Enhancement of Textual Images Classification using their Global and Local Visual Contents

Sabrina Tollari, Hervé Glotin, and Jacques Le Maitre

Laboratoire SIS - Equipe informatique,
Université de Toulon et du Var,
Bâtiment R, BP 132,
F-83957 La Garde cedex, France
{tollari,glotin,lemaitre}@univ-tln.fr

Abstract. This paper deals with the existence of a dependence between the textual indexation of an image (a set of keywords) and its visual indexation (color and shape features). This experience has been realized on a corpus of news photos manually indexed by keywords extracted from a hierarchically structured thesaurus. First, a reference classification of these photos was constructed from their textual indexation (regarded as relevant), then textual and visual features characterizing these classeswere constructed. Finally, they were used to evaluate performances of a content-based image retrieval combining textual and visual descriptions. Results of the visuo-textual classification show an improvement of 54% against classification using only textual information.

Keywords Information retrieval, content-based image retrieval(CBIR), visuo-textual fusion, vectorial model

1 Introduction

Within the research field of multimodal image indexing, visual modality is dominant, and despite the rich semantic of textual modality, it is largely ignored in combination with the visual one. Information retrieval systems based on textual modality are now very efficient [1]. The simple and common vectorial system given by Salton [15] has demonstrated its robustness. But this system requires the construction of an index (or a thesaurus) which is mostly carried out by documentalists who manually assign a limited number of keywords describing the image content.

On the other side, existing image engines allow users to search for images via a keywords interface or via query by image example [5], [6], [2], [13], [9], [12]. Most of them are based on visual similarity measures between an image reference and a test one. Nevertheless, most of WWW image engines allow the user to form a query only in term of keywords. To build the image index, keywords are extracted heuristically from HTML documents containing each image, and/or from the URL image. But giving too many keywords, for a precise query, the user may give information that narrows the scope of possible result images. Here again, the query must contain only a few amount of keywords in order to get few answers.

Unfortunately it is difficult to include visual cues within a WWW navigator framework. Therefore, it could be interesting to use a second filter stage, adding visual cues which have been put in correspondance with a given textual thesaurus, in order to refine the query.

In this paper we demonstrate such a system that combines textual and visual statistics in a single stochastic fusion for content-based image retrieval(CBIR). By truly unifying textual and visual statistics, one would expect to get better results than either used separately. Textual statistics are captured in vector form, used first in an Ascendant Hierarchical Classification (AHC) resulting in few semantic classes. Visual statistics are then drawn inside these classes, based on color and orientation histograms. The last stage consists in a fusion approach, taking advantage of coupling between the textual content of the document and its image content. Search performance experiments are reported for a database containing 600 images collected by Editing, a press agency, involved in the RNTL Muse Project [3]. All pictures are manually indexed by keywords from a hierarchical thesaurus and saved in an XML file following the MPEG-7 format [10]. Results of the visuo-textual classification show an improvement of 54% against a direct classification using textual information alone.

2 Construction of textual semantic reference classes

First, in order to map textual and visual information, we need to get a certain number of semantic classes containing few image samples. In this purpose textual statistics are captured in vector form, and we run the Ascendant Hierarchical Classification (AHC) [8] algorithm described in this section. One can use another method such as a Hopfield network to build semantic classes [18].

Let $D = \{d_1, d_2, \ldots, d_m\}$ a document set and $T = \{t_1, t_2, \ldots, t_n\}$ a keyword set, the vectorial model ([14],[16], [15], [1]) describes the document d_i as:

$$\boldsymbol{d_i} = (\omega_{1,i},\ \omega_{2,i},\ \ldots,\ \omega_{j,i},\ \ldots,\ \omega_{n,i})$$

