Keyword dependant selection of visual features and their heterogeneity for image content-based interpretation

Sabrina Tollari Hervé Glotin

Research report LSIS.RR.2005.003, 2005 (update 2006)

Contents

Abstract

To improve text-based image retrieval system, we propose to use visual content of images to filter their textual indexing. We propose first to generate new visual feature based on entropy measure (heterogeneity), and then we address the question of feature selection in the context of mislabeled images. We compare two methods of word dependant feature selection on mislabeled images: Approximation of Linear Discriminant Analysis (ALDA) and Approximation of Maximum Marginal Diversity (AMMD). A Hierarchical Ascendant Classification (HAC) as trained and tested using full or reduced visual space. Experiments are conducted on 10K Corel images with 52 keywords, 40 visual features and 40 new ones. We measure a classification gain of 59% and in the same time a reduction of 93% of the number of features.

1 Introduction

Query by example is a typical mode of request of image retrieval systems. User provides a query-image and the system searches for similar images based on a combination of low level features. But this multidimensional nearest neighbors search (NN-search) is not effective due to the high dimensional problem [2, 4]. Another mode is based on a textual indexing of image. This mode would be effective if the textual indexing of image (or classification) could be made automatically and correctly. In fact of the images are harshly classified by search engines on the Web: they are classified by words extracted from the same Web page, without any analysis content-based [5].

On the other hand, Content-Based Image Retrieval (CBIR) has attracted a lot of research interests in recent years [3, 12, 21, 18]. But most of the traditional techniques in CBIR are limited by the semantic gap between low level features (extracted from the images) and high level user's request. The increasing amount of data and diversity of the types of research contribute to widen this semantic gap. Moreover visual space is usually very high, as the number of keywords labeling objects in the image (also called “visems”), and only few training data with correctly full labeled visems are available as illustrated in Fig. 1.

In this paper we propose to automatically reduce the semantics gap by a pre-processing stage before classification or image indexation of mislabeled images, without any user intervention. Our strategy doesn't rely on a pre-selection of most difficult images to classify, but on the fact that visual features and keywords are strongly dependant. Therefore we first (i) generate higher order visual features based on entropy of all usual features (which generates a contextual visual analysis of visem), and secondly (ii) automatically optimize the visual space for each visem by selecting the most discriminant visual features. Thereby we produce a system which aims to give some solution to the problem claimed in [3]: “it remains an interesting open question to construct feature sets that (...) offer good performance for a particular vision task”. Our feature sets selection is based on an extension of Linear Discriminant Analysis (LDA) to the particular case of mislabeled data. We estimate the bias of this Approximation of LDA (ALDA). ALDA is applied before a Hierarchical Ascendant Clustering (HAC) to build visual clusters. Experiments conducted on COREL database (10K images, around 50 different keywords and three keywords by images) show that heterogeneity features are rich cues for the perceptual interpretation of ambiguous image and that great enhancement of classification results applying ALDA on combined usual and heterogeneity features. We then present the results of a filtering keywords model based on the HAC visual clusters, and we discuss in the final part of the expected enhancement using our visem specific feature subspaces.

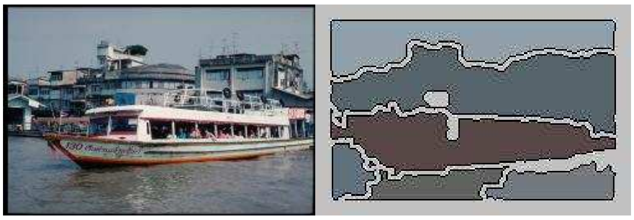


Figure 1: Example of automatic image segmentation by Normalized Cuts algorithm on Corel image labeled globally by *“water”, “boat”, “harbor”, “building”*}. It's difficult to know which blob may be labeled by a word. Moreover, a bijection between blobs and words is not possible.

In section 2, the paper presents first a short review of CBIR and feature selection systems. In section 3, we describe usual feature and new heterogeneity feature. In section 4, we present the features selection methods: LDA, approximated for mislabeled data. In section 5, we describe the corpus and show feature selection and HAC results for usual, heterogeneity and late fusion of feature space, and we apply our methods to keyword filtering. Finally, we discuss our results in the conclusion.

