Enhancement of Textual Images Classification using Segmented Visual Contents for Image Search Engine

Sabrina Tollari, Hervé Glotin, Jacques Le Maitre
Laboratoire SIS - Equipe informatique,
Université du Sud Toulon-Var,
Bâtiment R, BP 20132,
F-83957 LA GARDE cedex, France
{tollari,glotin,lemaitre}@univ-tln.fr

April 29, 2005

Abstract

This paper deals with the use of the dependencies between the textual indexation of an image (a set of keywords) and its visual indexation (colour and shape features). Experiments are realized on a corpus of photographs of a press agency (EDITING) and on another corpus of animals and landscape photographs (COREL). Both are manually indexed by keywords. Keywords of the news photos are extracted from a hierarchically structured thesaurus. Keywords of Corel corpus are semantically linked using WordNet database. A semantic clustering of the photos is constructed from their textual indexation. We use two different visual segmentation schemes. One is based on areas of interest, the other one on blobs of homogenous colour. Both segmentation schemes are used to evaluate the performance of a content-based image retrieval system combining textual and visual descriptions. Results of visuo-textual classifications show an improvement of 50% against classification using only textual information. Finally, we show how to apply this system in order to enhance a web image search engine. To this purpose, we illustrate a method allowing selecting only accurate images resulting from a textual query.

1 Introduction

Within the research field of multimodal image indexing, visual modality is dominant, and despite its rich semantics, the text content is rarely combined with the visual one. An image search engine would benefit from such visuo-textual fusion. We propose in this article a low cost system which is giving some partial solution to that task.

Existing image engines allow users to search for images via a keywords interface or via query by image example [7, 8, 4, 17, 12, 16]. Most of them are based on visual similarity measures between an image reference and a test one. Nevertheless, most of CBIR engines allow the user to form a query only by keywords. In the case of web search engine the image index is built with keywords extracted heuristically from HTML documents containing each image, and/or from the URL image. Then asking a too precise query giving many keywords, the user may narrow the scope of accurate images. Thus image indexing should incorporate keyword and visual features. Unfortunately it is difficult to include visual cues within a WWW navigator framework. Therefore in order to refine the query, one can use a second filter stage, by adding visual cues which have been tied with a given textual thesaurus.

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 rather than using them separately.

In our approach, textual statistics are captured in vector form [19] . We first cluster images using an Ascendant Hierarchical Classification. Visual statistics are then drawn inside these classes, based on colour, shape or orientation histograms. The last stage consists in a multimodal fusion approach. Search performance experiments are reported for two corpuses, for a total of 1200 images. The first corpus is collected by Editing, a press agency, which was involved in the french RNTL Muse Project [5]. All pictures are manually indexed by keywords from a hierarchical thesaurus and saved in an XML file following the MPEG-7 format [13]. The second corpus is from a pre-processed version of the COREL database [3, 15].

Our visuo-textual fusion model enhances by 50% a text query. Moreover, we apply our model to select most accurate images found on a web search engine with a textual query.

2 Semantic clustering

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) algorithm [11]. In the AHC algorithm, each class is initialised with exactly one element (one image), then two classes are merged if their distance is the smallest one. We stop to merge classes when all the distances between classes are higher than a certain threshold.

Let $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$ a document set and $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ a keyword set, the vectorial model [18, 20, 19, 1] describes the document d_i as $T(d_i) = (\omega_{1,i},\ \omega_{2,i},\ \ldots,\ \omega_{j,i},\ \ldots,\ \omega_{n,i})$ 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, thus $\omega_{j,i} \in \{0, 1\}$. The hierarchical structure (generated by an order relation $\prec$) 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 $T(d_i)$ [14] so that $\forall j, k \in [1, n]$, $\omega_{k,i} = 1$ if $\omega_{j,i} = 1$ and $t_j \prec t_k$, else $\omega_{k,i} = 0$. In this study, we use the cosine distance.

Finally, each semantic class is randomly split in two equal partitions: one is the reference set REF, the other is the test set $TEST$. As described later on, REF is providing example documents of each class used to estimate the class of any image of $TEST$ set, either using textual, visual or textuo-visual information. Then, hidden original semantic class of the $TEST$ images gives the classification error rate.