where $\omega_{j,i}$ is the term-weighting, the best known is tf-idf schemes. In this study, for each keyword of the thesaurus, a vector element is initialized to 1 if the keyword belongs to the image, to 0 if not. One thus has $\omega_{j,i} \in \{0, 1\}$. The hierarchical structure of the thesaurus implies that if an image is indexed by t_j and $t_j \prec t_k$ then it is also indexed by t_k. Therefore, using the thesaurus, one can extend the vector $\boldsymbol{d_i}$ [11] so that $\forall j, k \in [1, n]$, $\omega_{k,i} = 1$ if $\omega_{j,i} = 1$ and $t_j \prec t_k$ else 0. The usual similarity mesure in the vectorial model is the cosinus. Let d_k and d_l be two images:

$$cos(\boldsymbol{d_k}, \boldsymbol{d_l}) = \frac{\sum_{j=1}^{n} \omega_{j,k} \times \omega_{j,l}}{\sqrt{\sum_{j=1}^{n} \omega_{j,k}^2} \times \sqrt{\sum_{j=1}^{n} \omega_{j,l}^2}}$$

where $\omega_{j,k}$ and $\omega_{j,l} \in \{0,1\}$. In this case a simple distance is then defined as:

$$dist(d_k, d_l) = 1 - |cos(\boldsymbol{d_k}, \boldsymbol{d_l})|.$$

Two classes C_p et C_q are merged if the distance $D(C_p, C_q)$ is small enough. A first definition for $D(C_p, C_q)$ can be the distance of the nearest neighbour:

$$D(C_p, C_q) = \min\{dist(i,j); i \in C_p, j \in C_q\}$$

but results on our database generates too small or too large classes. The distance of the maximum diameter:

$$D(C_p, C_q) = \max\{dist(i,j); i \in C_p, j \in C_q\}$$

gives uniform classes, but without semantic homogeneity. A third usual distance, the average distance:

$$D(C_p, C_q) = \frac{\sum_{i,j}\{dist(i,j); i \in C_p, j \in C_q\}}{Card(C_p) \times Card(C_q)}$$

mainly gives same results as the first one. We then defined another one, thresholding the maximum diameter method by an empiric value (0.7).

The continuing criterion T in the final algorithm of the AHC (see below) is defined in order to assure semantic homogeneity within one class and enough image samples: classes are merged until the last distance obtained is higher than 0.55.

Ascendant Hierarchical Classification (AHC)

```
program AHC
  input
    E: the set of n elements to classify
    Dist: the array n*n of distances between elements
  ouput
    C: a set of semantic classes
  begin
    For each element e in E
      Add Classe(e) in C
    end For
    While T do
      Merge the 2 nearest classes
    end while
end.
```

Finally, after removing classes having less than 8 samples, we obtain 24 a priori classes (some are given in table 1), for a total of 517 images.

Classe	T_{f_1}	T_{f_2}	T_{f_3}
1	Mexique	Politique	Portrait
2	Israël	Judaïsme	Patrimoine
3	Constructeurs	Transport	Automobile
4	Contemporaine	Portrait	Rhône
5	Portrait	Armée de l'air	Aéronautique
6	Société	Famille	Enfant
7	Cameroun	Agriculture	Géographie physique
8	Municipalité	Portrait	Les Verts
9	Elevages	Santé	Police national
10	Portrait	Média	Administrations

Table 1. List of the more frequent terms of 10 classes

3 Methods

All the features (textual or visual) are vectors of various length described in the following sections. They will be compared after normalisation to the features of the reference set, according to the Kullback-Leibler distance[1].

Each of the semantic class is randomly divided in two partitions: a reference set B_{Ex} and a test set B_{Test}. As described later on, the reference set will be used to calculate the most probable textual, visual or visuo-textual class of any image of the test set. Automatic scoring of each classification method will be easely calculated according to the a priori semantic class of each image.

Let d_T be an image of the test set B_{Test} and d_E be an image of the reference set B_{Ex}. In the following sections, S_x refers to segments set of image d_x and F refers to the set of the F_j textual or visual features.

