Image annotation : which approach for realistic databases ?

Nicolas Hervé
INRIA Rocquencourt
78153 Le Chesnay cedex
France
nicolas.herve@inria.fr

Nozha Boujemaa
INRIA Rocquencourt
78153 Le Chesnay cedex
France
nozha.boujemaa@inria.fr

ABSTRACT

This paper describes an efficient approach to image annotation. It ranked first on the recent scene categorization track of the ImagEVAL[1] benchmark. We show how homogeneous global image descriptors combined with a pool of Support Vector Machines achieve very good results. We also used this approach on several well known object recognition databases to emphasize two main aspects of this research domain : the importance of contextual information in object recognition and the unsuitability of many standard databases for this task.

Categories and Subject Descriptors

I.4.10 [**Image Processing and Computer Vision**]: Image Representation; I.5.4 [**Pattern Recognition**]: Applications—*Computer vision*

General Terms

Experimentation, Performance

Keywords

Scene Categorization, Image annotation, Object Recognition, Global Descriptor, Support Vector Machine

1. INTRODUCTION

Automatic or semi-automatic image annotation is a very active research field. Many professional end-users are asking for such technologies to help them in their work. They currently have to deal with the exponential increase in the production of new pictures while undertaking the digitization of archived pictures. The competition between photo agencies has reached such a level that they should not only provide nice pictures but also very well documented ones. Observing archivists at work is a very enriching experience, and it reveals the technological gap between problems that are considered to have been virtually solved in our community and the tools that are used in such companies. For example, it is quite surprising to see that information such as the orientation of the picture and black and white versus color are still entered manually, even when EXIF metadata are available. Although we are only in the early stages of automatic object recognition, the scene categorization problem now has technologies that are mature enough to be implemented.

Several research communities are involved in the automatic images annotation problem. In addition to the computer vision community, many researchers from the machine learning or natural language processing fields have proposed new solutions. Being at a crossroads, we face an abundance of task definitions and corresponding evaluation strategies. With such diversity, knowing what really works and what doesn't is an almost impossible task for newcomers. Moreover, and paradoxically, the availability of large-scale image databases for research purposes is compromised by the uncertainty of copyright ownership. This leads researchers to work on a small number of commonly available databases. Unfortunately, these databases have huge drawbacks : they are quite different from real databases and, more importantly, they tend to direct the research orientations away from the end-users real needs. Whereas their usefulness was obvious in the early stages, we should now move toward considering more realistic data. Benchmark campaigns such as ImagEVAL are an attempt to unify datasets, evaluation strategies, task definitions and end-user needs.

Scene categorization has been studied for a long time now with, notably, the classical *indoor vs. outdoor* task. Numerous approches have been tested, focusing either on the low-level description of the pictures or on the learning strategies involved. In the latter case, building complex learning models has too often been presented as a way to bridge the semantic gap. We want to emphasize the fact that using off-the-shelf visual descriptors that are not appropriate for a given task or that are not able to capture all the visual information of pictures (including contextual information) is a problem that should be addressed before considering the use of new learning strategies. We call this "the numerical gap", meaning that the information is in the pictures but it is not extracted due to the weakness of the descriptor.

We propose to challenge our global descriptors in the scene categorization problem. We will show how efficient descriptors could perform very well using simple SVMs, in terms of both relevance and time. This approach ranked first on the

[1]http://www.imageval.org

CIVR'07, July 9–11, 2007, Amsterdam, The Netherlands.

ImagEVAL scene categorization track. Furthermore, we will discuss the contributions this method may have on object recognition tasks.

This paper is organized as follows. We first present the ImagEVAL benchmark and describe in detail the scene categorization task. Then we describe our low-level descriptors, our approach for this benchmark and our results. Finally, we present some evaluations achieved on other datasets using the same approach, including many object recognition databases. We then draw some conclusions on the relevance of global descriptors for image annotation.

2. IMAGEVAL BENCHMARK

ImagEVAL is a new image retrieval benchmark initiative that was launched in France in 2006. This campaign has been driven by the need for shared evaluation in our community. Despite the fact that some tracks dedicated to image retrieval tend to appear in well established evaluation campaigns (TREC, CLEF, etc), the task specifications are rarely dedicated to the image domain. They are often linked to the main objective of the campaign, leading typically to cross-media retrieval tasks in text evaluation campaigns. ImagEVAL is fully focused on the image retrieval domain. A second interesting aspect that distinguishes ImagEVAL is that its specification and organization were established by both a research team and professional archivists [17]. The task definitions were discussed in order to address the real problems that photo agencies face. The images on which the evaluation were conducted were also professional ones. They were selected by professionals, allowing the groundtruth to be established in a confident way : as expected by the users, not by the researchers [22]. We are therefore close to real-life scenarios, with challenging image collections. Several French teams participated in the evaluation, as well as private companies. The benchmark campaign was also open to European teams. ImagEVAL has five main tracks : transformed images, web based image retrieval, text area detection, object recognition and attribute extraction. We will focus on this last track, explain the method we used and present our results.