3 Classification methods

The class estimation of any document of the $TEST$ set is based on the Kullback-Leibler[1] distance L between its normalized[2] features and the corresponding normalized features of example documents of the REF set.

3.1 Text-only classification

Each class C_k of REF is represented by the textual vector $T(C_k)$ averaging the textual vectors of images of class C_k. Then, the text-only class of a document d of $TEST$ is estimated by:

$$C^T(d) = \text{argmin}_{all\ C_k} L(T(d), T(C_k)). \tag{1}$$

3.2 Visual-only classification: early fusion

An image is defined by a set of segments S. Let P be the set of their physical characteristics (colour, area, etc), then $V_p(s^d)$ is the visual features set with $p \in P$, for the segment s of the document d.

As we have for a given document d_i a set S_i of visual segment features, the comparison of two documents d_1 and d_2 generates $card(S_1) * card(S_2)$ distances. If one considers only the $q = min(card(S_1), card(S_2))$ first segments, q^2 distances are required.

Let $\text{L}_{V_p}(s_i^1, s_j^2)$ be the Kullback-Leibler distance for the visual features V_p between the segment s_i^1 of image d_1 of TEST and the segment s_j^2 of image d_2 of REF.

In order to reduce the complexity of the system, we define the visual distance between an image d_1 of $TEST$ and an image d_2 of REF as the average of the N smallest[3] visual distances ($N \in [1, q^2]$). Therefore, we obtain for each image d the distance:

$$\gamma_{V_p}(d_1, d_2) = \text{moymin}_\text{N}(\{L_{V_p}(s_i^1, s_j^2); \forall s_i^1, s_j^2 \in [1, q]\}). \tag{2}$$

Now, one can estimate the global distance between a test image d of $TEST$ and all images contained in a class C_k of REF, averaging the I first minimal distances. Then we have:

$$\delta_{V_p}(d, C_k) = \text{moymin}_\text{I}(\{\gamma_{V_p}(d, d_k); \forall d_k \in C_k\}) \tag{3}$$

where d_k is a document of the class C_k and $I \in [1, \text{card}(\text{C}_\text{k})]$. Finally, the visual-only class of d is estimated by:

$$C^{V_p}(d) = \text{argmin}_{all\ C_k} \delta_{V_p}(d, C_k). \tag{4}$$

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

[1] The relative entropy of Kullback-Leibler between two distributions a and b on an alphabet χ is: $divK(a, b) = \sum_{y \in \chi} a(y) \log \frac{a(y)}{b(y)}$. Then the Kullback-Leibler distance is $L(a, b) = divK(a, b) + divK(b, a)$.

[2] Features normalisation estimates the distribution of the probability that a feature has the highest possible value. Another metric would be convenient as for as it takes in account feature distribution.

[3] To express the average of the Z smallest values, we define the function:

$$\text{moymin}_\text{Z} : \{\alpha_1, \alpha_2, \ldots, \alpha_M\} \rightarrow (\alpha_{min_1} + \alpha_{min_2} + \ldots + \alpha_{min_Z})/Z$$

where $Z \leq M$ and α_{min_p} is the p^{th} smallest element of $\{\alpha_1, \alpha_2, \ldots, \alpha_M\}$.

3.3 Combining visual and textual classifications: early and late fusion

3.3.1 Early visuo-textual fusion

One can extend the previous process to larger features vectors resulting from the early fusion of textual and visual features [23]. For each image, we define a new vector $TV_p(s_i)$ which is the concatenation of its textual vector and its visual vector V_p, for its segment s_i. Let $\mathrm{L}_{TV_p}(s_i^1, s_j^2)$ be the distance between the vectors TV_p of the segment s_i^1 of image d_1 of TEST and the segment s_j^2 of image d_2 of REF. Then, as in equation (2), we have:

$$\gamma_{TV_p}(d_1, d_2) = \mathrm{moymin}_{\mathrm{N}}(\{L_{TV_p}(s_i^1, s_j^2); \forall s_i^1, s_j^2 \in [1, q]\}). \tag{5}$$

Finally, the distance $\delta_{TV_p}(d, C_k)$ and $C^{TV_p}(d)$ are estimated as in equation (3) and (4).

