

LEARNING OPTIMAL VISUAL FEATURES FROM WEB SAMPLING IN ONLINE IMAGE RETRIEVAL

Sabrina Tollari* and Hervé Glotin**

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

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

We optimize an image retrieval system using Web data and an online approximation of the Linear Discriminant Analysis (LDA) to select query dependant features. Results on a reference database show both significant improvement of the Mean Average Precision and reduction of the computation time.

1 ImagEVAL evaluation campaign

The task 2 of the campaign ImagEVAL 2006 (www.imageval.org) consists to retrieve Web images using textual and visual information. The corpus, extracted from the Web is composed of texts and 10K images extract from 700 urls. The goal of the task is to find all the images answering each of the 25 queries. A query q is composed of a set of keywords $\mathcal{K}(Q)$ and a set of few positive images $\mathcal{I}(Q)$ (Fig. 1).

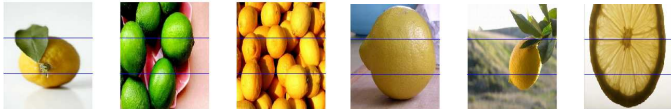


Fig. 1: The query 4 of ImagEVAL task 2 is composed of the keyword “lemon” and of 6 query images divided into 3 equal horizontal subbands, then 15 visual features are extracted from each subband: mean and std of normalized colors, entropy of the horizontal ($NHhor$), vertical ($NHvert$) and of the surface ($NHsurf$) color pixel distributions.

2 Feature selection on mislabeled data using web sampling

Most of available images (for example images in web pages) are mislabeled, i.e. there is no objective bijection between surrounding words and the concept in the images. In this context, we apply an Approximation of LDA (ALDA) [1], using additional Web training data. For each query, we split training data into 2 sets: $\Psi(Q)$ including positive images and Ω including all the training images. For each query Q and each visual feature X , we estimate the discriminant power $\hat{J}(X; Q)$:

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

where $\hat{B}(X; Q)$ is the between variance and $\hat{W}(X; Q)$ the within variance. We showed that feature ranking errors are small as long as enough samples are given in $\Psi(Q)$ and if Ω is very large [1].

Because query image sets are very small, we use a Web image search engine to use, as training samples, images matching with the keywords $\mathcal{K}(Q)$ of each query. Note that as the Web image search engines index images according to text information, the so built train set contains images which don’t visually correspond to Q .

To retrieve relevant images for each query, we first reduce visual vectors to their N most discriminant features. Then, test set images are sorted according to the geometric mean of their visual L2 distance to each query image. Second, we average visual and textual informations applying the text rate weight t to visual D_V and textual (tfidf) D_T distances: $D = t \times D_T + (1 - t) \times D_V$.

3 Results and Discussion

In [2] we concluded that the use of the visual/textual fusion to retrieve images gives better MAP scores than Text only ($t=100\%$) or than Visual only ($t=0\%$) (see Tab. 1) (we did not use feature selection).

Analysis of the query feature selection (Fig. 2) show that the Surface pixel distribution $NHsurf$ and the Mean of each color are the best features, before the Horizontal $NHhor$ feature (i.e. the entropy of the sum of the pixels across the lines of the image band). The poor selection rate of $NHvert$ may be due to its integration across large image lines (contrary to $NHvert$ features). The 10 best features types distribution shows a strong selection variations between each query.

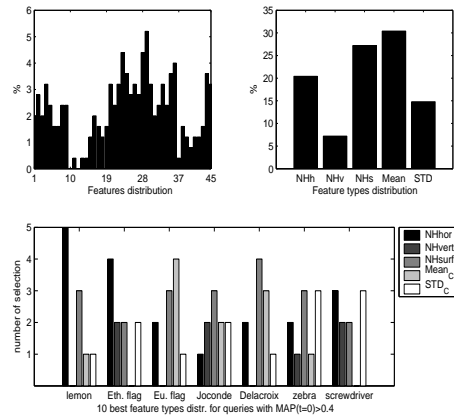


Fig. 2: Distributions of the first 10 most discriminant visual features for each query selected with ALDA using the query image set + 5 web images (features are ordered from band 1 to 3, and Horizontal, Vertical, Surface entropy, normalized R, G, B Mean and STD features)

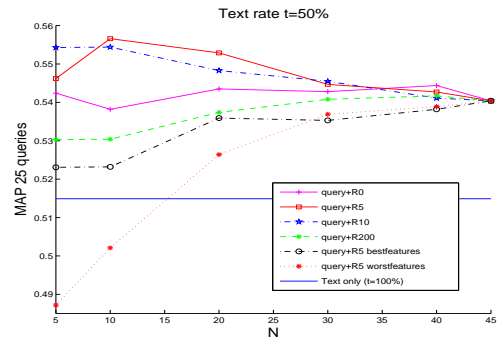


Fig. 3: MAP curves for $t=50\%$ according to different $\Psi(Q)$ set used to select the N most discriminant features by ALDA. All curves converge to the fusion MAP value without feature selection ($N=45$).

	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

Tab. 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

For all experiments, Ω is composed of sampled 17k Web images. We show (Fig. 3) that if $\Psi(Q) = \mathcal{I}(Q)$ (“query+R0”) then there is no MAP degradation when the number of dimension N decreases. Next, if $\Psi(Q)$ is composed of the union of $\mathcal{I}(Q)$ and of the first R Web training result images (“query + Rx ” where $x \in \{5, 10, 200\}$), we obtain the best MAP ($N = 10$ and $R = 5$). Moreover our adaptive ALDA selection takes the Bestfeatures selecting the N most discriminant features in average on all queries. Finally, Tab. 1 shows that we improve MAP jointly with the time needed to compute the visual distance.

[1] H. Glotin and S. Tollari and P. Giraudet, Shape reasoning on mis-segmented and mis-labeled objects using approximated Fisher criterion, Computers and Graphics, Special Shape Reasoning and Understanding, Elsevier, Vol30, N2, pages 177-184, 2006.
[2] S. Tollari, H. Glotin, Web Image Retrieval on ImagEVAL: Evidences on visualness and textualness concept dependency in fusion model, ACM Conference on Image and Video Retrieval (CIVR), pages 65-72, 2007