3.1 Text-only classification

Each class C_k of B_{Ex} is represented by an average textual vector $\boldsymbol{C_k^t}^*$, which is the average of the textual vector of each images that it contains. Then the class of an image d_T described by some normalized textual vector $\boldsymbol{d_T^t}^*$, is calculated as:

$$C^t(d_T) = \text{argmin}_{k\in\{1,2,\ldots,c\}} DKL(\boldsymbol{d_T^t}^*, \boldsymbol{C_k^t}^*).$$

3.2 Visual-only classification: early fusion of visual features

We generate visual feature for each segment of x. Therefore, for a given visual feature F_j, each image x has $card(S_x)$, and so, for an image d_T and for an image d_E, there exist $card(S_T) \times card(S_E)$ distances.

[1] The relative entropy of Kullback-Leibler between two distributions d and g is: $KL(d,g) = \sum_{y\in\chi} d(y)\log\frac{d(y)}{g(y)}$. Then the Kullback-Leibler distance is $DKL(d,g) = KL(d,g) + KL(g,d)$

If one considers only the $L \in [1, min(card(S_T), card(S_E))]$ first areas, there exist $L \times L$ distances between possible areas of the image. In order to reduce the complexity of the system, we will define a distance between the visual features of two images which takes into account the best score among the smallest number calculation.

Let $\mathrm{moymin_K}$ be the function:

$$\mathrm{moymin_K} : \{\alpha_1, \alpha_2, \ldots, \alpha_M\} \rightarrow (\alpha_{min1} + \alpha_{min2} + \ldots + \alpha_{minK})/K.$$

To calculate the visual distance between an image d_T of B_{Test} and an image d_E of B_{Ex}, we calculate the L^2 possible distances and we calculate the average of the N smallest values ($N \in [1, L^2]$). We obtain for each image the distance:

$$\gamma_{F_j}(d_T, d_E) = \mathrm{moymin_N}(\{DKL_{F_j}(S_a, S_b); \forall S_a, S_b \in L\}).$$

Now, if one considers the distances between an image d_T of B_{Test}, and all images contained in a class C_k of B_{ex}, one can calculate the final distance between d_T and C_k averaging only the I first minimal distances. Then we have:

$$\delta_{F_j}(d_T, C_k) = \mathrm{moymin_I}(\{\gamma_{F_j}(d_T, d_{E_k}); \forall d_{E_k} \in C_k\})$$

where d_{E_k} is an element of the class C_k of the base of examples and $I \in [1, \mathrm{card(C_k)}]$ is the number of minimal values taken among the $\mathrm{card(C_k)}$ distances. Again the class of d_T considering feature F_j is given by:

$$C^v_{F_j}(d_T) = \mathrm{argmin}_{k \in \{1,2,\ldots,c\}} \delta_{F_j}(d_T, C_k).$$

This method allows to reject too large distances which would penalize the system, and to keep the best distances which increase the probability of being in the good class.

3.3 Combining visual and textual classifications

We now merge the textual and visual indices in order to improve the results obtained with textual classification. The main fusion strategies are early and late fusion. The first one is usual in CBIR [18], the second allows more freedom for adaptive weighting in a stochastical framework [7]. We choose the second one in this study.

For each image d_T and each class C_k , one calculates the textual distance $DKL(\boldsymbol{d_T}^*, \boldsymbol{C_k}^*)$ as explained in section **??**. Then, it is normalized and we estimate the probability of membership with the class C_k as :

$$P^t_{d_T}(C_k) = 1 - \frac{DKL(\boldsymbol{d_T}^*, \boldsymbol{C_k}^*)}{\sum_k DKL(\boldsymbol{d_T}^*, \boldsymbol{C_k}^*)}.$$

We use the same formula for each visual feature F_j:

$$P^v_{d_T}(C_k|F_j) = 1 - \frac{\delta_{F_j}(d_T, C_k)}{\sum_k \delta_{F_j}(d_T, C_k)}.$$