3.3.2 Late visuo-textual fusion

A second fusion scheme consists in merging the textual and visual classification results. This late fusion allows more freedom for adaptive weighting in a stochastic framework [10] as follows.

For any image d of $TEST$ we calculate the textual and visual classifications $C^T(d)$ and $C^{V_p}(d)$. Thus, for each image d the probability to belong to C_k is for the textual modality:

$$P('d \in C_k'|T) = 1 - \frac{L(T(d), T(C_k))}{\sum_{all\ C_k} L(T(d), T(C_k))}. \tag{6}$$

For the visual modality, we have:

$$P('d \in C_k'|V_p) = 1 - \frac{\delta_{V_p}(d, C_k)}{\sum_{all\ C_k} \delta_{V_p}(d, C_k)}. \tag{7}$$

Finally, the combination of visual and textual posteriors is estimated by:

$$P('d \in C_k'|T \vee V) = \sum_{f \in (T \vee V)} P('d \in C_k'|f) \times \omega'(f) \tag{8}$$

where $\omega'(f) = \frac{\omega(f)^h}{\sum_{f \in (T \vee V)} \omega(f)^h}$, $\omega(f) = \frac{1 - ER(f)}{\sum_{f \in (T \vee V)} 1 - ER(f)}$ and $ER(f)$ is the classification error rate given by the feature f. The parameter h increases contrast between weights. The final class is given by the maximum a posteriori criterion:

$$C^{T \vee V}(d) = \mathrm{argmax}_{all\ C_k \in REF} P('d \in C_k'). \tag{9}$$

4 Experimentations

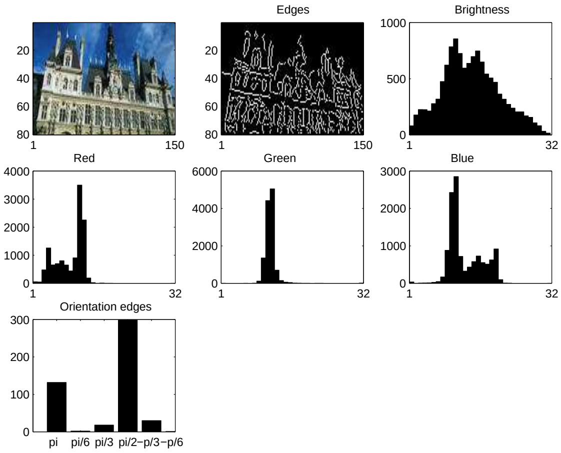

Figure 1: The five visual features for one image. Photo ©Editing.

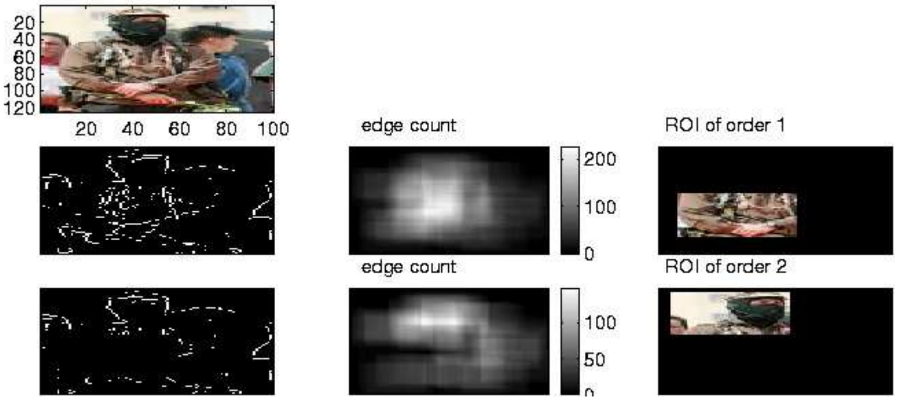

Figure 2: Selection of the ROI. Photo ©Editing.

4.1 Corpuses definitions

We made experiments on two corpuses: EDITING and COREL. Both are composed of textual and visual indexation. Each image of these corpuses are segmented into visual segments respectively.

