WISTI: a simple efficient Textuo-Visual Web Image Retrieval model - Specifications and Benchmarks

Sabrina Tollari
CNRS Lab. Systems and Information Sciences
University Sud Toulon-Var
BP 20132
F-83957 La Garde cedex, France
tollari@univ-tln.fr

Hervé Glotin
CNRS Lab. Systems and Information Sciences
University Sud Toulon-Var
BP 20132
F-83957 La Garde cedex, France
glotin@univ-tln.fr

ABSTRACT

In this article we specify the textual and visual algorithms that we developed and merged to design a simple efficient Textuo-Visual Web Image Retrieval System that we call WISTI (Web Image Search by Text and Image content). It integrates subband entropy profile visual features and usual mean and color standard deviation, which distributions are depicted. A simple weighted norm fusion is done with tf-idf web page analysis. WISTI is the second best model (after XEROX) according to the official European ImagEVAL Technovision 2006 campaign evaluation. We depict analyses of the fusion behavior of each of the 25 queries on 700 URLs, with Mean Average Precision curves for different text ratio. These precise analyses bring us to propose and discuss on "visualness" versus "textualness" of each query concept.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Image retrieval, query by example, query by keywords

General Terms

Experimentation, measurement, performance

Keywords

Web image retrieval (WIR), visuo-textual fusion, ImagEVAL, visual features extraction, CBIR, visualness, textualness

1. INTRODUCTION

Since content-based image retrieval is still considered very difficult, web image search engines exploit text information, such as title, filename, adjacent text to "understand" the content of Web images. However, web text information is not always reliable and informative for retrieving images. To enhance web image retrieval a good idea might be to combine textual and visual information.

Some previous works [13, 8, 12, 17, 10, 1, 9, 15] propose interesting experiments to combine textual and visual informations, but none of them propose experiments which:

(i) use a large image corpus extracted from real Web pages

(ii) measure recall and precision according to human ground truth.

The new campaign called ImagEVAL [6] gives a framework for such studies. It aims at "assessing image processing technology needed for sorting, finding and describing still images contained in vast data bases. The assessment focuses on features relating to what collection holders expect in terms of how images may be used (as described by a panel of participants from the defence, industrial and cultural sectors)."

The second task of ImagEVAL [6] aims at assessing "how techniques involving text and images work together in order to improve similar image searches within the framework of information searches on text/image data." This task is strongly "Web oriented" because it runs on real Web multimodal pages (containing text and images). Our model WISTI obtained the second best Mean Average Precision (MAP), just after Xerox's system, and before CEA's system as detailed in the results section (see also [7]).

In this paper, we present our approach, emphasising on the behavior of the visual and textual fusion according to each query. We discuss the fact that even a simple system, using simple visual features and usual tfidf textual informations, and a linear fusion of both, generates interesting and efficient results, and opens discussions on the complexity of the efficient of content based generic Web image retrieval system.

2. THE IMAGEVAL TASK 2 CAMPAIGN : COMBINED TEXT / IMAGE SEARCH

2.1 Task description

The second task of ImagEVAL [6] is "Web oriented". It consists in a retrieval task of Web images using textual and visual informations.

The database has been created by extraction of Web pages, especially from Wikipedia for copyright reasons. The Web pages have been found using classical search engines (Google and Alltheweb). An automatic segmentation of pages into text and image is performed. For this first edition, the Web pages are in French. The link between the image and its

	Query	Number of query images	Number of relevant images
1	bee	7	39
2	avocado	7	39
3	tennis ball	4	20
4	lemon	6	94
5	ladybird	6	19
6	Ethiopian flag	1	13
7	European flag	1	31
8	Picasso Guernica	3	19
9	Joconde	2	14
10	lava flow	7	66
11	Delacroix Liberty	3	11
12	Great Wall China	7	88
13	Perce Rock	6	33
14	clown fish	7	51
15	Siamese cat	6	33
16	tennis playground	9	40
17	Ayers Rock	6	41
18	zebra	6	30
19	Eiffel Tower	5	53
20	Statue Liberty	4	18
21	Niagara falls	6	51
22	teddy bear	6	9
23	screwdriver	5	20
24	poplar tree	5	19
25	map Norway	6	8

Table 1: Queries information. For the 700 URLs of the ImagEVAL Task2 Official Test set, there are 5153 images

position in the text is kept. The database is composed of a list of 700 URLs and the corresponding text and images files. The participants were also invited to use personal Web page segmentation tools. Pages were selected using topics like: "Eiffel Tower", "Lemon", "Clown Fish", "Uluru rock", "Ethiopian flag". Using Wikipedia, they focused on more "encyclopaedic" and "picturable" topics: animals, places, monuments, objects.

The goal of the task is to find all the images answering the query composed of keywords and few positive images. A query is a composition of keywords (for instance: "Eiffel Tower") and few relevant images (that did not come from the database). 25 queries have been selected. Notice that for the official run the target results was unknown. Table 1 gives the queries, the number of query images and the number of relevant images for each query.

For example, figure 2(a) gives the 7 "clown fish" query images and figure 2(b) gives the first 20 under 51 relevant images for "clown fish".

2.2 Corpus and TREC Evaluations

The Task2 official Test set is composed of 700 URLs. The Web pages of that URLs contain 10264 images. There are 25 queries that are listed in table 1.

For the 700 URLs of the official set, the total number of images is 10264 among which we only retain 5153 that are "interesting" images based on these criteria:

- not a logo, not a blank images, not small arrows... given by the official campaign list of useless images,
- a direct RGB visual extraction is running.

The Mean Average Precision (MAP) is the principal metric. All the metrics are calculated with the TREC software for each query on the first 300 images returned by the system. The TREC evaluation provides for each task: MAP, Average Precision for each run and for each query and Recall & Precision values.