2.1 Attribute extraction task

The purpose of this task is to allow image classification. Two kinds of semantics are targeted : the nature of the image (artistic representation, color photograph, black and white photograph, colored black and white photograph) and the context of the image (indoor/outdoor, day/night, natural/urban).

Figure 1 shows the organization of these concepts as they are presented in the task description.

A database containing 5416 images was provided for training purposes. The typical size of these pictures was about 1000x700 pixels. A groundtruth file was also provided. The binary partition of the photograph contexts was not as clear as it should have been. There were photographs in the database that were more related to dawn or twilight than day or night. The same concept ambiguity also appeared for photographs of natural and urban scenes and even for *indoor/outdoor* classification. But these ambiguities reflect the real-life cases and should be dealt with. The only constraint for the archivists that annotate these pictures was to provide all the semantics to reach the leaves of the concepts tree.

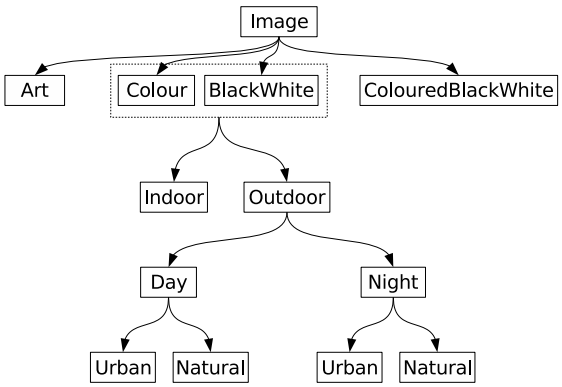

Figure 1: Semantics as presented in the task description

Semantics	Nb. pictures
ART	429
BlackWhite, Indoor	498
BlackWhite, Outdoor, Day, NaturalScene	159
BlackWhite, Outdoor, Day, UrbanScene	449
BlackWhite, Outdoor, Night, UrbanScene	16
BlackWhite, Outdoor, Night, NaturalScene	0
Color, Indoor	1129
Color, Outdoor, Day, NaturalScene	946
Color, Outdoor, Day, UrbanScene	1092
Color, Outdoor, Night, NaturalScene	3
Color, Outdoor, Night, UrbanScene	368
ColoredBlackWhite	327

Table 1: Task 5 learning database

Table 1 summarizes the distribution of the pictures in the learning database. It can be seen that these data are hugely unbalanced, but this simply reflects the natural distribution of these semantics in real databases. Finding black and white photographs taken in a natural environment by night is actually quite rare. Figure 2 shows some examples from this learning database. The final evaluation database contains 23572 images.

For the evaluation, the queries were paths of the semantic tree (e.g. *Art* or *Color/Indoor*). For each query, the first 5000 images should be retrieved. As an image retrieval task, the measure used to evaluate the algorithms emphasizes retrieving pertinent documents earlier. The MAP (Mean Average Precision) is used. It differs from the measures used in classification tasks. Therefore, the confidence we have in a semantic prediction is important to rank the results. As usual, the precision is defined as :

$$P = \frac{\text{number of relevant documents retrieved}}{\text{number of documents retrieved}} \quad (1)$$

The Average Precision is the average of the precision after each relevant document is retrieved. For a query Q, with r the rank, N the number retrieved, $rel()$ a binary function on the relevance of a given rank, and $P()$ precision at a given cut-off rank, we have :

$$AP_Q = \frac{\sum_{r=1}^{N}(P(r) \times \text{rel}(r))}{\text{number of relevant documents}} \quad (2)$$

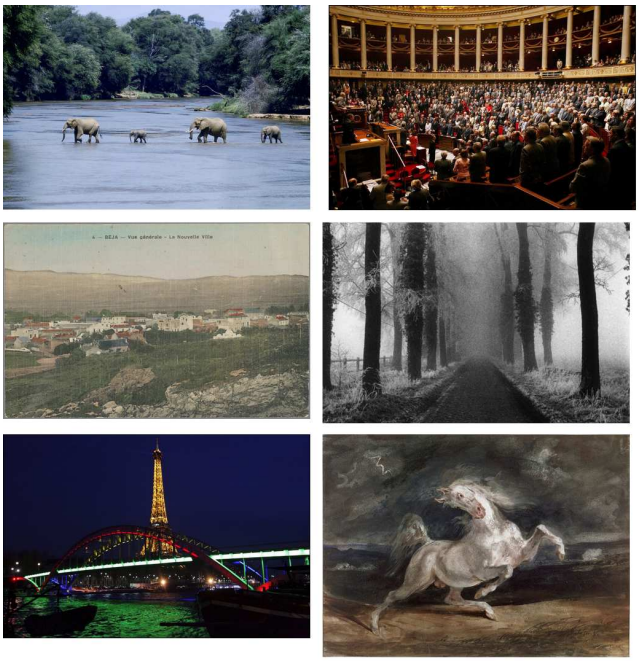

Figure 2: Task 5 learning database samples

The MAP is the mean of the AP over a set of queries. It is computed by the Treceval software v7.3.