4.1.1 EDITING's corpus

The first corpus is a database containing 600 images collected by Editing, a press agency, which was involved in the RNTL Muse Project [5]. All pictures are indexed manually by keywords from a thesaurus and saved in an XML file following the MPEG-7 format [13]. The hierarchical Thesaurus is composed of 1200 keywords with an average depth of 3.

We chose to use the simplest visual features as possible. Therefore we use colours (red(V_1), blue(V_2) and green(V_3)), brightness(V_4) and direction histograms(V_5)[4](figure 1). In order to deal with image scale variations we extracted the visual features from the original image and from four local regions. The segmentation approach, proposed in [22], performs an unsupervised and fast segmentation based on the Canny edge detection [6]. 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,

[4] Details on the direction feature can be found in [22].

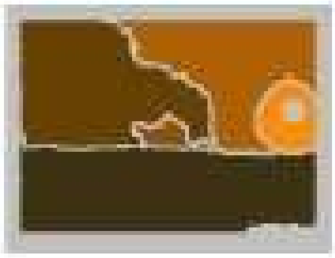

Figure 3: Example of five regions (also called blobs) of a sunset image extracted from [3].

the ROI of first order is extracted from the rectangle window of fixed size which maximises the 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 to 25% of the surface of the global image. The extraction of the first two ROI is illustrated in figure 2.

4.1.2 COREL's corpus

The second corpus is a pre-processed COREL database containing 600 labelled images. Since there is no hierarchical thesaurus for this database, legend keyword have been extended using the hypernyms of WORDNET database [9]. This thesaurus is composed of 1500 keywords with an average depth of 6.

Data were preprocessed by Computer Vision Group of University of California (Berkeley) and Computing Science Department of University of Arizona as described in [3, 2]. Each image is segmented using normalized cuts [21] (figure 3). This segmentation method tends to produce small or unstable regions. Thus, only the 10 largest regions in each image are selected. For this database, the order of any region is defined as its decreasing size rank.

Each region is described by a set of 46 features. Size is represented by the portion of the image covered by the region. Shape is represented by the ratio of the area to the perimeter squared, the moment of inertia (about the center of mass), and the ratio of the region area to that of its convex hull. Colour is represented using the average and standard deviation of (R,G,B) and (L,a,b) over the region. In the case of RGB, the 3 extracted bins are: S=R+G+B, r=R/S, g=G/S, named rgS. Texture and shapes are extracted as in [3]. Authors has chosen these features to be computable for any image region, and be independent of any recognition hypothesis.

4.2 Results of the semantic clustering

For both corpuses, we compute the semantic clustering algorithm described in section 2. We remove from the classification the classes which have less than 8 images. According to the class distribution, we estimate the error rate of the random classifier[5]. We finally obtain the classification describe in table 1.

Corpus	Minimum number images by class	Maximum number images by class	Mean number images by class	Stdev	Number of classes	Total images number
EDITING	8	98	21.5	21.4	24	517
COREL	12	46	24.3	10.10	23	559

Table 1: Results of the clusters for EDITING and COREL corpuses.

4.3 Text-only classification results

Table 2 gives the text classification Error Rate (ER) obtained for EDITING and for COREL corpuses. We notice that, in the case of EDITING, when the vectors are extended by the thesaurus, the classification is significantly improved. This is due to the high quality of the manual thesaurus. In the case of COREL, when the vectors are extended by WORDNET, the error rate is similar with the thesaurus, maybe because we use all the senses of WORDNET for any COREL keyword, as far as we do not know in what sense a keyword is associated to a COREL image.

	Textual with thesaurus	Textual without thesaurus	Origine of thesaurus	Random classifier
EDITING	1.17	13.72	manual	91.60
COREL	18.97	17.15	automatic	94.80

Table 2: Classification Error Rate in %, with or without thesaurus extension for EDITING and COREL, and for a theoretical error rate of a random classifier.

4.4 Visual-only classification results

We start by calculating the distance between the segments of equal order. Table 3 shows the results. One notices that, in general, ER on the global images are better, except for the direction where the first order ROI gives better results. Area 1 contains the highest edges density. 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, areas 1 and 2 have the weakest error rates, which is explained by the fact that they contain most of the visual information (high number of edges or large area).