Therefore, the combination of the posteriors is given by:

$$P_{d_T}^{v \vee t}(C_k) = \sum_{j=1}^{card(F)} P_{d_T}^{v}(C_k|F_j) \times \omega'(F_j) + P_{d_T}^{t}(C_k) \times \omega'(F_0)$$

where $\omega'(F_j) = \frac{\omega(F_j)^p}{\sum_{i=0}^{card(F)} \omega(F_i)^p}$ and $\omega(F_j) = \frac{1-TE(j)}{\sum^{c} ard(F)_{i=0} 1-TE(i)}$ and $TE(j)$ is the ER given by F_j. The parameter p increases contrast. The final class is given by:

$$C^{v \vee t}(d_t) = \text{argmax}_{k \in \{1,2,\ldots,c\}} P_{d_T}^{v \vee t}(C_k).$$

4 Experiments

4.1 Corpus

The first corpus is a database containing 600 images collected by Editing, a press agency, involved in the RNTL Muse Project [3]. All pictures are indexed by keywords from a thesaurus and saved in an XML file following the MPEG-7 format[10]. The hierarchical Thesaurus is composed of 1200 keywords with an average depth of 3. See below an example of an XML file including "Telephone" and "Radio"(simplified MPEG7 schema).

```
<?xml version="1.0" encoding="UTF-8"?>
<mpeg7:Mpeg7 xmlns:xsi="http://www.w3.org/1999/XMLSchema-instance"
             xmlns:mpeg7="http://www.mpeg7.org/2001/MPEG-7_Schema">
  <mpeg7:DescriptionMetadata>
    <mpeg7:LastUpdate>2002-10-2</mpeg7:LastUpdate>
    <mpeg7:PrivateIdentifier>BAR9501001C-1</mpeg7:PrivateIdentifier>
    <mpeg7:CreationTime>2002-10-2</mpeg7:CreationTime>
  </mpeg7:DescriptionMetadata>
  <mpeg7:ContentDescription xsi:type="ContentEntityType">
     <mpeg7:Creation>
       <mpeg7:Title>Developpement of mobile</mpeg7:Title>
       <mpeg7:KeywordAnnotation>
         <mpeg7:Keyword>Telephone</mpeg7:Keyword>
         <mpeg7:Keyword>Radio</mpeg7:Keyword>
       </mpeg7:KeywordAnnotation>
     </mpeg7:Creation>
  </mpeg7:ContentDescription>
  <mpeg7:ContentDescription xsi:type="ViewDescriptionType">
    <mpeg7:Image>
      <mpeg7:MediaUri>BAR9501001C-1.jpg</mpeg7:MediaUri>
    </mpeg7:Image>
  </mpeg7:ContentDescription>
</mpeg7:Mpeg7>
```