2.2 Previous work

This task is close to what is generally presented in the literature as scene categorization. The classical *indoor vs. outdoor* problem has been studied for the past ten years. The *city vs. landscape* is also a common categorization problem addressed in numerous papers. Several databases have been used and a wide range of approaches has been explored. Among the first attempts, Szummer [29] extracted color and texture descriptors on rectangular regions obtained on a fix grid over the pictures. A two-stage approach was used to separate *indoor* and *outdoor* pictures. The first layer used KNN classifiers, the second layer was based on a majority vote. The tests were made on a Kodak database of 1300 pictures. Serrano [26] used a similar approach on the same database with SVMs for both layers. He also included semantic cues [27] in a bayesian network to replace the second layer. Concepts like grass, sky or clouds are detected. Vailaya [30] [31] worked on the global pictures. He first used global descriptors and KNN classifiers. The method was refined with the introduction of a hierarchy of binary bayesian classifiers. Maron [16] introduced a Multiple Instance Learning strategy based on small blobs of 2x2 pixels. The Diverse Density algorithm was used to create the model. The evaluation was made on the Corel database. Dugué [13] used the Local Dominant Orientations extracted from the power spectrum using a scale space to find man-made structures in the pictures. Oliva introduced the Spatial Envelope [20] [21] based on perceptual dimensions like naturalness, openness or expansion. Some papers present approaches in the context of the aceMedia European project [19] [28]. By using multiple MPEG-7 low-level descriptors, these approaches face the problem of fusing non-homogeneous descriptions of the pictures. Zhang [37] studied boosting approaches compared to SVM approaches on the Corel database. Recently Cutzu [8] studied several new low-level descriptors to distinguish paintings from photographs. This is to our knowledge the only work done on the subject. A comparison between these approaches is quite difficult as the databases used are different and rarely publicly available, as are the implementations of the algorithms. Sometimes, even the metrics used are different (classification rates with or without the learning examples kept, precision-recall curves, ROC equal error rate, etc). Generally, the best rates reported for the *indoor/outdoor* classification task are around 90%. The processing times are almost never reported.

3. OUR APPROACH

3.1 Visual content descriptors

The quality of the low-level descriptors used in any CBIR or automatic image annotation system is dominant over any other component. As the image signatures are the raw material on which all the algorithms of our domain are based to express visual similarity or to find some semantic concept, it is of a great importance that they can be trusted. Global descriptors have long been used to characterize the visual aspect of images. Designing visual descriptors for specific databases, with *a priori* knowledge that can be encapsulated in the descriptors definition is already a difficult task. Finding good descriptors definition for generic databases is challenging. Colors, textures and shapes have been identified as the main low-level aspects that can be characterized in images. We believe that good descriptors should be faithful to the image content and try to keep a low dimensionality to avoid the 'curse of dimensionality' problems that may occur. The speed of extraction is important. More critically, the comparison processing time between signatures needs to be as quick as possible. An other important point is that we focused on designing structurally homogeneous descriptors. Concretely, they are all histograms, and therefore can be used simultaneously, with the same distance functions or the same kernels. This possible combination of our descriptors is one of the powerful points on which we will come back later. Most of these descriptors already existed in our IKONA CBIR engine [3]. They have already been widely tested in visual similarity search and relevance feedback scenarios. We will present new results on scene categorization and object recognition on the ImagEVAL benchmark database and on some common publicly available research databases.

For the color description, we used a standard HSV histogram (*hsv*, 120 bins). We also use weighted color histograms [32] which make it possible combine both color and structure information in a single representation. It is well known that usual color histograms do not keep any spatial information about the pixels. But it is also known that pixels with the same color do not have an equal visual importance depending on their localization in the image. Thus came the idea to include some local activity information, measuring local uniformity or non-uniformity, in color histograms. Shape and color are merged by weighting pixel colors with the Laplacian (*lapl*, 216 bins). Texture and color are merged by weighting colors with a probability measure (*prob*, 216 bins). Texture information is gathered by a Fourier histogram [11]. After obtaining the 2D Fourier

transform of an image, two histograms are computed in the complex plane. They represent two types of distributions of the energy. The first histogram is computed with a disks partition, the second uses a wedges partition. Both have an equal importance in the final signature. We slightly adapted this descriptor for ImagEVAL in order to have a constant radius increase rather than a constant disk surface increase for the disk partition version (*four*, 64 bins). Shapes are characterized with a histogram inspired by the Hough transform [11]. For each pixel, the gradient orientation and the projection size of the pixel vector onto the tangent vector to the local edge are used to build a 2D histogram (*hou*, 49 bins).