3. VISUAL AND TEXTUAL FEATURES

3.1 Visual features

Content based Web images retrieval systems must be time efficient. Thus visual features must be quickly calculated, and should generate low dimensional features. On the other hand, we expect our generic visual features to be easily normalized and to allow to quantify their information content. We found interesting then to extract features giving also an idea of the amount of visual information content (that we call "visualness", see section 4.4 for more discussion). Note that a recent web CBIR used different entropic visual features [16], like some other previous works. Thus we develop a simple horizontal and vertical profile entropy based features described in the following algorithm. These features avoid object segmentation, nevertheless extract information about the projected shape of any object. For reason of efficiency, we don't use any RGB color conversion (in HUV dimension for example). Instead, we simply use the normalised $r = R/(R+G+B)$, $g = G/(R+G+B)$, and $T = R+G+B$. We also integrate in our visual features usual mean and standard deviation (std) of these three normalised colors.

for each image of L_{Lines}=$C_{Columns}$ pixels **do**
  split it in 3 equal horizontal bands ($L/3 \times C$ pixels),
  for each band b **do**
    r = R/(R+G+B)
    g = G/(R+G+B)
    T = R+G+B
    for each feature $F \in$ {r, g, T} **do**
      sl = vector of the sum of F value for each pixel of each line of band b
      hl = histogram of sl on $\sqrt{C}$ bins
      $X_{b,F,l}$ = entropy(hl)
      sc = vector of the sum of F value for each pixel of each column of band b
      hc = histogram of sc on $\sqrt{L/3}$ bins
      $X_{b,F,c}$ = entropy(hc)
      $hsurf$ = histogram of all F pixel values in band b on $\sqrt{(C \times L/3)}$ bins
      $X_{b,F,surf}$ = entropy($hsurf$)
      $X_{b,F,\mu}$ = mean of all F pixel values in band b
      $X_{b,F,\sigma}$ = std of all F pixel values in band b
    end for
  end for
end for

Finally, the concatenation of $X_{b,F,l}$, $X_{b,F,c}$, $X_{b,F,surf}$, $X_{b,F,\mu}$ and $X_{b,F,\sigma}$ for the three colors and for the three bands generates a total of 45 features for each image.

All features distributions are shown in figure 1. We see that usual mean color features for the three bands in Luminance (R+G+B) have similar distributions than Luminance normalized entropy (NH) vertical or horizontal profile, or than NH on all the surface band for Red and Green bands. These subset of features have reasonable behaviors and are

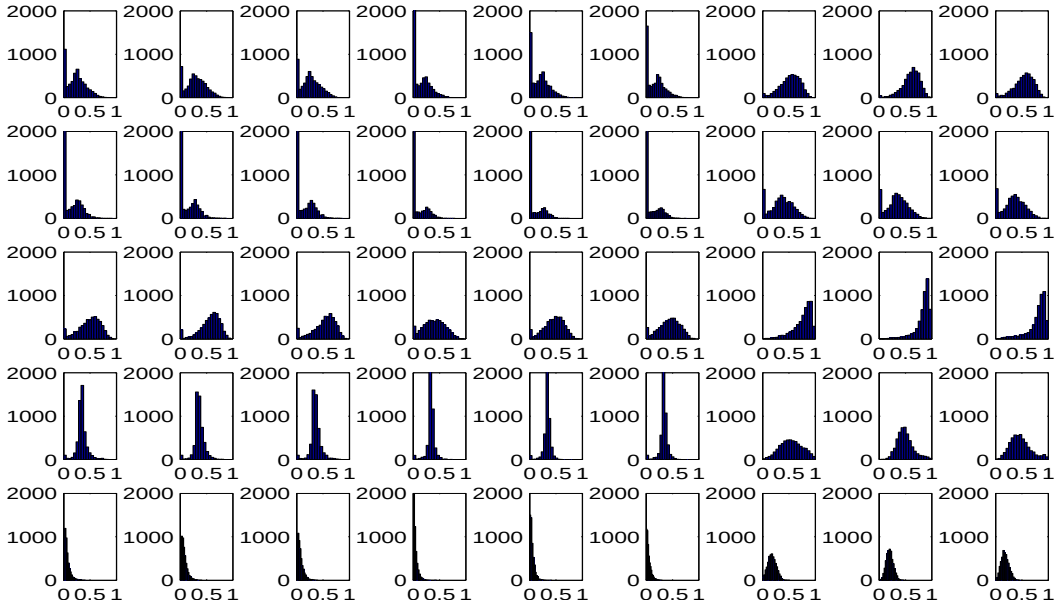


Figure 1: Distributions of the visual features. Columns from left to right: Red in the highest band (b1), in the medium band (b2) and the lowest one (b3), Green in b1, b2, b3, and Luminance (R+G+B) in b1, b2, b3. First row: normalized entropy (NH) of horizontal profile; 2nd row: NH of vertical profiles; 3rd row: NH of all the band surface; 4th row: mean of color in the bands; 5th row: std of color in the bands

certainly discriminative ones (note that NH surface of Luminance for all 3 bands has non gaussian distribution but may also be efficient). On the contrary, The Red and Green NH profiles for all bands generate an artifact with a distribution pic at 0, which may reduce their discrimination power. The color STD distribution have also a pic near 0 for Red and Green, and should be less discriminative than the color mean as it is often denoted.

3.2 Textual features

The extraction of textual features from Web pages is also simply and fastly done in two phases.

First, Web pages are converted into text without HTML tags, so that all HTML tags information (boldness, text size, distances between tags...) are lost, but ulterior treatment will be faster. Notice that all tags contains, including URLs and image name, are retained. Then, all capital letters are replaced by minuscule ones; all accentuated letters are replaced by their equivalent not accentuated (for example, 'è', 'é', 'ê' and 'ë' are replaced by 'e'; remember that page contains french text); all other special characters are removed. Finally, classical french stop words are removed. The same treatment is applied to each text query.

Second, we count each word occurrence in each Web page. Then, we suppose that each query word have the same importance and seek for texts that include at least one word of the query. We calculate the standard tfidf values [11] and associate them with each image of each Web page.

Notice that in our fast text feature extraction, all the images of a Web page are associated with the same words and the same tfidf values which can be considered as suboptimal because we do not use the information given by the distance between the image and the surrounding words (nevertheless, the efficiency of the use of this information remains an open question).