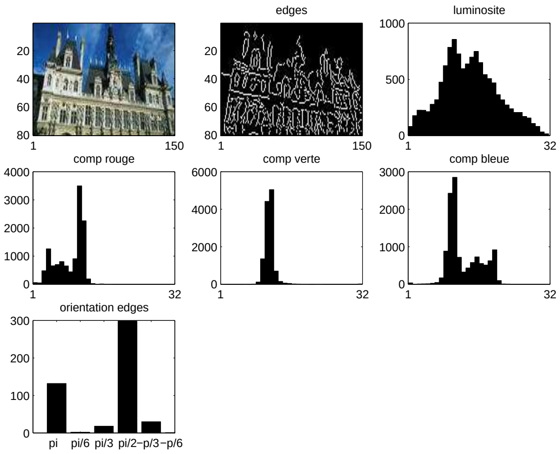

Fig. 1. The five visual features for one image

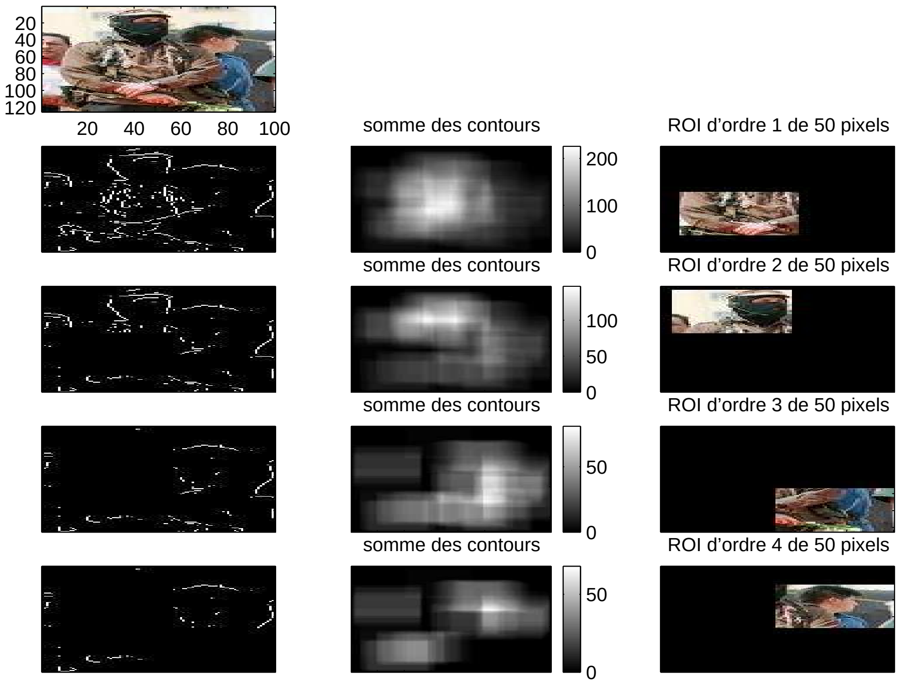

Fig. 2. Selection of the ROI

We chose to use the simplest visual features as possible. Therefore we use colors (red(F_1), blue(F_2) and green(F_3)), brightness(F_4) and direction histograms(F_5)[2]. After normalisation, these histograms are taken as visual vectors.

In order to deal with image scale variations we extracted the visual features from the original image (called "global level"), and from four local regions. The adopted segmentation approach, proposed in [17], performs an unsupervised and fast segmentation based on the Canny edge detection[4]. The local Regions Of Interest (ROI) of four different orders are automatically extracted from the global image as follow. After calculation of the edge matrice of the global image, the ROI of first order is extracted from the rectangle window of fixed size which contains the maximum number of edges. Then the ROI of second order is extracted using the edge matrice where edges corresponding to the first ROI have been removed. Other ROIs of third and fourth orders are processed iteratively. For this experiment we fixed the surface of each ROI as equal to 25%[3] of the surface of the global image. The extraction of the two first ROI is illustrated in figure 2.

A first experiment consists in classifying the test set using DKL criterion. Then this estimated classification will be easily compared to the a priori class obtained by AHC.

We then run two textual experiments: the first consists in extending the textual vector using the thesaurus as explained in section 2, the second in using directly the textual without any extension.

Table 2 gives the Error Rate (ER) obtained in the two cases. We notice that

Textual with thesaurus	Textual without thesaurus
1.17	13.72

Table 2. Classification Error Rate in %, with or without thesaurus extension

when the vectors are extended by the thesaurus, the error rate is very low. On the contrary, we see that vectors without the information from the thesaurus produces nearly 14% ER. Aiming to use our system in the case of reduced textual information as described previously, we won't extend textual vectors by the thesaurus in the following section.

4.2 Definitions of global and local visual features

4.3 Visual classification

Let $\mathrm{DKL}_{F_j}(r_t, r_e)$ be the distance for the visual features F_j between the ROI r_t of image d_T of the test set and the ROI r_e of image d_E of the reference set. We

[2] Details on the direction feature can be found in Tollari's master[17].

[3] This amount is dicussed in section 5.

start by calculating the distance between the areas of interests of equal order. Table 3 shows the results. One notices that, in general, the distances on the global indices are better, except for the direction where the first order ROI gives better results. Indeed, area 1 contains more edges, it is thus the most significant. For the green feature, the good result obtained for ROI 2 is explained by an artifact from the data (one class contains more green than the others). However, our assumption supposing that the most descriptive local areas are those which contain the most edges is confirmed, because areas 1 and 2 have the weakest error rates.