2 Features Selection and Semantic Gap Reduction

Recent approaches to reduce the semantic gap is the Active Learning (AL) asking the user to label some images closest to the classifier boundary. Support vector machine AL for image retrieval has been proposed by [19]. This approach treats the relevance feedback problem as a supervised learning problem. A binary classifier is learned by using all relevant and irrelevant labelled images as input training data. SVM classification method used in AL has been compared to Bayes and kNN classification methods in [10]. Authors claim that for category search in very large databases, efficient exploration process before classification process will become crucial. Unfortunately active learning requires a lot of manual user feedback, many hundreds for about only 10 visems [10], and therefore AL can't be applied to a large image data base with large visem lexicon.

On the other hand the most famous method of dimensionality reduction is Principal Components Analysis (PCA). This technique searches for directions in the data that have largest variances and subsequently project the data onto it. But PCA does not include label information of the data. For instance let imagine two cigar like clusters in 2 dimensions. If the cigars are positioned in parallel and very closely together, such that the variance in the total dataset, ignoring the labels, is in the direction of the cigars, then PCA for classification would be a terrible projection, because all labels get evenly mixed and we destroy the useful information. Although PCA finds components that are useful for representing data, there is no reason to assume that these components must be useful for discriminating between data in different classes.

Nevertheless, where PCA seeks directions that are efficient for representation, Linear Discriminant Analysis seeks directions that are efficient for discrimination ([6] page 117). In this paper we adapt LDA by approximation to real mislabeled general image database.

3 Visual features

Image processing are usually based on color, texture and shape features representing, rather roughly, major visual properties. Moreover, images are often segmented into regions (called 'blobs' that are in our paper automatically extracted by Normalized Cuts [17]; see Fig. 1). In this section we present the usual feature set we use, and we propose to generate from it a new one set motivated by psychovisual studies.

3.1 Major visual properties and usual features

Visual feature set are often chosen to be computable for any image region, and to be independent of any recognition hypothesis. As in [3], we use for each blob the 40 features listed below. Color is represented using the average and standard deviation of (R,G,B), (L,a,b), r=R/(R+G+B), g=G/(R+G+B). Texture is represented using the average and variance of 16 filter responses. We use 4 differences of Gaussian filters with different sigmas, and 12 oriented filters, aligned in 30 degree increments [17]. Shape is represented by the ratio of the area to the perimeter squared, the moment of inertia, and the ratio of the region area to that of its convex hull. Size is the image portion covered by the blob, and position is the coordinates of blob center of mass normalized by the image size. But as said in [3], it is not clear that these image features are canonical.

3.2 Heterogeneity of features

According to the experiments carried out in psychovision [14], heterogeneity criterion applied to surfaces has more or less impact in visual descriptions of objects. The value of the heterogeneity of the visual feature v_j of the image d contenting the b_p blobs is the entropy of the distribution of its probalized values $b_{p,j}$:

$$H_j = -\sum_{b_p \in d} b_{p,j} \times \log_2(b_{p,j}). \qquad (1)$$

In [13], heterogeneity is only defined on the area feature. Based on neurobiological studies [1], we propose in this paper to extend heterogeneity concept to all features. Recent advances in cognitive sciences claim that human interpretation is based on a contextual visual analysis. As pointed out in [1]: "This context-dependent transformation from image to perception has profound but frequently under-appreciated implications for neurophysiological studies of visual processing". Content-based image retrieval systems should take into account this context-based neuronal bases of visual scene perception. Red color can be discriminant for 'tomato', but it is much more the heterogeneities of color features that are discriminant for 'market'. Thus we extend the visual space applying heterogeneity to all normalized usual features.