4. IMAGE RETRIEVAL BY COMBINING TEXTUAL AND VISUAL INFORMATION

4.1 Image retrieval by visual features only

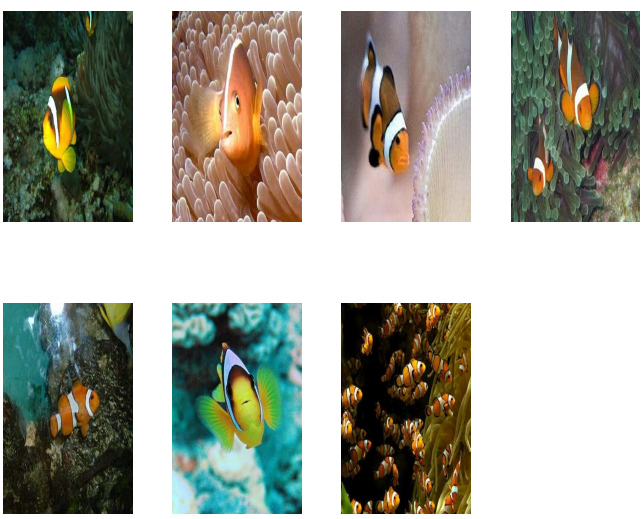
Visual vectors are simply compared using a L2 norm. For each query, the Web images are sorted from the closest to the farthest, according to the mean of their visual distance to each query image.

for each query Q **do**
  for each data image I_i **do**
    for each query image q_j of Q **do**
      calculate the distance δ as the L2 norm between the visual vector of q_j and the visual vector of I: $\delta_{i,j} = L2 - norm(q_j, I_i)$
    end for
    calculate δ_i^* the distance between the set of query images of Q and I_i, according to the criterion γ: $\delta_i^* = \gamma_{j=1}^{n_Q}(\delta_{i,j})$
  end for
  return the first 300 images which have the smallest δ^*
end for

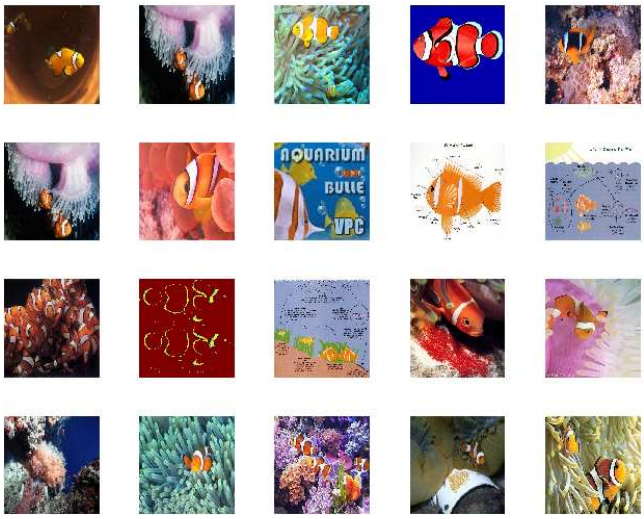
The criterion γ can be the arithmetic mean (mean), the minimum (min), the geometric mean (gm) or the harmonic mean (hm). The L2-norm could be calculate on the whole vectors or on part of the vectors.

4.2 Image retrieval by textual features only

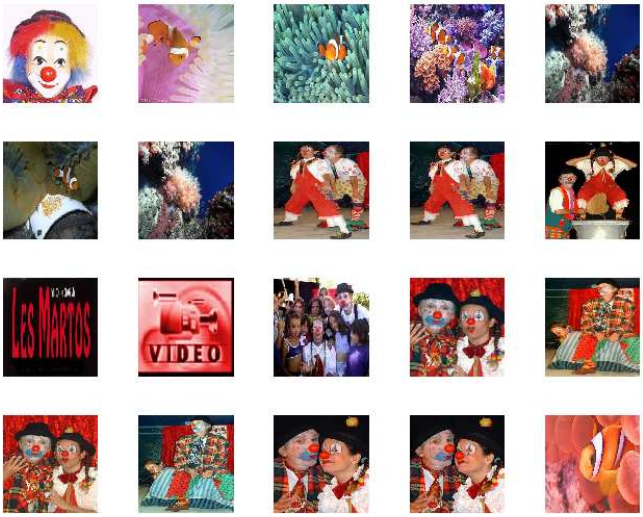
First, we extract textual features as explained in section 3.2 and for each query, we keep only images that have at least one of the word of that query in the text associated. This constraints the number of answer images as illustrated in figure 3. Thus it may decreases system performance by filtering too few number of images which have anyway strong visualness (A LSA model [2] would not generate such ar-



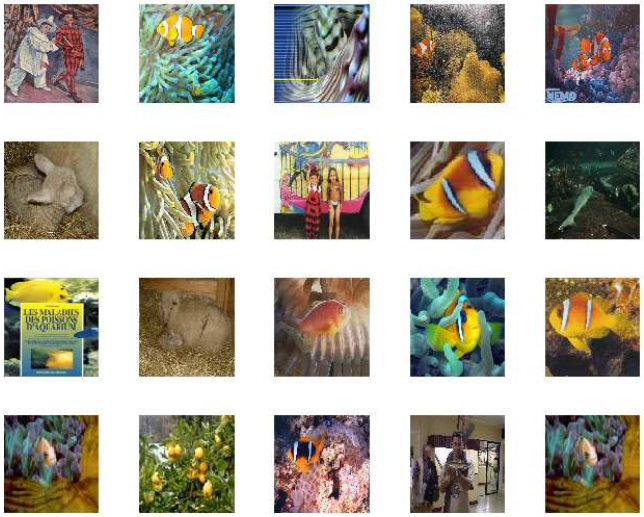
(a) The 7 “clown fish” query images



(b) First 20 “clown fish” relevant images (under 51 relevant images)

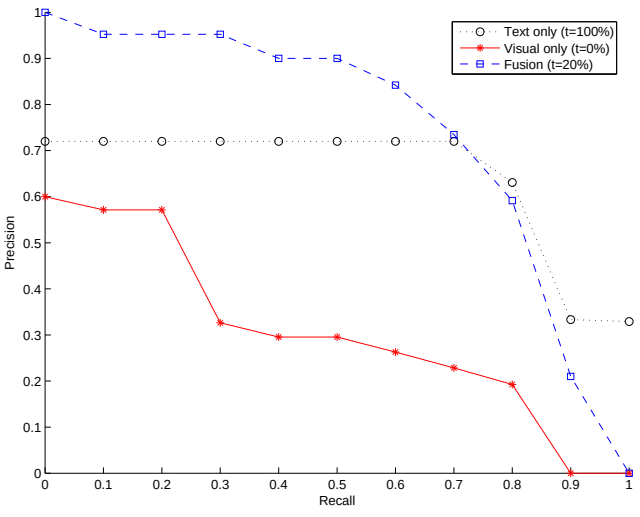


(c) First 20 Text only result images (MAP=0.54)