[5] Let P_k be the frequency of the class C_k, the error rate of a random classifier is:

$$ER_{rand} = 1 - \sum_{k=1}^{c}(P_k)^2 = 1 - \sum_{k=1}^{c}\left(\frac{card(C_k)}{\sum_{k=1}^{c} card(C)}\right)^2$$

where c is the number of classes and $card(C_k)$ is the number of images in the class C_k.

Corpus	Features	$L_{V_p}(g,g)$	$L_{V_p}(1,1)$	$L_{V_p}(2,2)$	$L_{V_p}(3,3)$	$L_{V_p}(4,4)$
EDITING	Red	81.17	79.21	81.17	82.35	**73.33**
EDITING	Green	83.13	**78.03**	86.66	80.78	78.43
EDITING	Blue	82.35	80.39	83.92	84.70	**74.50**
EDITING	Brightness	80.39	81.17	81.56	83.52	**76.40**
EDITING	Direction	**79.60**	81.56	80.00	84.31	85.49
COREL	rgS	-	**83.57**	86.86	84.30	86.86

Table 3: Error Rates (ER in %) for different visual features for segments of equal order. $L_{V_p}(g,g)$ refers to the experiment considering only global images.

4.5 Early fusion of visual features

Parameters N, I and q could be optimized on a developpement test set, but they can be roughly estimated from table 4 and 5, and curve in figure 4. Table 4 and 5 give the error rates obtained by early fusion while varying the parameters N and I for EDITING and COREL. Table 4 shows that parameter N has little influence on the error rate both on EDITING and COREL for colour features, but is significant for direction features. Table 5 shows that it is better to check if the test image is similar to four images of the same class rather than to only one.

N	1	2	3	4	5	6
EDITING Red	**71.76**	72.54	72.54	73.72	76.47	77.64
EDITING Green	**76.07**	77.64	77.64	76.86	76.86	76.47
EDITING Blue	77.64	**77.25**	79.60	80,00	79.60	81.56
EDITING Brightness	**77.64**	79.21	77.64	77.64	79.21	79.21
EDITING Direction	83.52	80.39	80.39	80,00	79.21	**78.82**
COREL rgS	79.92	80.65	80.65	79.56	78.46	**77.73**

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

I	1	2	3	4
EDITING Blue	78.03	77.64	78.03	**77.25**
COREL rgS	85.40	79.92	78.83	**77.73**

Table 5: Influence of the parameter I on the Error Rates (ER in %) for two visual features. N is fixed according to the best result in table 4

4.6 Textuo-visual fusion: early and late fusion

Early fusion is conducted on COREL as mentioned in section 3.3.1 and gives an error rate of 6.56%. Figure 4 describes the results obtained with late fusion using textual classification without thesaurus[6] (E.R. 13.72%) and all visual classifications. We use the best parameters of early fusion. We plot error rates for different h constrast factor defined in equation 8. Naturally, all methods converge to the textual error rate for large h ($h > 8$). Table 6 summarizes rising of textual classification by the visual classification.

[6] One should use a thesaurus for the reference semantic clustering, but during a text process above various search engine, he can't.

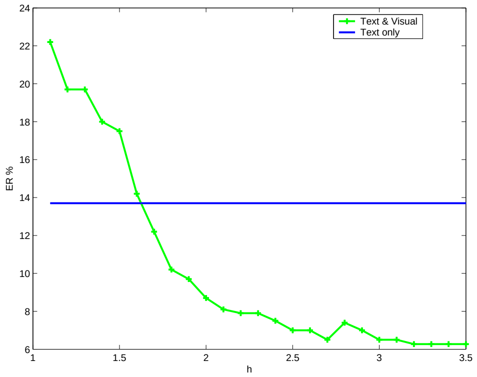

Figure 4: Error Rates of the different systems for various h factor: text only and combining of textual with various visual contents (see text for details).

