LEARNING OPTIMAL VISUAL FEATURES FROM WEB SAMPLING IN ONLINE IMAGE RETRIEVAL

Sabrina Tollari

Université Pierre et Marie Curie-Paris6
UMR CNRS 7606-LIP6
Paris, F-75005, France
sabrina.tollari@lip6.fr

Hervé Glotin

Université du Sud Toulon-Var
UMR CNRS 6168-LSIS
La Garde cedex, F-83957, France
glotin@univ-tln.fr

ABSTRACT

In this article, we propose to use an online Approximation of Linear Discriminant Analysis (LDA) to improve a Web images retrieval system. Our work takes place in the official European ImagEVAL 2006 campaign evaluation. The task consists to retrieve Web images using both textual (Web pages) and visual information. Our visual features integrate subband entropy profile, usual mean and color standard deviation. A simple weighted norm fusion is done with standard tf-idf Web page text analysis. Our model is the second best model of the ImagEVAL task2. We show how, sampling online image sets from the web, one can estimate by approximated fisher criterion an optimal visual feature subsets for some query concepts and then enhance their mean average precision by 50%. We discuss on the fact that some concept may not so nicely be enhanced, but that in average, this optimization reduces by 10 the visual dimension, without any MAP degradation, yielding to a significant CPU cost reduction.

Index Terms— Information retrieval, Statistics, Image processing, Image analysis

1. INTRODUCTION

Since content-based image retrieval is still considered very difficult, web image search engines exploit text information, such as title, file name, adjacent text to "understand" the content of Web images. However, web text information is not always reliable and informative for retrieving images, so a fusion of visual and text information may be accurate.

Previous works [1, 2, 3, 4] show interesting approaches to combine textual and visual information, but none of them use a large image corpus extracted from real Web pages and measure recall and precision according to human ground truth. The new campaign called ImagEVAL [5] gives an ideal framework for such studies.

We thank CEA LIST team for organizing ImagEVAL campaign, and particularly P.-A. Moellic. This research is conducted within the AVEIR project funded by the French National Agency for Research (ANR)

In this article, we first present the task2 of the ImagEVAL campaign. Second, we describe the visual and textual feature extraction process. Then, we give methods to estimate feature discriminant power, using for example mislabeled online web sampled images. We next present our image retrieval model and baseline Mean Average Precision (MAP) results. Finally, we show that our feature discriminant power approximation is efficient for some query concepts, and we discuss on further studies.

2. IMAGEVAL TASK 2

The second task of ImagEVAL [5] consists to retrieve Web images using textual and visual information. The database has been created by extraction of Web pages, especially from Wikipedia for copyright reasons. The Web pages (in French) have been found using classical search engines. The database is composed of a list of 700 URLs and the corresponding text and images files (around 10k images among with only 5k where not small images or blank ones). Pages were selected using 25 topics: "bee", "avocado", "tennis ball", "lemon", "ladybird", "Ethiopian flag", "European flag", "Picasso Guernica", "Joconde", "lava flow", "Delacroix Liberty", "Great Wall China", "Perce Rock", "clown fish", "Siamese cat", "tennis ground", "Ayers Rock", "zebra", "Eiffel Tower", "Statue Liberty", "Niagara falls", "teddy bear", "screwdriver", "poplar tree", "map Norway".

The goal of the task is to find all the images answering the query Q. Each query is composed of a set of keywords $\mathcal{K}(Q)$ (for instance: "Eiffel Tower") and a set of few positive images $\mathcal{I}(Q)$ that did not come from the database (we call these images "reference query images"). For example, Fig. 1 gives the 6 "lemon" reference query images. Notice that for the official run the target results were unknown. For each query, the Mean Average Precision (MAP) is calculated (with the treceval software) in function of the first 300 images (over the 10k images) returned by the system.

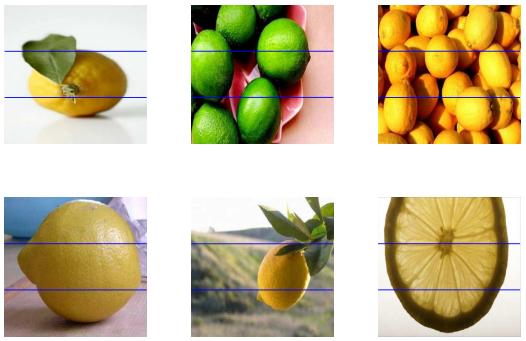

Fig. 1. The 6 "lemon" reference query images. Each query is composed of a set of keywords and a set of images. All images are divided into 3 equal horizontal subbands, then 15 visual features are extracted from each subband