	DKL($r1, r1$)	DKL($r2, r2$)	DKL($r3, r3$)	DKL($r4, r4$)	DKL(g, g)
ER Red	81.17	79.21	81.17	82.35	**73.33**
ER Green	83.13	**78.03**	86.66	80.78	78.43
ER Blue	82.35	80.39	83.92	84.70	**74.50**
ER Brightness	80.39	81.17	81.56	83.52	**76.40**
ER Direction	**79.60**	81.56	80.00	84.31	85.49

Table 3. Error rates (ER in %) between the areas of interests of equal order

4.4 Results of early fusion of visual features

Tables 4, 5 and 6 give the error rates obtained by early fusion while varying the parameters N, I and L.

Results in table 4 give the influence of parameter N for the values of I and L giving best results. It is noticed that parameter N has little influence for the Red, Blue, Green, and Brightness features. On the other hand, for the direction, one observes a real improvement of the ER when one takes large N. Results in table 5 show that it is better to check if the test image is similar to several images of the same class rather than to only one. Lastly, in table 6, one notices that first order ROI alone ($L = 1$) is not sufficient and that fourth order ROI brings only littleinformation as expected, because ER for $L = 4$ is worse than for $L = 3$.

N	1	2	3	4	5	6	7	8
ER Red	**71.76**	72.54	72.54	73.72	76.47	77.64	77.64	76.07
ER Green	**76.07**	77.64	77.64	76.86	76.86	76.47	78.82	78.82
ER Blue	77.64	**77.25**	79.60	80,00	79.60	81.56	81.96	81.96
ER Brightness	**77.64**	79.21	77.64	77.64	79.21	79.21	78.82	78.03
ER Direction	83.52	80.39	80.39	80,00	79.21	78.82	78.43	**76.86**

Table 4. Influence of the parameter N on the Error Rates (ER in %) ($I = 4, L = 5$)

I	1	2	3	4
ER Red	75.68	74.50	**71.76**	**71.76**
ER Green	79.60	78.03	76.86	**76.07**
ER Blue	78.03	77.64	78.03	**77.25**
ER Brightness	79.21	78.03	**76.07**	77.64
ER Direction	84.70	78.03	**76.86**	**76.86**

Table 5. Influence of the parameter I on the Error Rates (ER in %) ($L = 5$)

L	1	2	3	4	4+g
Dimension L^2	1	4	9	16	25
ER Red	81.17	78.82	76.07	76.07	**71.76**
ER Green	83.13	78.82	**75.68**	79.60	76.07
ER Blue	82.35	80.00	79.60	81.56	**77.25**
ER Brightness	80.39	79.60	78.03	**77.64**	**77.64**
ER Direction	79.60	78.03	**76.07**	76.47	76.86

Table 6. Influence of the parameter L on the Error Rates (ER in %) ($I = 4$)

It is also noticed that, for $L = 5$ (4+g), the global indices make a clear improvement of the ER, except in the case of the direction feature, which was foreseeable. If one compares these results with those of table 3, one notices an ER decrease of 5% to 10% using the local indices, and an improvement of 2% on the global ones. Moreover, the early fusion reduces the time computation.

Figure 3 describes the results obtained for the fusion of textual classification without thesaurus (E.R. 13.72%) and several visual classifications. The first result (T+Vis[Local]) is obtained using only best classifications of early fusion of the ROI ($L \in [1, 4]$) only. The second (T+Vis[Global]) considers only classifications on the global indices. The third (T+Vis[Local+Global]) uses the best parameters of early fusion of the local and global indices ($L \in [1, 5]$). The last (T+Vis[Dir+Global]) takes into account the global features for the red, green, blue and brightness features, and the local direction calculated by DKL(r1,r1). On this figure, one notices that our simple ROI fonction generally improves classification compared Global for the same p. Naturally, all methods converge to the textual ER($p > 8$). Table 7 summarizes rising of textual classification by the visual classification.