We also implemented a generic shape descriptor called 'Local Edge Orientation Histogram' (*leoh*). The basic idea for the definition of this descriptor has been driven by the ImagEVAL scene categorization task where a distinction should be made between natural and urban pictures. It appears that human structures in such pictures were of importance but did not necessarily take up a huge proportion of the picture. The classical Edge Orientation Histogram [14] could be used to detect the characteristic horizontal and vertical structures induced by the presence of human buildings. But if this building is too small in the picture, its presence will be swamped by the surrounding noise. We follow the idea expressed in [25] to extend blob histograms to local orientations and adapted it to contours gradient only. Local Edge Orientation Histogram has the advantage of containing both local and global information. We quantized the orientations using 8 bins. We used 4 bins for the relative proportions. We thus have a 32 bins signature. This descriptor has been compared to a 32 bins classical EOH on the ImagEVAL dataset and it leads to better results (presented in section 3.3).

3.2 Learning strategy

We choose an early fusion approach with a soft-margin Support Vector Machine as the learning algorithm. Using generic global descriptors, we have no specific idea on which characteristic is important to discriminate each concept (apart for the Local Edge Orientation Histogram where we inject a little *a priori* knowledge by focusing on horizontal and vertical edges). We believe that it is the responsibility of the learning algorithm to find which information is important to build a model for these concepts. As we may have some correlation between low-level features from different types (color, shape and texture), we think the early fusion approach is appropriate and the SVM is powerful enough to select the pertinent features. Thus, by staying away from specific solutions, both for the visual descriptors and the learning strategy, our system remains generic and it is able to adapt to any new content and/or concept. By concatenating the six global descriptors described in section 3.1, we obtain a 697-bin signature per picture.

As shown in figure 3, the concept tree can also be represented as a complete partition. It can be seen as a flattening of the tree. This raises the question of knowing which type of SVM strategy should be chosen. We have two possibilities. We can choose to learn each concept separately with one-against-all approaches. In this case, we have one model per concept. The second option is to consider the concepts of the semantic tree in extension. Each leaf of the tree is then a single concept. Both strategies have been tested and provide similar results. This is perfectly understandable as

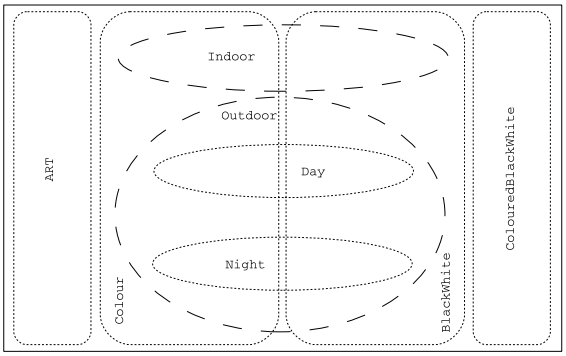

Figure 3: Other representation of the concepts

the SVM focuses on the boundaries between concepts in the feature space. The only difference is that in the first case the same boundaries will be learnt several times and appear in different models. This will lead, globally, to heavier models (more support vectors, implying longer prediction time). But this approach is more flexible as adding new concepts is easier.

We have tested four kernels (see table 2). The non-parametric kernels have the huge advantage of quickening the learning process as the parameter optimization phase is reduced to finding the optimal regularization factor C of our soft margin SVM. This is done by 5-fold cross-validation on the learning database. More information on these kernels may be found

Kernel	Parametric	
Laplace	X	$k(x,y) = e^{-\sigma \sum_i \lvert x_i - y_i \rvert}$
RBF	X	$k(x,y) = e^{-\sigma \sum_i (x_i - y_i)^2}$
Triangular		$k(x,y) = \sum_i \lvert x_i - y_i \rvert$
GHI		$k(x,y) = \sum_i min(\lvert x_i \rvert, \lvert y_i \rvert)$

Table 2: Tested kernels

in [2].