(d) First 20 Visual only result images (MAP=0.30)

(e) First 20 Fusion t=20% result images (MAP=0.73)



(f) Recall/precision curves

Figure 2: Examples for the query “clown fish”

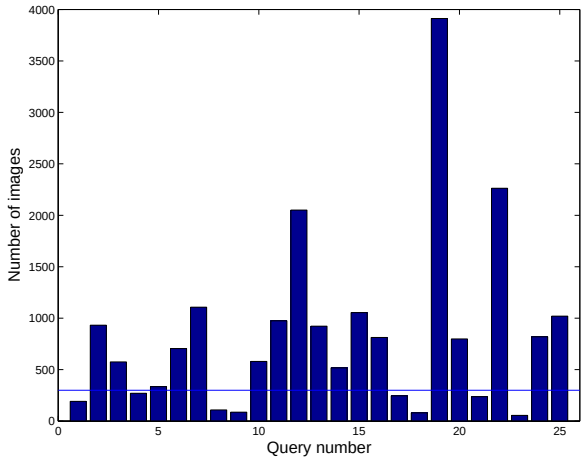


Figure 3: Number of result images by query for Text only. The line figures the 300 images level

	Text only	Visual only	Fusion
1	0.559(Run5)*	0.271(Run 16)	0.613(Run1)*
2	0.513(Run9)**	0.261(Run17)**	0.536(Run7)**
3	0.455(Run12)	0.181(Run20)	0.517(Run8)***

Table 2: Three best Official MAP results for ImagEVAL Task2 (22 runs among with 11 fusion runs, 7 Text only runs and 5 Visual only runs). *Xerox, **LSIS WISTI, ***CEA

tifact and should enhance the system). On the contrary three queries ("Eiffel Tower", "teddy bear" and "Great Wall China") obtain more than 1500 (over 5153) result images. One possible explanation is that some of the query words are frequent words (such as "tower", "bear" or "wall").

Second, for each query, we sort images by decreasing order of their tfidf values and we keep, when they exist, the first 300 images as required in ImagEVAL campaign.

4.3 Visuo-textual fusion

As for audiovisual speech recognition [4], we merge visual and textual informations using a weighted average of the visual and textual distances previously presented. A weight t is applied to the normalized textual distance D_T. The visual weight is equal to $1-t$ and is applied to the normalized visual distance D_V. Thus these two weights sum to the unity and we have the final distance $D = t \times D_T + (1-t) \times D_V$.

Ideally t could be optimized on a development set for each query, but the lack of dataset makes it impossible here. In the experimental study, we are discussing about the impact of the text rate t to the global average results over all the request, and to each request score.

4.4 Visualness versus Textualness

In [16] it is proposed to measure "visualness" ($\mathcal{V}$) of concepts, that is, to measure what extent concepts have visual characteristics discriminant for image annotation task by generic image recognition systems, since not all concepts are related to visual contents. For example, "animal" and "vehicle" are concepts that are not tied with the visual properties represented in their images directly, because there are many kinds of animals and vehicles which have various appearance

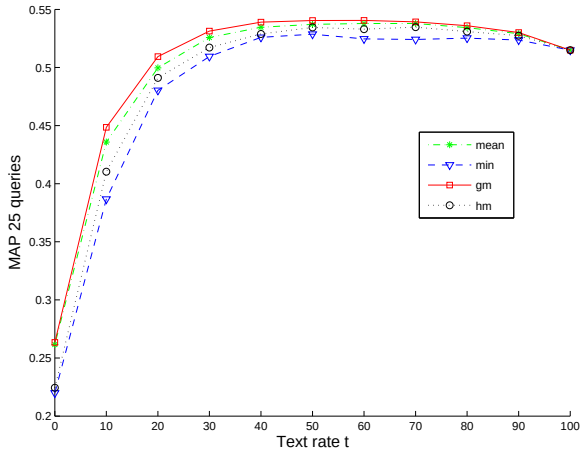


Figure 4: MAP curves varying text rate in the fusion. Best MAP is obtained with the geometric mean (gm) for $t = 60\%$

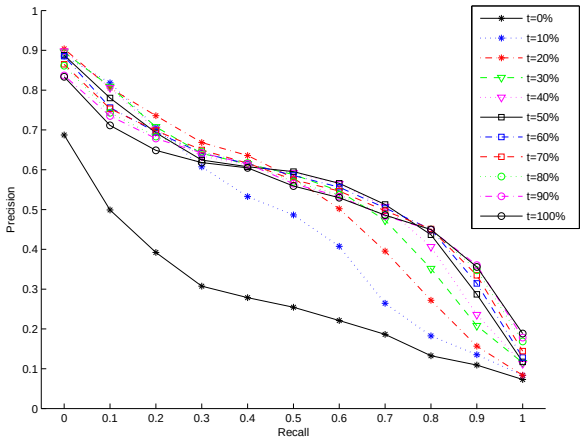


Figure 5: Recall/Precision curves for different values of text rate t with gm

in the real world. In our experiments, $\mathcal{V}$ can be illustrated by the complementary of the text rate t. We then propose to extend this notion of "visualness" to the one of "textualness" $\mathcal{T}$, that is, what concepts have textual discriminative power for example a web page image retrieval task. These notions are illustrated in the following sections.

5. RESULTS

5.1 Global results and comparison to XEROX system

We test four fusions model: *minimum*, *arithmetic*, *geometric* and *harmonic means*. As expected, the *minimum* ones is the worst fusion, whereas *arithmetic* and *geometric* fusions are the two best fusion models according to the MAP results over the 25 queries (figure 4).