Corpus	Fusion scheme	Textual without thesaurus	Fusion visuo-textual	Relative gain
EDITING	late	13.72	6.27	54.3
COREL	early	18.97	6.56	65.4

Table 6: Best results of early and late fusions of visuo-textual classification in %. Both early and late fusion enhance by more than 50% the text-only classification.

5 Application to image filtering

We demonstrated in previous sections that visual and textual informations can be efficiently merged in a multimodal search engine. We then show in this section how our system could be applied to filter any set of images found from a textual query. We load 20 images found on "Google" search engine with the textual query "Black Bear Snow". These terms are present in our reference database. Google is presenting various images, from bears to birds. Like in many other search engines, the accuracy of each image result is given to the user only by the order of each image. Thus, the user has to choose himself which images are not accurate. To give more information to the user, we compute the distance between the visual content of each image and the content of the reference database (COREL). The distances distribution is given in figure 5.

We see clearly that the distribution is bi-modal, one mode is centred on 0.03, the other on 0.05. This indicates, that all images of the first mode could be considered as accurate, and the other not. Thus we present to the user the accurate images sorted by increasing distance, from left to right and top to bottom. We see that all images containing a bear are selected.

6 Conclusion

We showed that early and late visuo-textual fusion are significantly improving text-only classification (+54%). Our method must be tested on more data in order to refine the results. Many criteria and parameters remain to be studied to improve visual description,

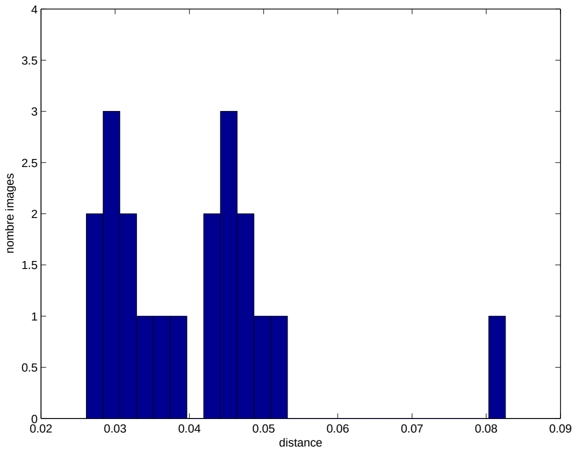

Figure 5: δ_{V_p} distances distribution for each image found by Google with the textual query: 'black bear snow' (see figure 6). This distribution is bimodal, thus, in the filter stage, images with $\delta_{V_p} < 0.04$ are considered as relevant, other are rejected.