As all our descriptors are histograms, we guarantee, by construction, that the sum of all the bins is equal to one for each signature. Thus, initially, our six descriptors have the same relative importance in the kernel computation. It is also possible to apply some preprocesses on the full vectors that will break this equity but may help the SVM to discriminate between concepts. We have tested four different preprocesses [2] :

- none : no preprocess
- scale : each bin is scaled between 0 and 1 for the full learning database
- normalize : each bin is normalized according to its standard deviation on the full learning database
- pow : each bin is raised to a given power (typically 0.25)

Once the models have been computed for each concept, they can be used to predict the semantics on new pictures. Obtaining only the predicted concepts is not enough. We also need confidence levels regarding the predictions in order

to rank the results. Some research has been done on SVMs to have probabilistic outputs [23]. This is quite convenient as it allows a comparison between the confidence levels of different SVMs. Despite this, we choose the simplest approach that assimilates the score of the SVM decision function in a confidence level, following the intuitive idea that being far from the decision boundary is equivalent to being less ambiguous for the concept. As all our models are based on the same feature space, we found this approach appropriate to compare and combine the predictions of the different concepts. We did not deal with the problem of a learning database that doesn't fully cover the possible feature space, leading to potentially biased models.

We also introduce a specific classifier that allows us to combine the outputs of several earlier classifiers. This is useful in order to associate a global confidence level to complex concepts (eg. *Color/Indoor*). For each combination of concepts, we keep the minimum of the confidence levels.

All the preprocesses, optimization, learning, prediction and querying phases have been implemented in C++ in our CBIR engine using a home-patched version of LibSVM [4].

3.3 Evaluation results

Six different teams participated in this task. Each team could submit up to five different runs. The teams were allowed to provide runs with additional pictures for the learning database, but none did. There was a total of 13 queries : (1) Art; (2) ColoredBlackWhite; (3) BlackWhite / Indoor ; (4) BlackWhite / Outdoor; (5) Color / Indoor ; (6) Color / Outdoor ; (7) BlackWhite / Outdoor / Night ; (8) BlackWhite / Outdoor / Day / Urban ; (9) BlackWhite / Outdoor / Day / Natural ; (10) Color / Outdoor / Day / Urban ; (11) Color / Outdoor / Day / Natural ; (12) Color / Outdoor / Night / Urban , (13) Color / Outdoor / Night / Natural. Each team had to return the 5000 first pictures for each query.

We submitted 5 runs, corresponding to the different options presented in the previous section.

Run	Options
imedia01	old version used for the blank tests
imedia02	GHI kernel, pow 0.25 preprocess
imedia03	triangular kernel, scale preprocess
imedia04	laplace kernel, scale preprocess
imedia05	extension concepts, triangular kernel, scale

Table 3: Task 5 runs

The full results are available on the campaign web site. We provide here the MAP of all the runs.

Run	MAP	Run	MAP
imedia04	0.6784	etis01	0.4912
imedia03	0.6556	anonymous	0.4907
imedia05	0.6532	anonymous	0.4931
imedia02	0.6529	anonymous	0.3676
imedia01	0.5979	anonymous	0.3141
cea01	0.5771	anonymous	0.1985

Table 4: Task 5 MAP

We can notice that our first run is 0.1 point better than the second team. Then came the three runs using non-parametric kernels. It is interesting to see that using a parametric kernel, leading to a time intensive parameter tuning phase, only brings a small improvement in the results. Triangular and GHI kernels obtain the same performances. Further tests, made after the evaluation campaign when the groundtruth was available, also indicate that the different preprocesses bring a small but not significative increase of 0.015 point for the MAP. As explained earlier, there is almost no difference between runs 3 and 5. Using the extension concepts leads to a quicker model. The total number of support vectors is 11252 for run 3 (learning time is 372 sec, prediction time is 0.1 sec per picture), and 6595 for run 5 (learning time is 176 sec, prediction time is 0.05 sec per picture). Run imedia01 can be forgotten as it corresponds to hypotheses that were true for the blank tests but which no longer hold. All the processing was done on a Pentium 4, 2.8GHz, 2 Go, Linux. The low-level features extraction time is 6 sec per picture. We can notice in the detailed results [17] that this approach is one of the quickest.

If we except the categories where only very few training examples were available, the MAP is above 0.75, which represent highly satisfactory results.

Using run imedia03 as a baseline, we made some more tests with the same options to see the individual contributions of each low-level descriptor to the global result. We used each descriptor alone and ran the same procedure.

Q	run3	hsv	prob	lapl	four	hou	leoh	eoh
1	.93	.60	.60	.55	.49	.58	.30	.17
2	.85	.53	.50	.52	.07	.24	.06	.02
3	.86	.69	.73	.59	.14	.26	.13	.06
4	.81	.57	.63	.46	.06	.17	.06	.06
5	.74	.49	.51	.49	.24	.35	.38	.11
6	.55	.50	.48	.45	.28	.39	.31	.21
7	.07	.02	.04	.02	.00	.00	.00	.00
8	.79	.56	.61	.41	.06	.14	.08	.08
9	.57	.10	.13	.08	.02	.05	.02	.02
10	.74	.44	.49	.48	.27	.30	.29	.12
11	.87	.54	.57	.54	.52	.58	.67	.48
12	.62	.44	.44	.37	.04	.06	.06	.02
13	.13	.02	.01	.01	.01	.01	.01	.00
MAP	.66	.42	.44	.38	.17	.24	.18	.10