Figure 5 compares recall/precision curves in function of the text rate in the fusion. Visual only ($t = 0\%$) is allways the worst one. Text only ($t = 100\%$) is, for $recall = 0$, 0.1 and 0.2, one of the worst curves, but is the best one for $recall = 1$. The curve for $t = 20\%$ is the best one for

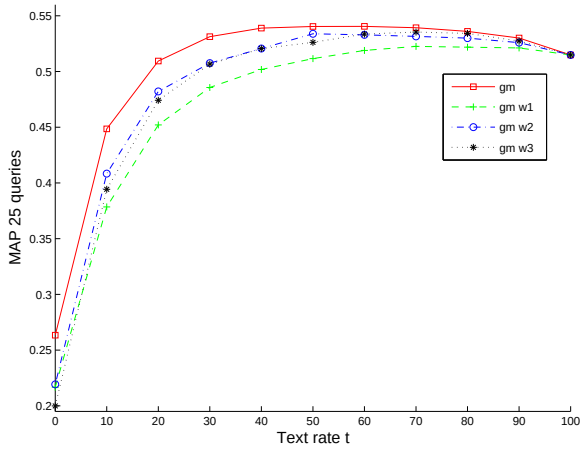


Figure 6: MAP curves in function of the visual band of the images used: w_1 (top), w_2 (center) and w_3 (bottom) compared with the geometric mean over all bands

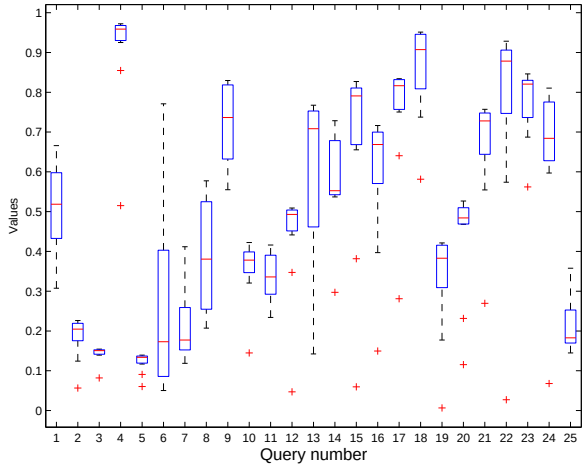


Figure 7: Mean and standard deviation of MAP values for each query for the text rate t varying from $t = 0\%$ to $t = 100\%$ by 10% step (boxes have lines of the lower, median and upper quartiles)

$recall = 0.2$ and 0.30. So, in function of the precision or the recall we want to obtain, it could be more interesting to use a different value of t. In average, the best ratio for merging visual and text distances is for $t = 60\%$.

We present in table 2 (see for details the public official ImagEVAL web site [6, 7]) the three best results in ImagEVAL campaign official test: XEROX, WISTI, and CEA systems.

Our model WISTI, whereas its simplicity, is second in Text only, Visual only and Fusion evaluations, just after XEROX which uses a powerful language model. CEA text features are of the same kind than LSIS ones (tfidf). For the three systems, visual features are around 50 dimensions, with more or less usual features that are all chosen for their low CPU cost. Thus WISTI represents a simple but nearly state of the art image Web query model.

About the processing time, the ImagEVAL consortium proposed three main processing time: (1) Features Extraction (2), Learning, Modelling processing, (3) Retrieval processing. The average time processing of these three ones, shows that WISTI runs with a similar time processing to CEA and Xerox systems, with nearly 1.5 Web page analysis per second.

5.2 Global Fusion Behaviors

The MAP variation for each query across the fusion rate t is illustrated in figure 7. We see clearly that the dynamics of the MAP is varying from a std of merely 0.5 to 0.05.

In order to analyse in detail the reason of theses differences, we split in two main classes each result: a first class with strictly monotonic MAP function of t, and the other containing non-monotonic MAP (constant or varying $MAP(t)$).

We depict strictly monotonic MAP in figures 8(a) and 8(b). Most of the queries are increasing their MAP with t. We can then infer that $\mathcal{V} << \mathcal{T}$. But we see that "Ethiopian flag", "European flag" and "Delacroix Liberty" queries are decreasing. One can assume that these three last queries are mostly 'visual concept' that are better discriminated by their visual features than by their tfidf word features ($\mathcal{V} >> \mathcal{T}$).

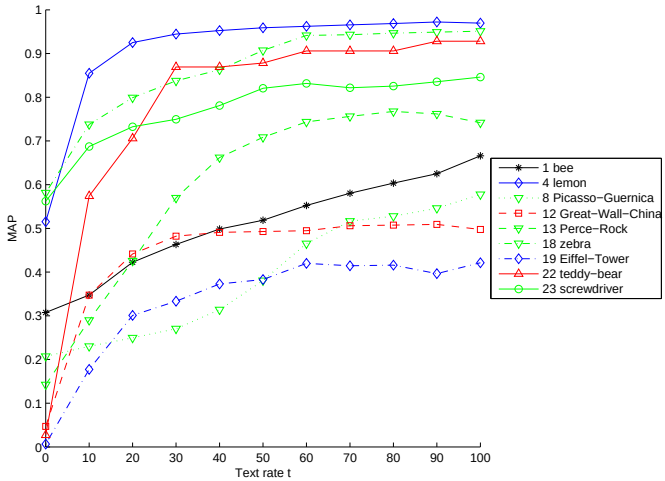
Nearly stable and low MAP values are generated for queries 2, 3, 5, 10 and 25. The pure visual MAP ($t = 0$) are weak and similar to most of the other queries, but we see that textual features do not increase MAP performances. In this case one could set that $\mathcal{V} \neq \mathcal{T}$ are weak. This can be due to the low level textual features that we use, compared to LSA or PLSA models [2, 5] that may would have enhanced and generalised textual informations. MAP, and $\mathcal{T}$, would certainly be increased using informations extracted from the word surrounding this picture.

In order to quantify $\mathcal{V}$ for each image subband we run our model for only the top band of each image (w1), the center band (w2), or the lower band (w3). Results with the geometric mean in figure 6 show that $\mathcal{V}(w1)$ is the lowest one. One interpretation could be that the sky or other irrelevant background is often present in this band. Experiments show that for this generic WIR, the joint subband information union of w_1, w_2, w_3, even if it is reaching a high number of dimension, is better than any individual subband information.