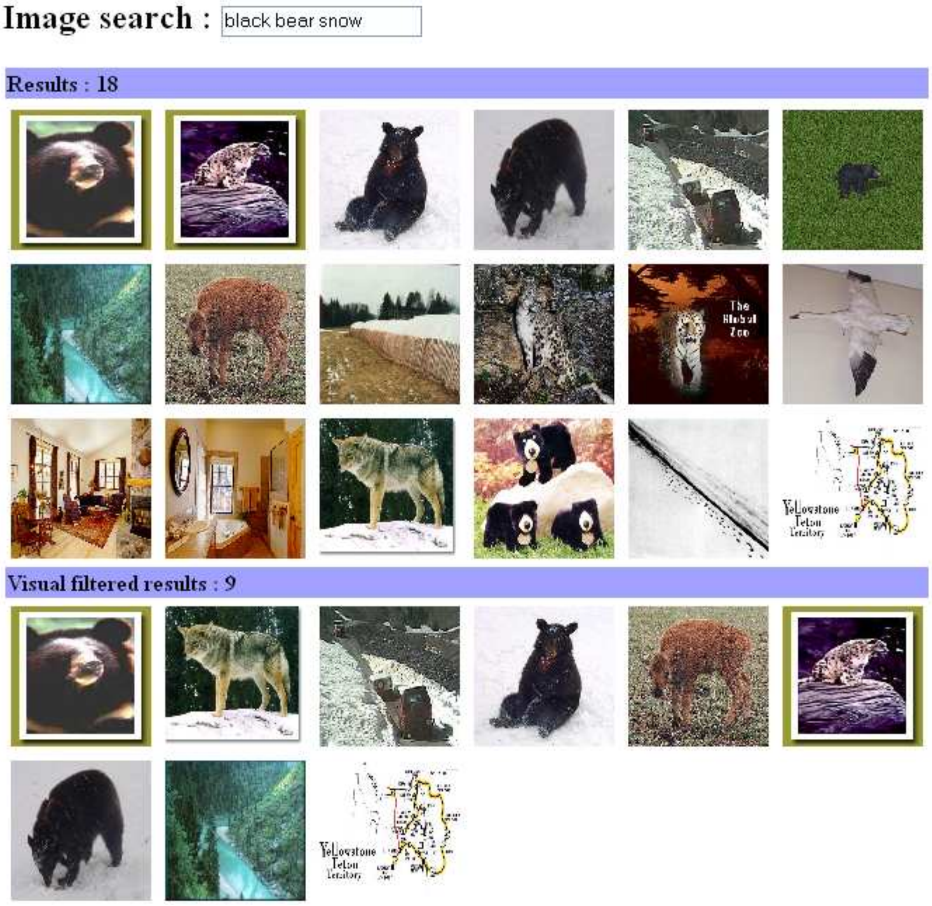

Image search : black bear snow

Results : 18

Visual filtered results : 9

Figure 6: Results of image filtering according to the distances distribution (figure 5). Top: Google results for the textual query: 'black bear snow'. Bottom: Filtered images extracted from the first cluster of the distances distribution. They are sorted by increasing distance, from left to right and top to bottom. We see that all images containing a bear are kept (the accurate one), the others are rejected.

as the influence of the size and the shape of the areas of interest. 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 consist on the adaptive size of the local images according to the edge density of the image or an entropic criterion.

We presented a simple system to unify textual and visual informations. Our system can be added like a fast visual filter on the result of a request of images on a search engine (such as *Google*), where textual requests are using a small number of keywords, and can't be extended by a thesaurus.

We could reverse the method by comparing the textual indices to visual classes. This would allow 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] K. Barnard. http://vision.cs.arizona.edu/kobus/research/data/jmlr_2003.

[3] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. In *Journal of Machine Learning Research*, volume 3, pages 1107–1135, 2003.

[4] 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.

[5] 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.

[6] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.

[7] 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.

[8] V. Castelli and L. D. Bergman, editors. *Image Databases.* John Wiley & Sons, 2002.

[9] Christiane Fellbaum, editor. *WordNet - An Electronic Lexical Database.* Bradford books, 1998.

[10] 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.* Phd thesis, ICP/Institut National Polytechnique de Grenoble & IDIAP/EPF, Lausanne, 2001.

[11] G.N. Lance and W.T. Williams. A general theory of classificatory sorting strategies: I. hierarchical systems. *Computer Journal*, 9:373–380, 1967.

[12] 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.

[13] B.S. Manjunath, Philippe Salembier, and Thomas Sikora, editors. *Introduction to* MPEG-7. John Wiley & Sons, 2002.

[14] 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.

[15] Henning Müller, Stéphane Marchand-Maillet, and Thierry Pun. The truth about corel – evaluation in image retrieval. In *The Challenge of Image and Video Retrieval (CIVR2002)*, 2002.

[16] C. Nastar. Indexation d'images par le contenu : un état de l'art. *Actes de CORESA'97*, 1997.

[17] 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.

[18] G. Salton. *The SMART Retrieval System ; Experiments in Automatic Document Processing.* Englenwood Cliffs, Prenctice-Hall, New Jersey, 1971.

[19] G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. *Information processing and management*, 24(5):513–523, 1988.

[20] G. Salton and M.J. Lesk. Computer evaluation of indexing and text-processing. *Journal of the ACM*, 15(1):8–36, 1968.

[21] Janbo Shi and Jittendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and machine Intelligence*, 22(8):888–905, 2000.

[22] S. Tollari, H. Glotin, and J. Le Maitre. Rehaussement de la classification textuelle d'images par leur contenu visuelle. *Actes du 14ième Congrés Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, 2004.

[23] Xiang S. Zhou and Thomas S. Huang. Unifying keywords and visual contents in image retrieval. *IEEE Multimedia*, 2002.