4 Automatic word dependant feature selection on mislabeled data

Due to the high dimension problem [2, 4], a good visual indexing would be made up with the visual features which have the strongest discriminating capacities. To determine which are the most relevant visual features to annotate an image with a word is a difficult problem because available (mostly mislabeled) data do not correspond with traditional statistical methods requirements. Previous works showed

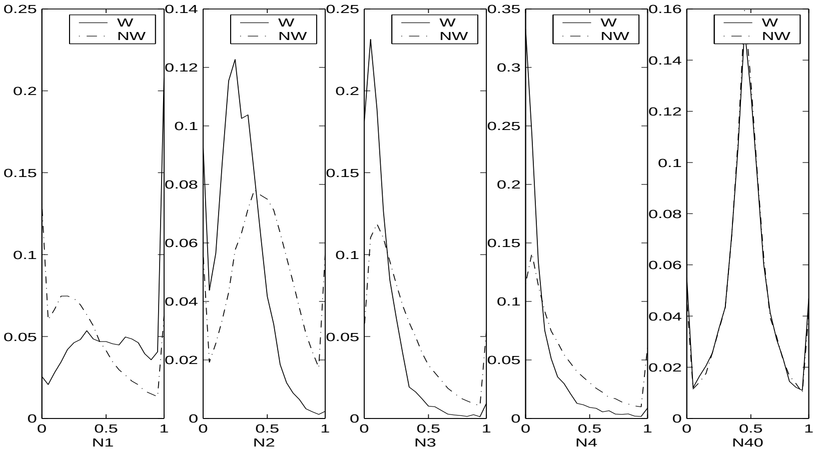


Figure 2: Conditional likelihoods $p(v_j|w_i)$ and $p(v_j|\neg w_i)$ of 5 visual features for WORD (W) versus NON-WORD (NW) approximated classes for keyword *"snow"*. Features are sorted from the best discriminative (N1) to the worst one (N40) (estimated by ALDA): N1 ('B' of RGB), N2 ('B' of LAB), N3 ('std A' of LAB), N4 ('std G' of RGS) and N40 ('3rd sigma texture'). We see clearly likelihood differences for discriminant features between W versus NW classes, and overlapping for N40.

that simple methods like LDA[1] (Linear Discriminant Analysis) or Maximum Marginal Diversity (MMD) [20] can discriminate acoustic [16] and visual features [22], but these methods were applied on well labeled corpuses describing a univocal relation between a conceptual class and a feature. The main difficulty for applying this kind of methods on large general images corpus is that they do not have a label for each blob, but a words set for an image (Fig. 1). We make however the following assumption: *if an image database presents each concept with a rather broad contextual variety, then LDA or MMD methods can estimate the N best discriminant features of each concept.* Thus, for each word, we build a bipartition of the training set: the class WORD of images which are labeled by this word and the class NON-WORD of images which are not labeled by it. Fig. 2 gives some features distributions obtained on WORD and NONWORD classes for *"snow"*.

4.1 Approximation of Linear Discriminant Analysis

Based on our two classes WORD and NONWORD, we calculate for each word w_i and for each visual dimension v_j, the between variance $\hat{B}(v_j;w_i)$ (average variance of each class) and the within variance $\hat{W}(v_j;w_i)$ (weighted average of each class variance). Finally, we calculate for each word w_i and each feature v_j the discriminant power $\hat{F}(v_j;w_i)$ defined by:

$$\hat{F}(v_j;w_i) = \frac{\hat{B}(v_j;w_i)}{\hat{B}(v_j;w_i) + \hat{W}(v_j;w_i)} \qquad (2)$$

This method, called ALDA (Approximation of LDA), has been theoretically and experimentally shown in [8] that ranking errors due to this approximation are small as long as enough samples are used and the considered concept is presented in various concepts.

4.2 Approximation of Maximum Marginal Diversity

However, LDA makes the assumptions that class densities are gaussian, that are unrealistic for most problems involving real data. The best feature set characterizing word class w_i should contain those feature with large marginal diversities. The marginal diversities $\hat{MD}(v_j;w_i)$ of feature v_j in class w_i is defined as the Kullback-Leibler divergence between $p(v_j|w_i)$ the class-w_i conditional probability density of v_j and $p(v_j|\neg w_i)$ the probability densities of class NONWORD:

$$\hat{MD}(v_j;w_i) = \sum p(v_j|w_i) log \frac{p(v_j|w_i)}{p(v_j|\neg w_i)}. \qquad (3)$$

4.3 Adaptive Features selection

To automatically determine the number of best features to discriminate each word as well as possible,

[1] Whereas PCA seeks direction that are efficient for representation, LDA seeks ones that are efficient for discrimination ([6] p.117).