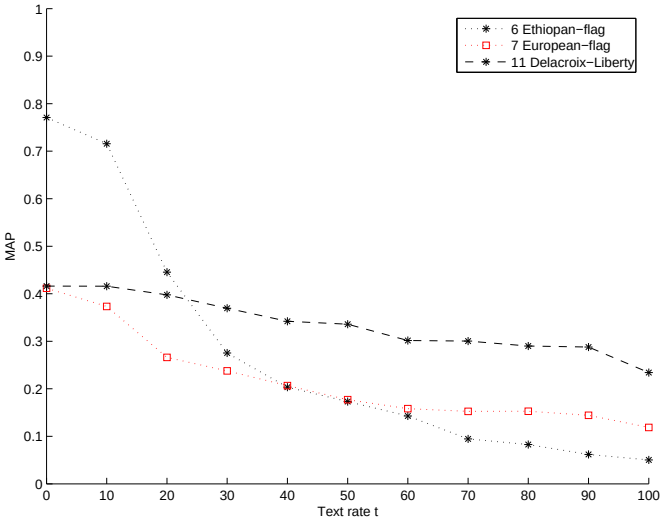
5.3 Why a visuo-textual fusion is necessary ?

The most interesting fusion behavior are illustrated in 8(c), generated for queries 9, 14, 15, 16, 21 and 24: the maximum MAP value is given for an intermediate t value. We have an optimal t around 20% (of textual information) for queries "poplar tree", "Joconde" and "clown fish", and an optimal one around 50% for "tennis ball", "Niagara falls" and "Siamese cat". Figures 2(c), 2(d) and 2(e) compare the first 20 images results for Text only, Visual only and Fusion image retrieval for "clown fish". We can see that Text only model makes mistakes because our model doesn't make the difference between the query "clown fish" and the query "clown or fish". Visual only model make mistakes because some images visually similar are not semantically similar. For these queries, neither a pure visual, nor a pure textual information is optimal. This demonstrates that for these concepts, visual and textual informations are complementary: $\mathcal{V} \perp \mathcal{T}$.

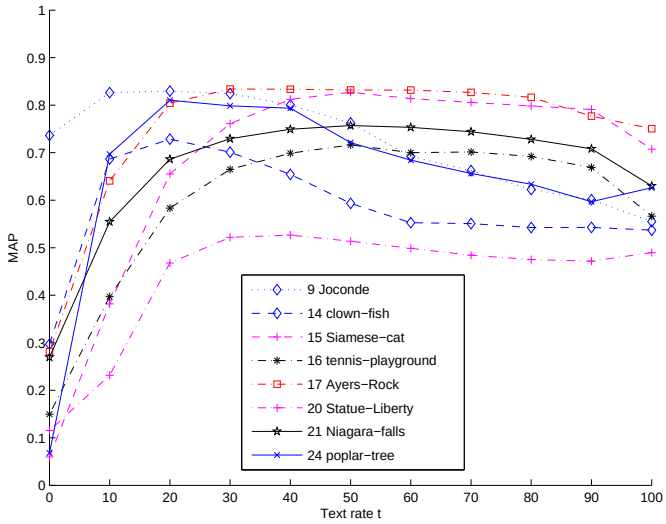
This may be due to the visual features that may be not



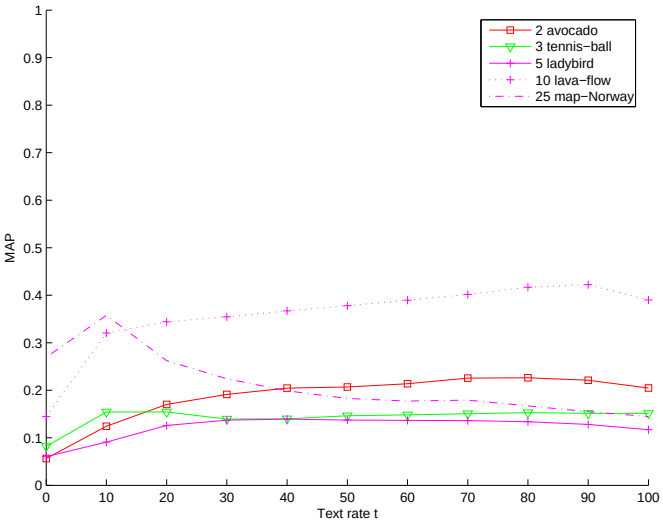
(a) Monotonic increasing MAP curves ($\mathcal{V} << \mathcal{T}$)



(b) Monotonic decreasing MAP curves ($\mathcal{V} >> \mathcal{T}$)

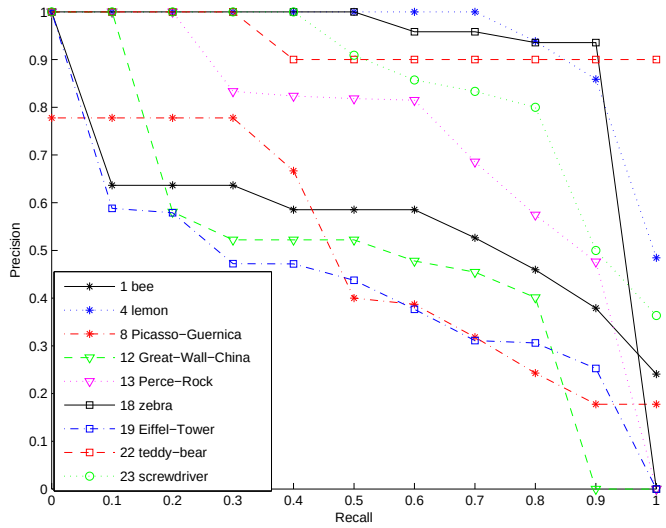


(c) Non-monotonic MAP curves ($\mathcal{V} \perp \mathcal{T}$)

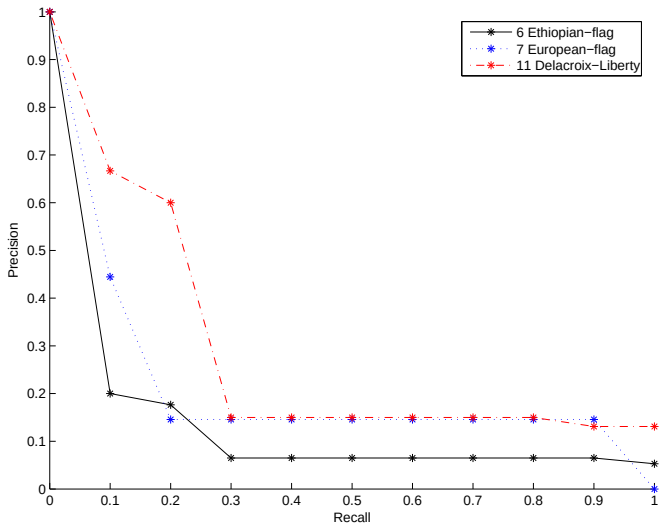


(d) Flat MAP curves (weak $\mathcal{V}$ and $\mathcal{T}$)