Textual without thesaurus	Fusion visuo-textual	Gain
13.72	6.27	+54.3

Table 7. Result of the late fusion of visual and textual classification in %

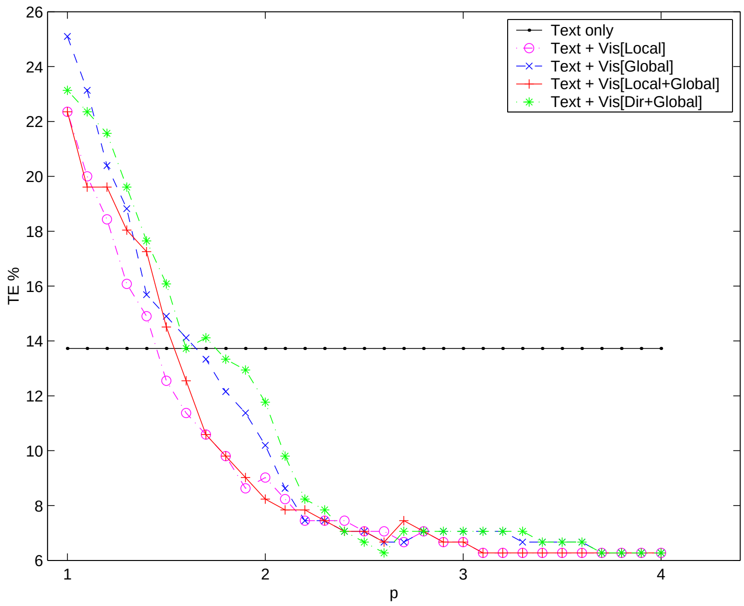

Fig. 3. Error Rate of the different systems for various p factor : text only and combining of textual with various visual contents (see text for details).

5 Discussion and Conclusion

Our corpus being only of 600 images, our method must be tested on a basis of more significant data in order to refine the results. Other visual attributes as texture or shape could be used. Many criteria and parameters remain to be studied to improve visual description, as the influence of the size and the shape of the areas of interest. The number of pixels of each local image is a parameter which could be optimized. In this first use, it is fixed a priori at 1/4 of the number of pixel of the global image. It would be interesting to compare the performances of the system by taking more reduced or focused local images, of about 1/16 of the number of pixels of the global image. Indeed, the more focused on the relevant areas of the image the visual features are, thus the more precise the classification should be.

Moreover the description of certain images by local area of interest can be more beneficial for certain types of images than for others. An automatic method determining if an image is of this type or not, would increase the performance of the system. A possible extension to this study would thus consist on the adaptive calculation of the size of the local images according to a measurement of the edge density on the image or an entropic criterion. Indeed, the more we expect an image to contain information, the more numerous but of reduced size the local images can be.

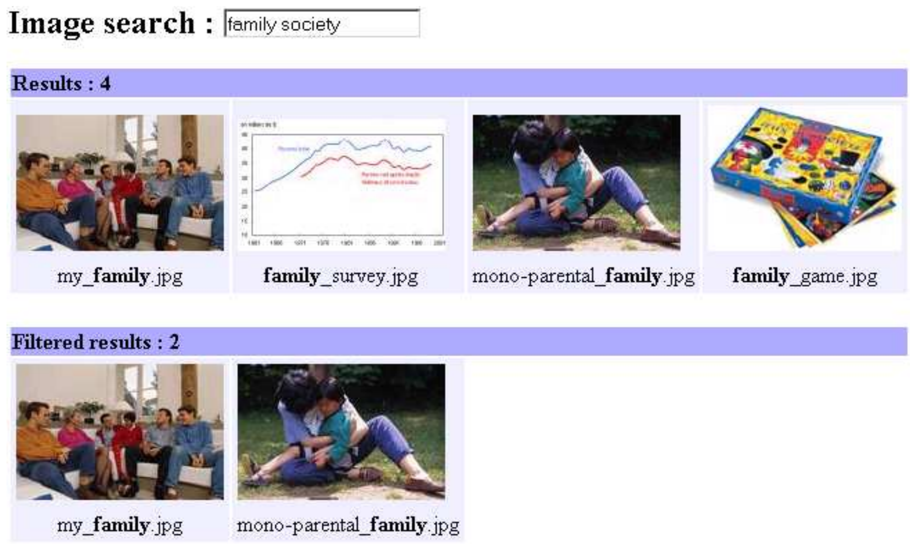

Fig. 4. The system is expected to be an efficient filter for image search results.