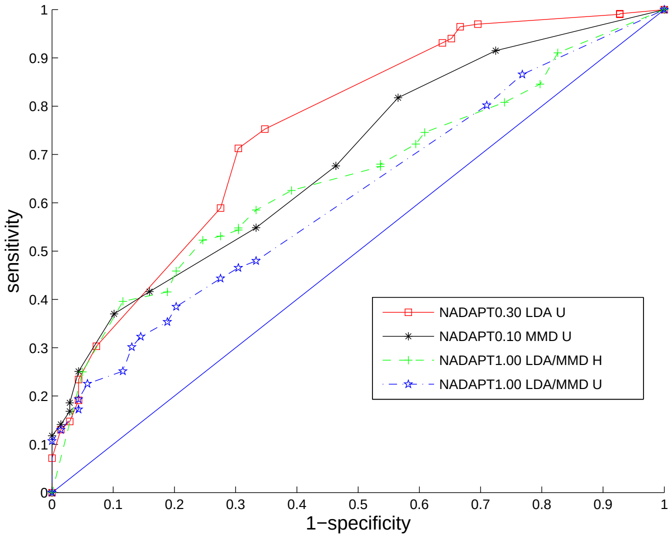


Figure 3: ROC curves of the HAC image classification for word *"woman"* applied for various methods to usual (U) and heterogeneity (H) features on DEV set. Between two points of the curve, 5% of the closest blobs are aggregated by HAC.

we propose to choose the N most discriminating ones which cumulate τ per cent of the total sum of the discriminating capacities over all the δ features for this word (method 'NADAPTτ'). We sort discriminant powers $\hat{DP}$ ($= \hat{F}$ or $\hat{MD}$) by descending order, then the system choose N such that:

$$\sum_{j=1}^{N} \hat{DP}(v_j; w_i) = \tau \sum_{j=1}^{\delta} \hat{DP}(v_j; w_i). \qquad (4)$$

5 Word dependant visual clustering

5.1 Clustering

In this first stage, we will associate a set of visual clusters $\mathcal{C}(w_i) = \{C_1(w_i), C_2(w_i), \cdots, C_K(w_i)\}$ to each word w_i of the lexicon. We define a visual cluster $C_k(w_i)$ as an hyperrectangle in the visual multidimensional space. To build the visual clusters of a word, we will seek out grouping of training blobs in visual space using a Hierarchical Ascendant Classification (HAC) [11] (not-supervised construction of clusters) with nearest neighbor aggregation criterion. The principle of this HAC is to gather the blobs having a weak distance in multidimensional space.

For each word w_i, we build a subset $\mathcal{A}(w_i)$ made with images d of the training set $\mathcal{A}$ labelled by the word w_i:

$$\mathcal{A}(w_i) = \{d | w_i \in W_{Ref}(d) \ and \ d \in \mathcal{A}\}. \qquad (5)$$

Figure 4(a) shows an example of training images labelled by the same word. On $\mathcal{A}(w_i)$, we carry out an HAC. Then we determine the level of the HAC by choosing the one giving best score on a development set (score calculation is explained section 5.3). We then keep only clusters which contain a significant number of blobs. Thus we associate visual clusters to word w_i.

Each class $C_k(w_i)$ is represented by a couple of same dimension vectors:

$$(\bar{C}_k(w_i), \vec{\sigma}(C_k(w_i))) \qquad (6)$$

where $\bar{c}_k(w_i)$ is the centroid vector of the visual cluster in multidimensional space and $\vec{\sigma}(c_k(w_i))$ is the vector of the standard deviations of the class for each dimension of space.

Let us notice that visual clusters of a word are disjointed, because no blob can belong to two clusters of the same word:

$$\forall k \neq \ k' \ C_k(w_i) \cap C_{k'}(w_i) = \emptyset \qquad (7)$$

what can be interpreted like a word can have several visual sens. For example, the sun can be yellow, but it is also often red at sunset. Let us also notice that two different words can have visual clusters not disjointed:

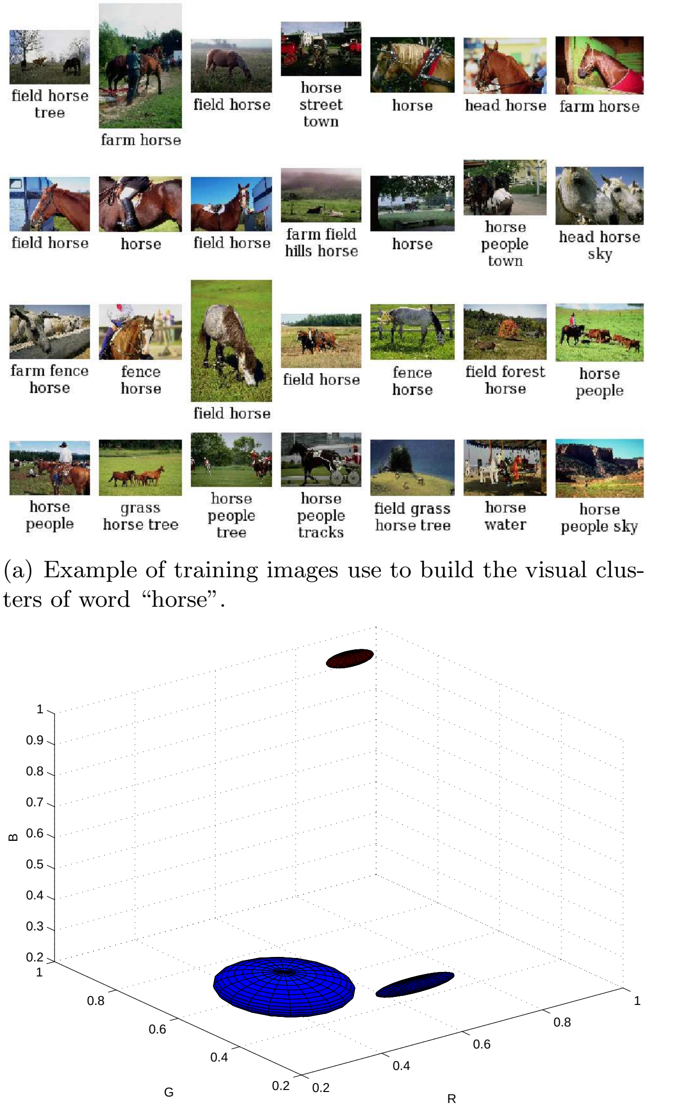
$$\exists \ k, k' \ and \ w_i, w_i' \ C_k(w_i) \cap C_{k'}(w_i') \neq \emptyset. \qquad (8)$$

Indeed, two words can have very close visual. For example, words: *"human"*, *"man"*, *"woman"*, *"child"* can have very close visual because of the presence of texture representing the skin (see also figures 4(a) and 4(b) for concrete example).

5.2 Quality evaluation of word dependant visual clusters

We have just described a method allowing to associate visual clusters to words, we will now evaluate the quality of this association in order to determine for each word the best value for stopping the HAC. For that, we use a test set. For each word, we classify initially the blobs images of test, then the images of test, in visual space. Lastly, we calculate the score.

By definition, a blob b_j of a test image belongs to a visual cluster $C_k(w_i)$ if the visual vector of b_j is



(a) Example of training images use to build the visual clusters of word "horse".

(b) The three visual clusters of word "horse" obtained by 40DIM method, represented in RGB space. Clusters are actually hyperrectangles.

Figure 4: Training examples and visual clusters for "horse".

in the hyperrectangle cluster $C_k(w_i)$, in other words if for each visual dimension p, the value of the visual vector of the blob b_j for dimension p is at a distance of the value of the centroid of cluster $C_k(w_i)$ for dimension p lower than the value of the standard deviation for this dimension multiplied by a constant:

$$b_j \in C_k(w_i)\ iif\ \forall p\ |\bar{C_{k,p}}(w_i) - \vec{V_p}(b_j)| \leq \alpha \times \vec{\sigma_p}(C_k(w_i)) \quad (9)$$

where α is an optimized constant. If we supposed that the distribution of blobs in a visual cluster built in training stage, follows a normal law then for $\alpha = 2$ we let us know that 95% of the individuals are in the cluster ([6] page 33). In our case, we do not have a normal law distribution, however we will suppose that $0 < \alpha \leq 4$, and we optimize α on development set (see ROC curves in figure 3).