3. FEATURE EXTRACTION

We propose in our system to extract features related to the amount of visual information content (called "visualness" [6, 7]). Thus we develop a simple horizontal and vertical profile entropy based features that avoid object segmentation, but extract informations from the projected shape of any object. For reason of efficiency, we don't use any RGB color conversion, we simply normalised by the luminance the three colors and use their mean and standard deviation.

for each image of $nblines \times nbcolumns$ pixels **do**
  split it in 3 equal horizontal bands
  for each band b **do**
    r = R/(R+G+B), g = G/(R+G+B), L= R+G+B
    for each feature $F \in \{$r, g, L$\}$ **do**
      sl = vector of the sum of F value for each pixel of each line of band b
      hl = histogram of sl on $\sqrt{nbcolumns}$ bins
      $NHhor$ = entropy(hl)
      sc = vector of the sum of F value for each pixel of each column of band b
      hc = histogram of sc on $\sqrt{nblines/3}$ bins
      $NHvert$ = entropy(hc)
      $hsurf$ = histogram of all F pixel values in band b on $\sqrt{(nbcolumns \times nblines/3)}$ bins
      $NHsurf$ = entropy($hsurf$)
      $Mean_C$ = mean of all F pixel values in b
      STD_C = std of all F pixel values in band b
    end for
  end for
end for

The 45 features ($3\times3\times$ ($NHhor, NHvert, NHsurf, Mean_C$, STD_C)) are normalized.

The extraction of textual features from Web pages and textual queries is done in 2 steps. First, HTML tags, special characters and stop words are removed. Second, we calculate the standard tf-idf weights. Notice that in our fast text feature extraction, all the images of a Web page are associated with the same words and the same tf-idf 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.

	t	Selection	N	MAP	Time
Text only	100%	-	-	0.515	-
Visual only	0%	without	45	0.263	309
Visual only	0%	with	20	0.271	237
Fusion	50%	without	45	0.539	309
Fusion	50%	with	10	**0.557**	**202**

Table 1. Best MAP results for $\Psi(Q)$="query+R5". Time: number of seconds used to calculate the distance between the 131 visual query images of the 25 queries and 100k synthetic vectors

4. QUERY DEPENDANT FEATURE SELECTION ON MISLABELED DATA USING WEB SAMPLING

Most of available images, like images included in web pages, are mislabeled, i.e. there is no objective bijection between surrounding words (labels) and the objects or concepts contained in the image. Another issue is the high dimension problem [8], which implies that a good visual indexing should be made up only with the visual features which have the strongest discriminating capacities. Previous works showed that simple methods like Linear Discriminant Analysis (LDA) can discriminate acoustic or visual features [9], but this method is applied on well labeled data describing a unique relation between a conceptual class and a feature.

In the context of mislabeled Web images, we apply an approximation of LDA using additional training data. For each query, we split training data into two classes. The first class (noted $\Psi(Q)$) is built using positive image examples of that query (query images $\mathcal{I}(Q)$ or/and additional training images). The second class (noted Ω) contains all the training images. For each query Q and for each visual feature X, we calculate the between variance $\hat{B}(X;Q)$ (average variance of each class), and the within variance $\hat{W}(X;Q)$ (weighted average of each class variance). Finally, we estimate for each query Q and each feature X the discriminant power $\hat{J}(X;Q)$ by:

$$\hat{J}(X;Q) = \frac{\hat{B}(X;Q)}{\hat{B}(X;Q) + \hat{W}(X;Q)} \quad (1)$$

This method, called ALDA (Approximation of LDA), has been theoretically proved in [10], and successfully tested on COREL database in [10, 11]. We showed [10] that ranking errors due to this approximation are small as long as enough samples are given for each concept in $\Psi(Q)$ set and if Ω is very large.

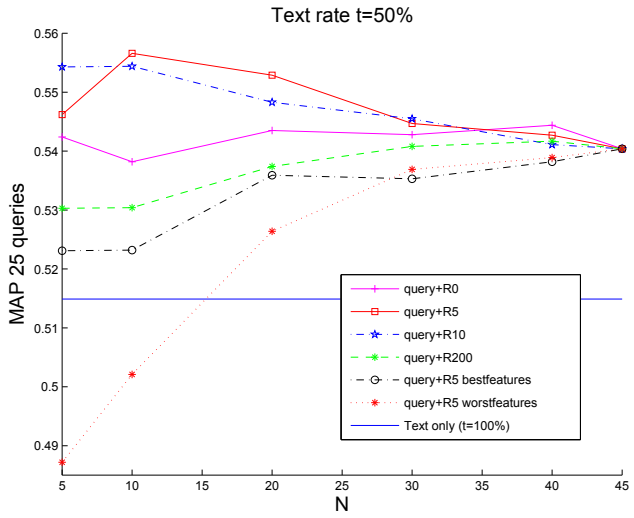

Fig. 2. MAP curves for t=50% according to different $\Psi(Q)$ set used to select the N most discriminant features by ALDA. All curves converge to the MAP value without feature selection (N=45)