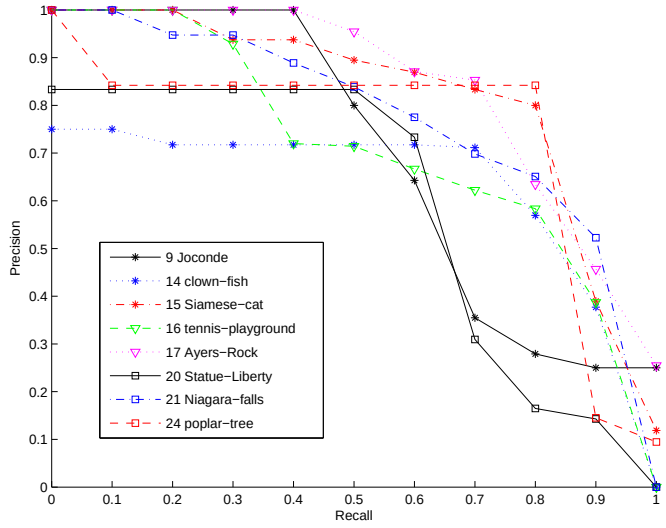
Figure 8: Monotonic, non-monotonic and flat MAP curves ($\mathcal{V}$: visualness and $\mathcal{T}$: textualness)



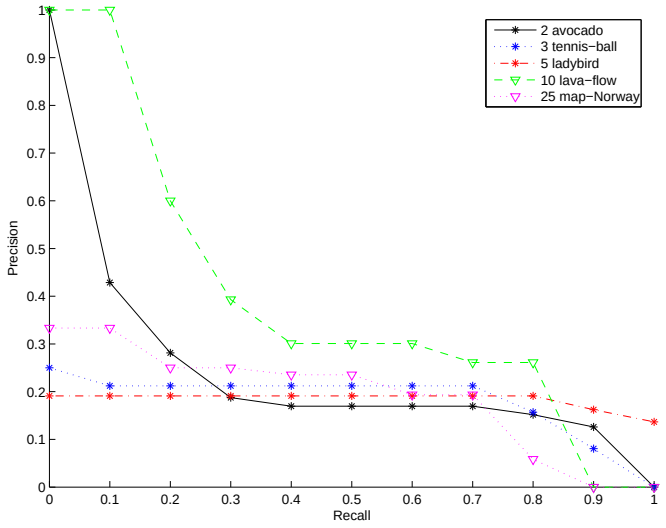
(a) Recall/precision curves which have a monotonic increasing MAP ($\mathcal{V} << \mathcal{T}$)



(b) Recall/precision curves which have a monotonic decreasing MAP ($\mathcal{V} >> \mathcal{T}$)

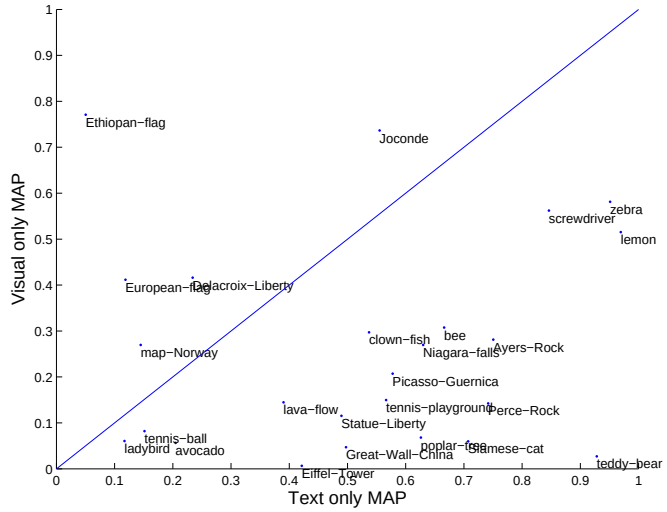


(c) Recall/precision curves which have a non-monotonic MAP ($\mathcal{V} \perp \mathcal{T}$)

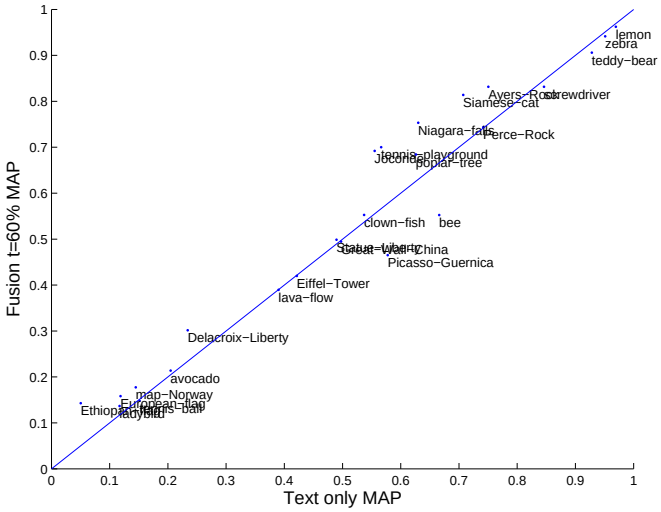


(d) Recall/precision curves which have a flat MAP (weak $\mathcal{V}$ and $\mathcal{T}$)

