = car	6	248	1	8	$0.9867(3/18)$
C = motorbike	4	3	208	1	$0.9903(5/18)$
D = people	4	9	7	64	$0.9508(4/16)$

Table V
CONFUSION MATRIX: TEST 1 PASCAL 2005

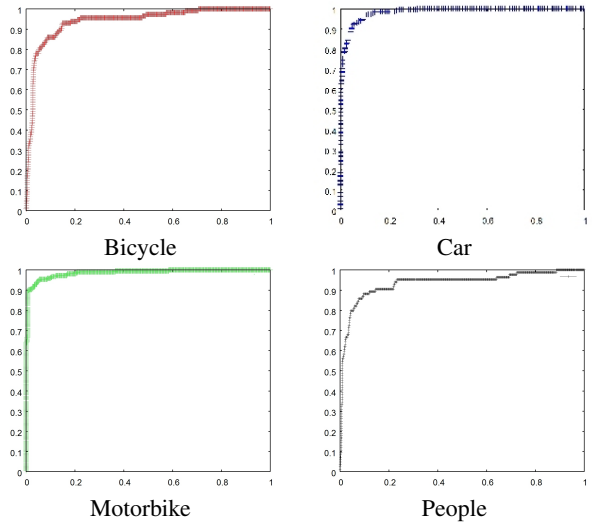

Figure 4. ROC curves: Test 1 PASCAL 2005

Method	Motorbike	Bicycle	People	Car
Aachen: ms-2048-histo	0.825	0.724	0.721	0.767
Aachen: n1st-1024	0.829	0.729	0.739	0.780
Darmstadt: ISM	0.706	•	•	0.572
Darmstadt: ISMSVM	0.716	•	•	0.683
Edinburgh: bof	0.710	0.606	0.552	0.655
HUT: final1	0.666	0.567	0.650	0.709
HUT: final2	0.693	0.647	0.661	0.740
HUT: final3	0.637	0.546	0.618	0.694
HUT: final4	0.675	0.645	0.630	0.744
INRIA-Zhang	**0.865**	**0.813**	**0.798**	**0.802**
MPITuebingen	0.765	0.654	0.655	0.717
Our approach 2	**0.812**	**0.736**	**0.811**	**0.752**
Rank	4/12	2/12	1/12	4/12

Table VI
AUC VALUES: TEST 2 PASCAL 2005 [3]

The figures 5 and 6 show respectively samples of good and bad classification with our strategy. It appears that there exists a large variety of color, texture and shape information to define one visual concept, a bicycle for instance. Nevertheless, the intermediate model seems to focus on the more redundant information and the final model combines each description to improve the classification results. The classification error of 9.04% may be explained by a slight confusion between bicycle and motorbike that shares many properties.
When two different objects interfere in the same image (cf. 3^{rd} image of motorbike on Figure 6), the classifier selects the more probable as a class winner. It would be interesting to focus the category learning process on a visual object, and to detect them during the classification process, to deal with such ambiguities.

VI. Conclusions and Future Studies

In this paper, we have presented a new system to classify visual concepts, jointly using several pieces of descriptor information contained in each region of interest. Based on the main properties of SOM, our scheme gives very promising results to identify the image class. The multi-modal architecture joins the color, texture and shape features to model the two categories in the context of adult image filtering better. We plan to experiment our fusion method on other issues such as face recognition.

The advantages of the proposed method are the following: the possible use of whatever image descriptor (color, texture, shape, etc.); several approaches are available: global, local, region-based; the number of descriptors is only limited by the classifier capacity; the final image feature vector is compatible with all classifiers; a feature selection; a heterogeneous feature fusion.

References

[1] Bres S. and Jolion J.-M. Detection of Interest Points for Image Indexation. In *VISUAL*, pages 427–434. Springer-Verlag, 1999.

[2] Broomhead D.S. and Lowe D. Multivariable Functional Interpolation and Adaptive Networks. *Complex Systems*, 2:321–355, 1988.

[3] Everingham M., Zisserman A., Williams C., Van Gool L., Allan M., Bishop C., Chapelle O., Dalal N., Deselaers T., Dorko G., Duffner S., Eichhorn J., Farquhar J., Fritz M., Garcia C., Griffiths T., Jurie F., Keysers D., Koskela M., Laaksonen J., Larlus D., Leibe B., Meng H., Ney H., Schiele B., Schmid C., Seemann E., Shawe-Taylor J., Storkey A., Szedmak S., Triggs B., Ulusoy I., Viitaniemi V., and Zhang J. The 2005 PASCAL Visual Object Classes Challenge. In *The First PASCAL Challenges Workshop*, 2006.