Table 5: AP - each descriptor alone

Table 5 shows the AP results in detail. The color descriptors achieve a good overall performance alone. The importance of color for such tasks has always been known. Furthermore, as *BlackWhite/Color* distinction is at the root of the concepts tree, this importance is increased in our context. We can also notice that *leoh* is almost twice as good as a simple *eoh* for the same signature size. As we have several descriptors for each type of characteristic, they are partially redundant. This overlap in the information extracted from the pictures is useful as it allows a wider area of features to be covered and lets the SVM select the most discriminating ones for a given concept to learn. In order to study the contribution of each descriptor mode deeply, we now run the same procedure removing each one of them individually. We report in table 6 the loss in the AP measure compared to run3 for the most relevant queries.

Actually, removing a single descriptor hardly affects the global results. In other words, it means that SVMs are able

Q	hsv	prob	lapl	four	hou	leoh
1	.01	.00	.01	.05	.03	.01
2	.02	.01	.01	.08	.01	.01
3	.01	.00	.02	.00	.01	.04
4	.01	.01	.00	.01	.01	.06
5	.00	.00	.04	.00	.01	.04
6	.00	.00	.00	.00	.01	.01
8	.01	.02	.01	.01	.01	.05
9	.03	.02	.01	.05	-.01	.09
10	.01	.03	.02	.01	.01	.04
11	.00	.01	.00	.04	.02	.05
12	.01	.02	.01	.02	.02	.04
MAP	.01	.01	.01	.02	.01	.04

Table 6: Loss in AP - removing each descriptor

to manage a surplus of redundant information. It can be noticed that certain descriptors are important for certain queries. For instance, *leoh* provides a huge contribution for query 9, which distinguishes between Natural and Urban scenes. Another interesting result is that *four* is important to identify the colored black and white pictures (almost all of them came from an old postcards archive) when combined with the other descriptors (0.08 loss in AP) although alone it has very poor performances on this concept (0.07 AP). It illustrates the great complementarity of our descriptors.

Based on this information, we have tested a new combination of 3 selected descriptors (*prob*, *four* and *leoh* - 312 bins) that produces almost equivalent results (MAP 0.63) with twice half as much information extracted from the pictures, leading to a method that is twice as fast.

4. OTHER EVALUATION DATABASES

What information could we obtain by applying the method described in the previous section to object recognition databases ? Finding objects in pictures involves using local descriptors obtained through segmented regions, points of interest or dense grid portions. Whatever the method, we need to focus on parts of the picture. But contextual information is also rather important to detect objects. Some descriptors try to integrate this local and global information [1]. We do not argue that object recognition problem is solved by using global descriptors. We do think, however, it is important to measure the suitability of the databases used by researchers to evaluate their methods. A global approach is a good way to obtain baseline results that provide pieces of information on the difficulty of the task and the database. In [24] these problems are also described and solutions to acquire new datasets are proposed. For the following databases, we used the same experimental setups as described by the authors of these studies. We used our reduced set of descriptors and SVMs with a triangular kernel. When necessary, we used a gray level version of the descriptors in order to make a comparison with the other methods possible.

4.1 Corel2000

The Corel dataset is probably the most widely-used in image retrieval and categorization. As early as 2002, papers explained the simplicity of this database [18] [33]. Some experiments use 2000 pictures from this Corel collection, divided into 20 categories. The database is randomly split into learning and testing sets, with 50 pictures per category. The operation is done 5 times, and the good categorization ratio is measured. The results clearly confirm that this database

Algorithm	Result
Our approach - 5 desc.	83.7
Our approach - *HSV* only	71.6
Chen - MILES [5]	68.7
Chen - DD-SVM [6]	67.5
Csurka [7]	52.3

Table 7: Results on Corel2000

is too simple. Even with a single *HSV color histogram*, the results are better than tested local approaches.

4.2 ImagEVAL object recognition task

The fourth task of the ImagEVAL benchmark was dedicated to object recognition. Ten classes of objects were proposed (armored vehicle, car, cow, Eiffel tower, minaret and mosque, plane, road signs, sunglasses, tree and US flag). The final database contains 14000 pictures, either color or gray level. Some contain objects from different classes, and 5000 contain no object. The learning database is composed of crops of these objects. This database is one of the most challenging available. The objects are in very different poses, contexts and sizes. As an example, one can see in figure 4 some pictures containing a US flag.