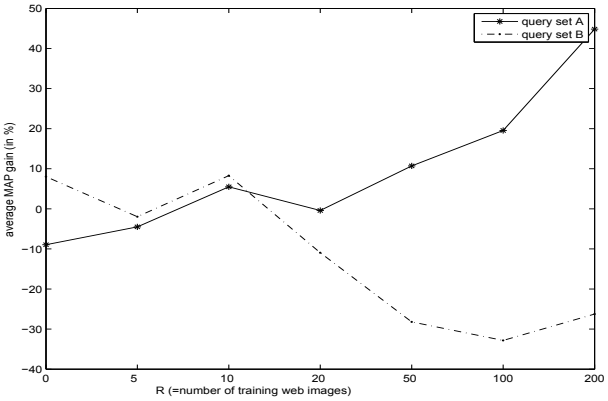

Fig. 3. MAP gains in function of the number R of training Web images used to calculate the ALDA discriminant power (we fixed $N = 10$, $t = 0\%$) Query set A ={"tennis ball", "lemon", "Euroflag", "Delacroix liberty", "tennis playground", "Ayers Rock"}. Query set B ={"Joconde", "Perce Rock", "Niagara falls", "map Norway"}. Mean MAP(A)=0.31, Mean MAP(B)=0.36. We see that queries in A give better MAP than queries in B when R increases.

Because query image sets are very small, we need additional image samples for each query. So we use a Web image search engine to retrieve images according to the keywords $\mathcal{K}(Q)$ of each query and used result images as training samples. As Web image search engines index images according to text information, the so built train set contains images which don't visually correspond to Q. Each $\Psi(Q)$ set is composed of $\mathcal{I}(Q)$ and of the first R training result images corresponding to this query. We noted these methods $\Psi(Q)$="query+Rx" where x is the value of R. If $R = 0$ ($\Psi(Q)$="query+R0") then only the references query images are used. The Ω set is for all experiments composed of all the training images.

First, to retrieve image according to visual features only (Visual only), we reduce, for each query, the visual vectors to their N most discriminant dimensions (from $N = 1$ to $N = 45$ (all features)) according to ALDA on $\Psi(Q)$ and Ω. Then, images of the official test set are sorted from the closest to the farthest, according to the geometric mean of their visual L2 distance to each query image. Second, we merge visual and textual informations by the weighted average of the visual distance D_V and the textual distance (estimated from the standard tfidf) D_T. Both distances are first normalized (by the max). Thus we have the final distance $D = t \times D_T + (1 - t) \times D_V$ where t represents the text rate in the fusion. t could be optimized on a development set for each query as we propose in further works.

5. EXPERIMENTAL RESULTS

In [7], we present in detail the official campaign results and discuss the impact of the fusion text rate t in the results. We concluded that the use of the fusion of text and visual information to retrieve images give better MAP scores than Text only retrieval (t=100%) or Visual only retrieval (t=0%) (see Tab. 1). These results were obtained without feature selection and so the time needed to calculate the visual distance is huge. We propose to estimate and then select optimal features for each query with ALDA. For all ALDA experiments, the Ω set is composed of sampled 17069 Web images. Fig. 2 shows that if we use only the reference query images to calculate the discriminant power ($\Psi(Q)$="query+R0") then when the number of dimension N decreases there is no MAP degradation. We propose next to use training images from the Web. For each query Q, $\Psi(Q)$ is composed of the union of the reference query images and of the first R result images ($\Psi(Q)$="query + Rx" where $x \in$ 5, 10 and 200). We obtain the best MAP when $N = 10$ and $R = 5$. Worstfeatures method means that only the N less discriminant features in average on all queries are always selected for each query. For example, for $N = 10$, the selected features are the 10 less frequent features in the distribution in the top left Fig. 4. In order to demonstrate the efficiency of query dependant selection, we also run a Bestfeatures method, selecting always the same N more frequently discriminant features. Feature selection by ALDA adapted to each query is better than Bestfeatures. To show the time reduction to use ALDA, we need more data, so we make the visual distances between the 131 query images and 100k synthetic visual vectors. In Tab. 1, we show that we could improve fusion MAP results (from 0.539 to 0.557) and as the same time reduce the time needed to calculate the visual distance (from 309 to 202 seconds).