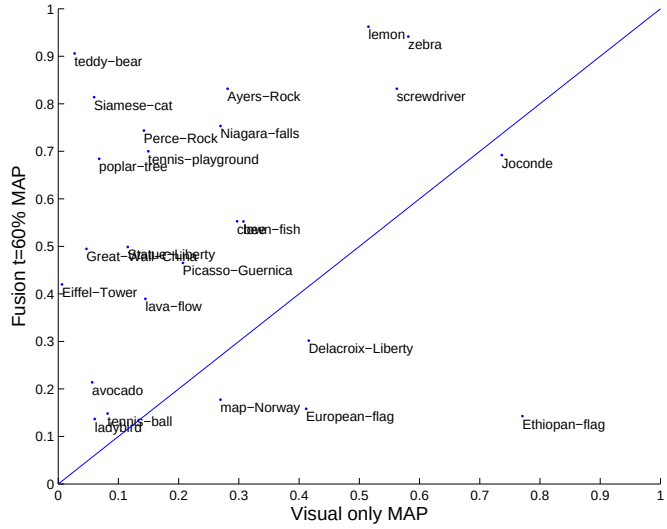
Figure 9: Fusion t=60% Recall/Precision curves



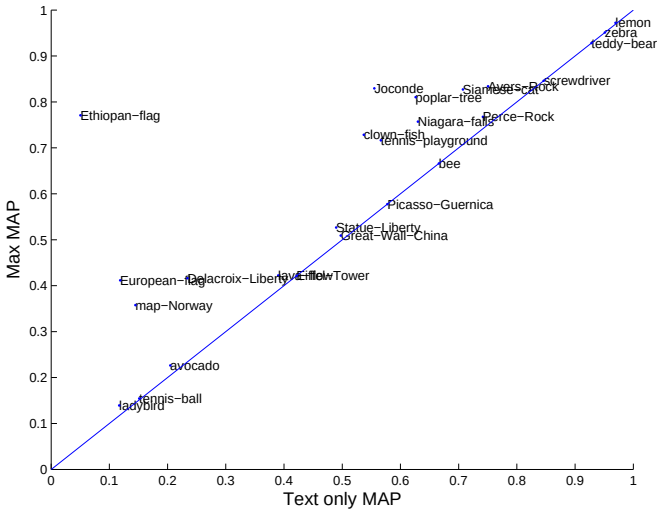
(a) Text only MAP versus Visual only MAP



(b) MAP Text only versus Fusion t=60%



(c) MAP Visual only versus Fusion t=60%



(d) Text only MAP versus Max MAP on all fusion

Figure 10: Comparison of Text only MAP, Visual only MAP, Fusion t=60% MAP and Max MAP on all fusion, for each query

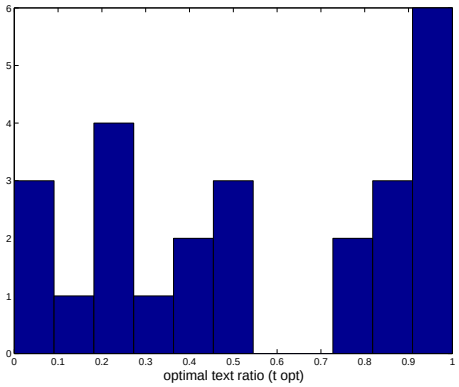


Figure 11: Distribution of the optimal t ratio over all the 25 queries

complex or precise enough, and then may give noisy informations at too low t levels. For a certain threshold visual and textual information are mixing efficient complementary informations. After this threshold, the textual information is less relevant than averaged visuo-textual information. For these queries, it is not sure that it may exist optimal visual features that would increase MAP for low t values like in the case of the ones in figure 8(b). Theses kind of concept queries shows the great interest of visuo-textual fusion model.

Figure 9 shows the recall/precision curves corresponding to the same cutting that in figure 8. Other representations are given in figure 10, where we resume each optimal MAP for each concept.

6. DISCUSSION AND CONCLUSION

This paper shows that a generic simple WIR model demonstrates non- and monotonic fusion behaviors of textual and visual informations. We then have been able to define and illustrate "visualness" and "textualness" of web page multimodal content. We argue that a good strategy for efficient WIR wouldn't be to generate complex visual features, but should be to estimate optimal fusion text ratio. It is interesting to note that if we set t_q for each query q at *topt* ratio that maximizes MAP on test set (see figure 11), we then obtain an average MAP of 0.615 (instead of 0.536), which is similar to the best model (XEROX) in ImagEval campaign. This optimisation should be made on a development set. Further study will be conducted in this direction, also for textualness estimation.

Finally, as we showed [3, 14], visual features are concept dependant, then we will in future work estimate the optimal visual features set for each concept.

Moreover it would be interesting to study if visual and/or textual ontologies could help in generalizing estimates of these parameters for any queries.

7. ACKNOWLEDGMENTS

We thank CEA LIST team for organizing IMAGEVAL campaign, and particularly Dr. Moellic.

8. REFERENCES

[1] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *International Conference on Computer Vision*, volume 2, pages 408–415, 2001.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6): 391–407, 1990.

[3] H. Glotin, S. Tollari, and P. Giraudet. Shape reasoning on mis-segmented and mis-labeled objects using approximated fisher criterion. *International Journal Computers and Graphics*, 30(2), April 2006.

[4] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin. Weighting schemes for audio-visual fusion in speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City-USA, 2001.

[5] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001.

[6] ImagEVAL. http://www.imageval.org.

[7] ImagEVAL. Nicephore days : imagEVAL international results symposium, 2006. http://www.imageval.org.

[8] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *IEEE Workshop on Content-based access of Image and Video*, 1998.

[9] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Neural Information Processing Systems (NIPS)*, 2003.

[10] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Pattern Analysis and Machine Intelligence*, 25(9):1075–1088, 2003.

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

[12] S. Sclaroff, L. Taycher, and M. L. Cascia. Imagerover: A content-based image browser for the world wide web. In *IEEE Workshop on Content-based Access of Image and Video Libraries*, 1997.

[13] R. K. Srihari. Automatic indexing and content-based retrieval of captioned images. *IEEE Computer*, 28(9):49–56, 1995.

[14] S. Tollari and H. Glotin. LDA versus MMD approximation on mislabeled images for keyword dependant selection of visual features and their heterogeneity. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2006.

[15] S. Tollari, H. Glotin, and J. Le Maitre. Enhancement of textual images classification using segmented visual contents for image search engine. *Multimedia Tools and Applications*, 25(3):405–417, March 2005.

[16] K. Yanai and K. Barnard. Image region entropy: a measure of "visualness" of web images associated with one concept. In *ACM Multimedia*, pages 419–422, 2005.

[17] X. S. Zhou and T. S. Huang. Unifying keywords and visual contents in image retrieval. *IEEE Multimedia*, 9, 2002.