Figure 4: Task 4 - American flag

Unfortunately, only three teams participated in this task (due to the high complexity of this real life database ?). The results are quite bad and reflect the improvements that need to be made in the local approaches for object recognition, especially for realistic pictures. We tried our global method on run imedia01, learning on the crops and predicting on the global pictures. The MAP obtained is not far from our best run. A detailed review of the results indicates that the global approach obtains good results on the *plane* and *tree* categories, which is quite coherent because both object classes involve a somewhat specific context that is captured

Run	MAP	Run	MAP
imedia05	0.2242	imedia03	0.1545
imedia04	0.2111	anonymous	0.1506
etis01	0.1974	cea01	0.1493
imedia01	0.1777	anonymous	0.14
imedia02	0.1733		

Table 8: Task 4 MAP

globally. The context is an important information element for some object categories.

4.3 Caltech4

This database contains four object classes. For each of them, background pictures are also available. The purpose is to separate pictures containing the object from the background pictures. It is a classification task *object vs. background*. We used the same training and testing sets as in [12]. We used the gray level descriptors : *lapl*, *prob*, *four* and *leoh* for a total of 84 dimensions. We obtained equiv-

Algorithm	Plane	Car	Face	Motorbike
Our approach	99.2	100	98.6	98.8
Chen [5]	98.0	94.5	99.5	96.7
Zhang J. [36]	98.8	98.3	100	98.5
Willamowski [34]	97.1	98.6	99.3	98.0
Fergus [12]	90.2	90.3	96.4	92.5

Table 9: Results on Caltech4

alent results with our global approach. Classification rates that reach almost 100% with a global approach tend to prove that this database is clearly not difficult enough for object recognition algorithms.

4.4 Xerox7

This database contains 1776 pictures from 7 classes. As in [34], we used a multi-class classification on a 10-fold cross-validation. The average accuracy is reported. We used the gray level descriptors. Our results are really close to the

Algorithm	Result
Our approach	92.5
Zhang J. [36]	94.3
Willamowski [34]	82.0

Table 10: Results on Xerox7

best ones published. As for the previous one, this database is not well-suited to this task.

4.5 Pascal VOC2005

We can find the complete description of this challenge and the two datasets in [9]. The first dataset is said to be quite easy while the second one is challenging. The Equal Error Rate on the ROC curve is used to compare methods. We only report the best published results within all.

VOC2005-1 is too simple, but when we move to a more difficult dataset, containing pictures collected with Google, the local approaches have a signifiant contribution.

Algorithm	Bike	Car	Motorbike	People
Our approach	88.7	92.2	95.8	86.9
Best score in [9]	93.0	96.1	97.7	91.7

Table 11: Results on VOC2005-1

Algorithm	Bike	Car	Motorbike	People
Our approach	57.9	66.3	64.8	69.2
Zhang J. [36]	68.1	74.1	79.7	75.3

Table 12: Results on VOC2005-2

4.6 Caltech101

A 101 object classes database, plus one background class that is generally not used [10]. The objects are always centered in the pictures. There are between 31 and 800 images per category, with huge drawbacks on some of them : two face classes, artificial 45 rotation of some classes, etc. We found two main experimentation protocols, with 15 or 30 images per class for the training database.

Algorithm	30 im./class	15 im./class
Our approach	39.6	32.7
Zhang H. [35]	66.23	59.08
Lazebnick [15]	64.6	56.4

Table 13: Results on Caltech101

5. CONCLUSIONS

We obtained good results on a scene categorization task with our approach involving global descriptors and pools of SVMs. We believe that the techniques used in this context are now mature enough to be implemented in real applications and could help end-users. For the object recognition task, databases such as Corel, Caltech4, Xerox7 and VOC2005-1 should now clearly be abandoned for testing local approaches as simple global methods achieve equivalent accuracy. This implies that these databases are not problematic ones. Caltech101 is a particular case. The use of local approaches is needed and their benefits have been clearly demonstrated. But the pictures are far from what may be found in real digital libraries. Actually, we doubt whether the same problem is being addressed in this database and in real cases. We believe that the use of realistic databases such as ImagEVAL-4 should now be standard. We will face very challenging problems that meet end-users scenarios. Processing time is also an important criteria that should be considered as professional databases commonly have millions of photographs and scalability problems will be a key issue. The importance of contextual visual information has also been shown for this type of database.

6. ACKNOWLEDGMENTS

We would particularly like to thank Itheri Yahiaoui who suggested implementing the LEOH descriptor for the ImagEVAL benchmark. Discussions with Alexis Joly and Marin Ferecatu were also beneficial. The work presented in this paper was partially supported by the European Commission

under contract FP6-045389 Vitalas. Picture copyrights are, in order of appearance : Shah-Jacana/Hoa-Qui, Bassignac-Gamma, CADN_N050041, Patrimoine Photo, Dufour-Gamma, Faillet-Keystone, Bassignac-Gamma, Keystone, Bassignac-Gamma, Bassignac-Gamma, Bassignac-Gamma.