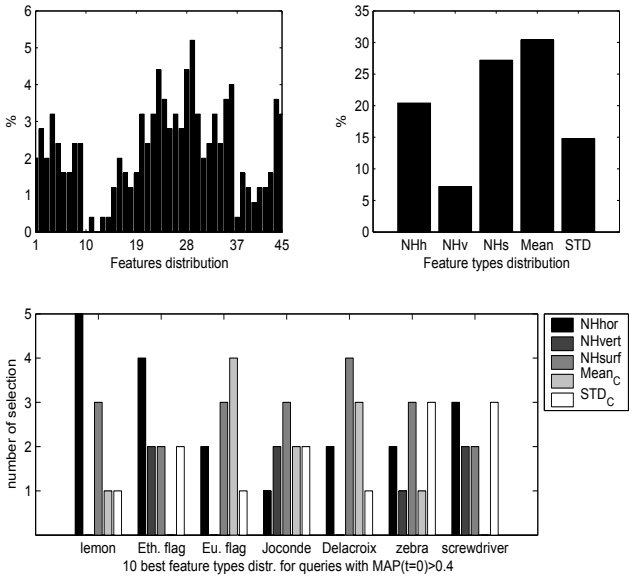

Fig. 4. Distributions of the first 10 most discriminant visual features of each query with ALDA running on "query + R5". Top left figures show features in the order that they are calculated (see algorithm section 3)

Theoretically, one could expect that the larger Ω is and the more it represents different concepts, the more the ALDA may be accurate. Actually we show in Fig. 3 the queries for which there is a significant MAP amelioration when R increases, versus other queries showing the contrary (others have no significant MAP variation). This may be due to the intrinsic concept "visualness" properties, which impact the quality of the web search engine results. Curves in Fig. 3 shows that average gain increases or decreases with R, showing two significant different kinds of query set. Set A could be said as a visual dependant concept, set B as higher level ones needing another information than visual features.

We analyse in Fig. 4 the query features selection in detail. First, in average on the whole query set we notice that $NHsurf$ and mean of the color are in average the best feature types. The second best feature type is $NHhor$, which is the entropy of the sum of the pixels across the lines of the image band. It is interesting to note that the usual STD color features are slightly less selected. The very poor selection rate of $NHvert$ can be explained by the fact that the sum across the whole image pixel line integrates on a too large domain (on the contrary of $NHvert$ which integrates pixels values on smaller band domain). We then show in Fig. 4 the 10 best features types distribution for the queries having a high MAP only with visual information ($t = 0, MAP > 0.4$). We clearly see strong selection variations between each query. The differences between Ethiopian and European flag come from their different orientations.

6. DISCUSSION AND CONCLUSION

As we showed [10] and tested on COREL experiments [11], optimal visual features are concept dependant. We proposed here an original Web based method to quickly estimate them from online mislabeled data. We show that the use of ALDA can improve the global MAP score on the 25 queries and as the same time divide by 15 the time processing. A more precise analysis shows that this method improves significantly MAP results for some concepts when R growths and, on the contrary, penalized other queries for high R.

MAP appears to be dependent on the number of features (N) and the number of training samples (R). However, the choice of these parameters would be very much dependent on the set of queries. It would be interesting to see how these parameters generalize. Experiments could be carried out by using a training set to choose the optimal parameters. This could be followed by evaluating the retrieval performance on a separate testing set and see how the performance changes. Further works will consist in defining if visual and/or textual ontologies could be useful to estimate which concept may have such properties.

7. REFERENCES

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

[2] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1075–1088, 2003.

[3] Kobus Barnard and David Forsyth, "Learning the semantics of words and pictures," in *IEEE ICCV*, 2001, vol. 2, pp. 408–415.

[4] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures.," in *NIPS*, 2003.

[5] ImagEVAL, "http://www.imageval.org," 2006.

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

[7] S. Tollari and H. Glotin, "Web image retrieval on imageval: Evidences on visualness and textualness concept dependency in fusion model," in *ACM CIVR*, 2007.

[8] L. Amsaleg, P. Gros, and S.-A. Berrani, "Robust object recognition in images and the related database problems," *Multimedia Tools and applications*, vol. 23, no. 3, pp. 221–235, 2004.

[9] F. Zuo, P. H. N. de With, and M. van der Veen, "Multistage face recognition using adaptative feature selection and classification," in *ACIVS2005, LNCS 3708*, 2005.

[10] H. Glotin, S. Tollari, and P. Giraudet, "Shape reasoning on mis-segmented and mis-labeled objects using approximated fisher criterion," *Computers and Graphics*, vol. 30, no. 2, April 2006.

[11] S. Tollari and H. Glotin, "LDA versus MMD approximation on mislabeled images for keyword dependant selection of visual features and their heterogeneity," in *IEEE ICASSP*, 2006.