We presented a simple system to unify textual and visual informations. We showed that visual information reduces the errors of textual information without thesaurus of about 50%, which is very promising because of the simplicity of the method. Our system can be added like a fast visual filter (see figure 4) on the result of a request of images on a search engine (such as *Google*), requested with a small number of keywords, and thus without the use of thesaurus (otherwise no image is found by the search engine).

We could reverse the experiment by considering the textual indices compared to visual classes. This method would make it possible to correct a bad textual indexing using the visual content. For example, if a statistical plot image of the working population was labelled automatically by ‘woman’ and ‘worker’, a comparison with visual classes representing ‘woman’ would highlight the indexation error. Therefore, it could automatically remove the word ‘woman’ from the keyword set of this image.

References

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, 1999.
2. Marinette Bouet and Ali Khenchaf. Traitement de l’information multimédia : recherche de média image. *Ingénierie des systèmes d’information (RSTI série ISI-NIS)*, 7(5-6):65–90, 2002.
3. E. Bruno, J. Le Maitre, and E. Murisasco. Indexation et interrogation de photos de presse décrites en MPEG-7 et stockées dans une base de données XML. *Ingénierie des systèmes d’information (RSTI série ISI-NIS)*, 7(5-6):169–186, 2002.

4. J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.
5. Marco La Cascia, Sarathendu Sethi, and Stan Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. Technical Report 1998-004, 9, 1998.
6. V. Castelli and L. D. Bergman, editors. *Image Databases*. John Wiley & Sons, 2002.
7. H. Glotin. *Elaboration et étude comparative de systèmes adaptatifs multi-flux de reconnaissance robuste de la parole : incorporation d'indices de voisement et de localisation*. Thèse de doctorat, ICP/Institut National Polytechnique de Grenoble & IDIAP/EPF, Lausanne, 2001.
8. G.N. Lance and W.T. Williams. A general theory of classificatory sorting strategies: I. hierarchical systems. *Computer Journal*, 9:373–380, 1967.
9. Ying Li, C.C. Jay Kuo, and X. Wan. Introduction to content-based image retrieval overview of key techniques. In V. Castelli and L. D. Bergman, editors, *Image Databases*, chapter 10, pages 261–284. John Wiley & Sons, 2002.
10. B.S. Manjunath, Philippe Salembier, and Thomas Sikora, editors. *Introduction to* MPEG-7. John Wiley & Sons, 2002.
11. Jean Martinet, Yves Chiaramella, and Philippe Mulhem. Un modèle vectoriel étendu de recherche d'informations adapté aux images. *Actes du XXème Congrès INFORSID*, pages 337–348, 4-7 juin 2002.
12. C. Nastar. Indexation d'images par le contenu : un état de l'art. *Actes de CORESA'97*, 1997.
13. W. Niblack. The QBIC project: querying images by content using color, texture and shape. *Proceedings SPIE: Storage and Retrieval for Image and Video Database*, pages 173–181, 1993.
14. G. Salton. *The SMART Retrieval System ; Experiments in Automatic Document Processing*. Englenwood Cliffs, Prenctice-Hall, New Jersey, 1971.
15. G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. *Information processing and management*, 24(5):513–523, 1988.
16. G. Salton and M.J. Lesk. Computer evaluation of indexing and text-processing. *Journal of the ACM*, 15(1):8–36, 1968.
17. S. Tollari. Rehaussement de la classification textuelle d'une base de données photographiques par son contenu visuel, 2003. Mémoire de DEA.
18. Xiang S. Zhou and Thomas S. Huang. Unifying keywords and visual contents in image retrieval. *IEEE Multimedia*, 2002.