7. REFERENCES

[1] J. Amores, N. Sebe, and P. Radeva. Efficient object-class recognition by boosting contextual information. In *IbPRIA*, 2005.

[2] S. Boughorbel. *Kernels for Image Classification with Support Vector Machines*. PhD thesis, Paris XI, 2005.

[3] N. Boujemaa et al. Ikona: interactive specific and generic image retrieval. In *MMCBIR*, 2001.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM : a library for support vector machines*, 2001.

[5] Y. Chen, J. Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *PAMI*, 2006.

[6] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *JMLR*, 5:913–939, 2004.

[7] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *ECCV'04 Workshop on Statistical Learning in Computer Vision*

[8] F. Cutzu, R. Hammoud, and A. Leykin. Distinguishing paintings from photographs. *Computer Vision and Image Understanding*, 100:249–273, 2005.

[9] M. Everingham et al. The 2005 pascal visual object classes challenge. In *Selected Proceedings of the First PASCAL Challenges Workshop*, 2006.

[10] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR 2004, Workshop on Generative-Model Based Vision.*

[11] M. Ferecatu. *Image retrieval with active relevance feedback using both visual and keyword-based descriptors*. PhD thesis, University of Versailles Saint-Quentin-En-Yvelines, 2005.

[12] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Computer Vision and Pattern Recognition*, 2003.

[13] A. Guérin-Dugué and A. Oliva. Classification of scene photographs from local orientations features. *Pattern Recognition Letters*, 21:1135–1140, 2000.

[14] A. K. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29:1233–1244, 1996.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[16] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *ICML*, 1998.

[17] P.-A. Moëllic and C. Fluhr. Imageval 2006 official campaign. Technical report, CEA List, 2006.

[18] H. Müller, S. Marchand-Maillet, and T. Pun. The truth about corel - evaluation in image retrieval. In *CIVR*, 2002.

[19] P. Mylonas, T. Athanasiadis, and Y. Avrithis. Improving image analysis using a contextual approach. In *WIAMIS*, 2006.

[20] A. Oliva and A. Torralba. Modeling the shape of the scene : a holistic representation of the spatial envelope. *IJCV*, 42:145–175, 2001.

[21] A. Oliva and A. Torralba. Scene-centered representation from spatial envelope descriptors. In *Biologically Motivated Computer Vision*, 2002.

[22] C. Picault. Constitution of the imageval database, an end-user oriented approach. Technical report, Paragraphe Laboratory, Université Paris 8, 2006.

[23] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ALMC*, 1999.

[24] J. Ponce et al. *Toward Category-Level Object Recognition*, chapter Dataset Issues in Object Recognition. Springer-Verlag Lecture Notes in Computer Science, 2006.

[25] R.J. Qian, P. Van Beek and M.I. Sezan *Image Retrieval Using Blob Histograms*. In *ICME*, 2000.

[26] N. Serrano, A. Savakis, and J. Luo. A computationally efficient approach to indoor/outdoor scene classification. In *ICPR*, 2002.

[27] N. Serrano, A. E. Savakis, and J. Luo. Improved scene classification using efficient low-level features and semantic cues. *PR*, 37:1773–1784, 2004.

[28] E. Spyrou, H. L. Borgne, T. Mailis, E. Cooke, Y. Avrithis, and N. O'Connor. Fusing mpeg-7 visual descriptors for image classification. 2005.

[29] M. Szummer and R. W. Picard. Indoor-outdoor image classification. *Workshop on Content-based Access of Image and Video Databases*, 1998.

[30] A. Vailaya, M. Figueiredo, A. Jain, and H. J. Zhang. Content-based hierarchical classification of vacation images. *IEEE Multimedia Systems*, 1999.

[31] A. Vailaya, H. Zhang, C. Yang, F.-I. Liu, and A. K. Jain. Automatic image orientation detection. *Ieee Transactions On Image Processing*, 11, 2002.

[32] C. Vertan and N. Boujemaa. Upgrading color distributions for image retrieval: can we do better ? In *International Conference on Visual Information Systems*, 2000.

[33] T. Westerveld and A. P. de Vries. Experimental evaluation of a generative probabilistic image retrieval model on 'easy' data. In *SIGIR Multimedia Information Retrieval Workshop*, 2003.

[34] J. Willamowski, D. Arregui, G. Csurka, C. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *ICPR Workshop Learning for Adaptable Visual Systems Cambridge*, 2004.

[35] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006.

[36] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classifcation of texture and object categories: An in-depth study. Technical Report RR-5737, INRIA Rhône-Alpes, 2005.

[37] L. Zhang, M. Li, and H.-J. Zhang. Boosting image orientation detection with indoor vs. outdoor classification. In *Workshop on Applications of Computer Vision*, 2002.