[4] Freund Y. and Schapire R. Decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[5] Harris C. and Stephens M. A Combined Corner and Edge Detector. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[6] Kohonen T. *Self-Organizing Maps*. Springer, 2001.

[7] 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*, 31:73–94, 2006.

[8] Lefebvre G. and Garcia C. Facial biometry by stimulating salient singularity masks. *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 511–516, 5-7 Sept. 2007.

[9] Lefebvre G., Laurent C., Ros J., and Garcia C. Supervised image classification by SOM activity map comparison. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 728– 731, 2006.

[10] Mallat S. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE TPAMI*, 11(7):674–693, 1989.

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

[12] Ros J. and Laurent C. Description of local singularities for image registration. In *ICPR*, volume 4, pages 61–64, 2006.

[13] Rosenblatt F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65:386–408, 1958.

[14] Shapiro J.M. Embedded Image Coding Using Zerotrees of Wavelet Coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, 1993.

[15] Spyrou E., Le Borgne H., Mailis T., Cooke E., Avrithis Y., and O'Connor. Fusing MPEG-7 Visual Descriptors for Image Classification. In *ICANN*, pages 847–852, 2005.

[16] Swain M.J. and Ballard D.H. Color Indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[17] Vapnik V. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[18] Zhang W., Shan S., Gao W., Chang Y., Cao B., and Yang P. Information fusion in Face Identification. In *CVPR*, pages 950–953, 2004.

[19] Zheng H., Lefebvre G., and Laurent C. Fast-learning adaptive-subspace self-organizing map : An application to saliency-based invariant image feature construction. *Neural Networks, IEEE Transactions on*, accepted for publication in may 2008.

Figure 5. Good classification: Test 2 PASCAL 2005

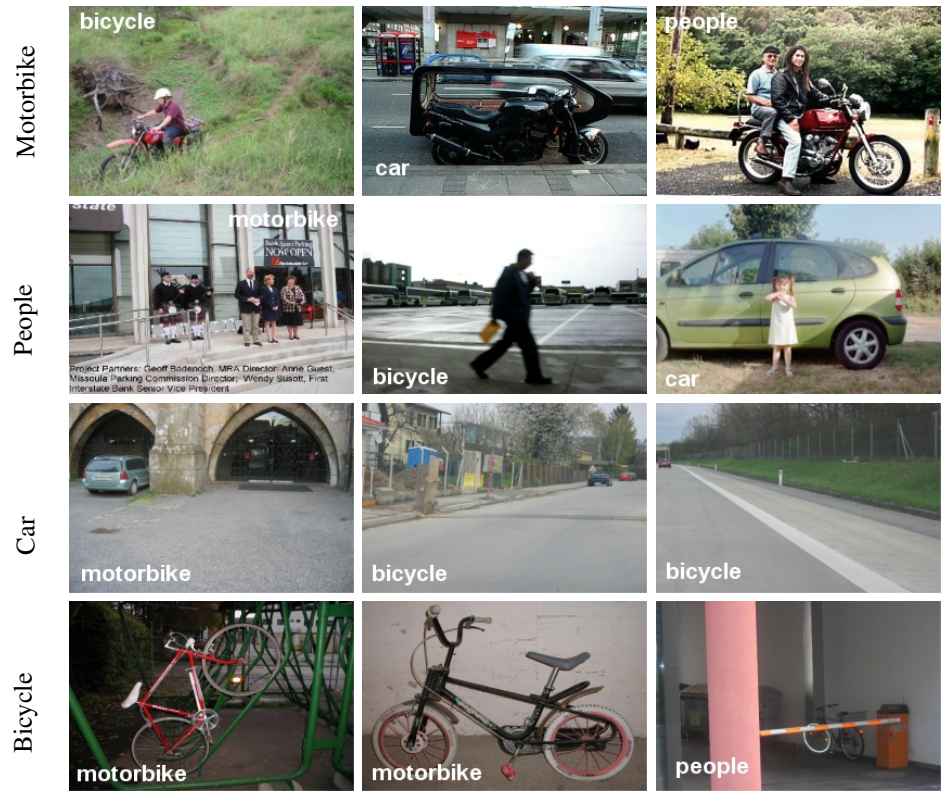

Figure 6. Bad classification: Test 2 PASCAL 2005