For each blob b_j of test, we associate the word w_i to b_j if b_j belongs to one of the visual clusters of this word:

$$w_i \in W_{Sys}(b_j)\ iif\ \exists k\ such\ as\ b_j \in C_k(w_i) \quad (10)$$

where $W_{Sys}(b_j)$ is words set associated by the system to blob b_j. As the visual clusters of a word are disjointed, a blob belongs to no more than one cluster. Thus if there exists k such as $b_j \in C_k(w_i)$ then it is the only one. Thus $|\{k|b_j \in C_k(w_i)\}| = 1$ if the word w_i is associated to b_j, 0 if not. So $|\{k|b_j \in C_k(w_i)\}| = 1$ if word w_i is associated to b_j, 0 else. Finally, we assume that word w_i is associated to a test image d if this word is associated to at least β blobs of this image:

$$w_i \in W_{Sys}(d)\ iff\ \sum_{b_j \in d} |\{k|b_j \in C_k(w_i)\}| \geq \beta \quad (11)$$

where $W_{Sys}(d)$ is the word set associated by the system to image d and β is a constant lower or equal to the minimal number of blobs of an image. This constant is dependent on the number of blobs in an image. Indeed, more there are blobs in an image and more the constant β will be large, because a word may correspond to several blobs of the image. However, we set $\beta = 3$ for usual features experiments and $= 1$ for heterogeneity ones.

5.3 Scoring

Each test image is initially labelled by a words sets $W_{Ref}(d)$. Thus we can calculate the rates of sensitivity and specificity. We also use the score "Normalized Score" (noted NS thereafter) employed in [3, 15].

Sensitivity (also called recall) is the number of relevant documents found among the number of relevant documents. We use it to measure the number of test images associated with w_i by the system and initially labelled by w_i. Specificity is the number of not relevant not found documents among the number of not relevant documents. The score NS is the sum of the sensitivity and specificity less 1:

$$NS = sensitivity + specificity - 1 \quad = \frac{right}{n} - \frac{wrong}{N-n} \tag{12}$$

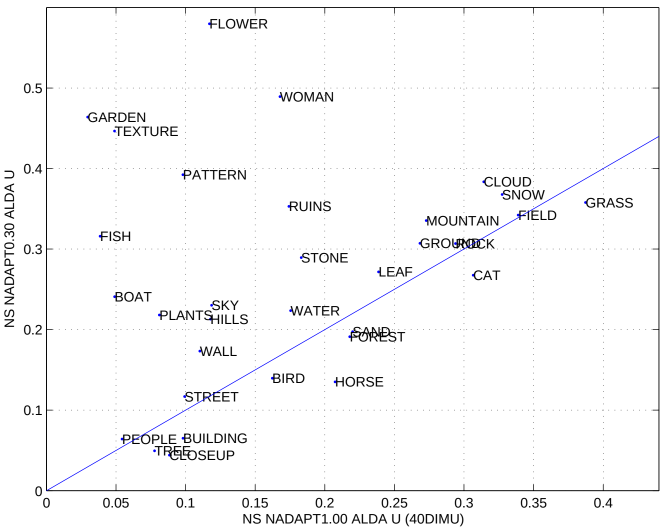
where $right$ is the number of found images labeled with w_i, $wrong$ is the number of found images not labeled by w_i, N is the total number of test images and n is the number of test images labelled with w_i. Let us notice that $-1 \leq NS \leq 1$. Score is 1 when the system finds the n words of references and none of the other words, -1 when it only finds the words which are not references, 0 when all the words of the lexicon are found. The gain measures given in our experiments is for method M:

$$gain(M) = (NS(M) - NS(40DIMU)/(NS(40DIMU)) \tag{13}$$

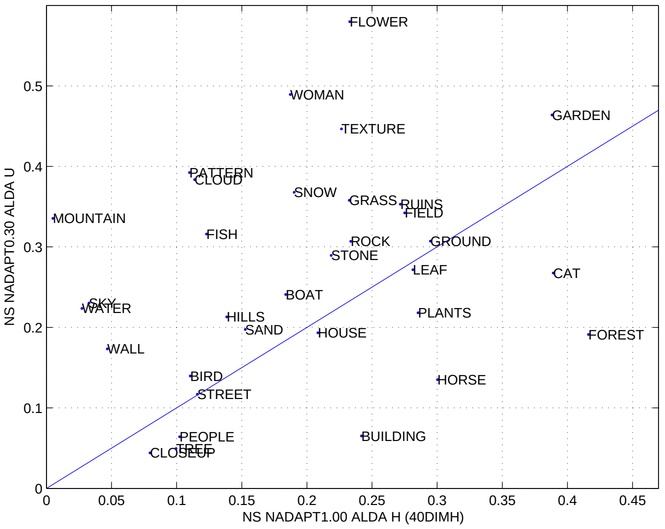
6 Experimental results

6.1 Corpus

We use the same data as in [3]. Experiments are made on Corel images database made of various 10K images, approximately 100 000 segments (called 'blobs') are preprocessed in [3] by 'Normalized Cuts' algorithm [17]. This segmenter has the occasional tendency to produce small, typically unstable regions. We keep the 10 largest regions in each image by computing, for each region, a set of 40 features described in section 3.1. Then we calculate the heterogeneity for each visual feature. In order to avoid artifact, we normalize both usual (U) and heterogeneity (H) vectors in 90% of their MLE Gamma distribution. Finally, each blob is represented by a vector of 80 dimensions where each component is in $[0, 1]$. Features pdf in eq. 3 are estimated by $\sqrt{256}$ bins histograms on TRAIN set. Each image of Corel is manually labeled by an average of 3.6 words from a lexicon of 250 different words. We choose to study in this article only the 52 keywords having more than 60 occurrences in our training set. The corpus is split by chance in a training set (TRAIN) of 5000 images, a development set (DEV) of 2500 images and a test set (TEST) of 2500 images.



(a) Comparison of the normalized scores (NS) obtained on TEST set for 40DIMU (NADAPT1.00 U ALDA) and NADAPT0.30 U ALDA methods. Some words are more discriminated by heterogeneity (H) features than by usual (U) ones.



(b) Comparison of the normalized scores (NS) obtained on TEST set for NADAPT0.30 ALDA U and NADAPT1.00 ALDA H methods. Some words are more discriminated by heterogeneity (H) features than by usual (U) ones.

Figure 5: Comparison of normalized scores.

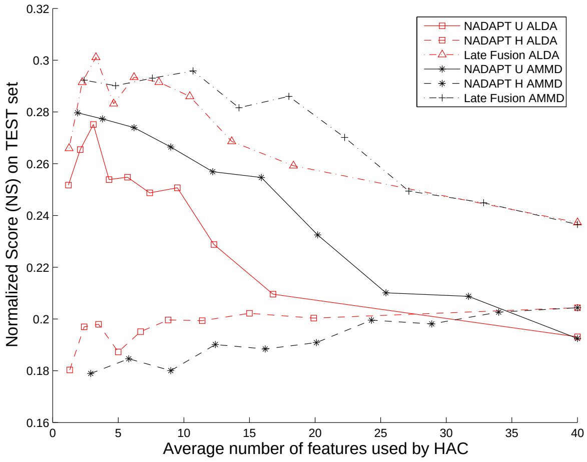


Figure 6: Averaged classification NS, over 52 keywords and 2500 TEST images, in function of the average number of features used. Each dot is for $\tau = 10\%$ to 100% (left to right). NADAPT U AMMD/ALDA both naturally converge to the reference model U on usual feature without feature selection (τ=1.00).

6.2 Word clustering by HAC

To model the association between visual features for a given word, we build visual clusters by Hierarchical Ascendant Classification (HAC) as explained in section 5. For each word, we cluster by HAC the visual vectors – reduced to the N best dimensions chosen by ALDA or AMMD – of TRAIN images labeled by this word. Each visual cluster is represented by mean and standard deviation vectors. Finally, the system indexes an image by a word if at least three blobs ($\beta = 3$ for U) (resp. one blob $\beta = 1$ for H) of the image are in one of the visual clusters of this word. Each DEV and TEST image is initially labeled by a words set, thus we calculate the Normalized Score (NS). We optimize parameters (clusters sizes, α) on DEV set maximizing NS (see Fig. 3).

6.3 Selection and fusion results on TEST

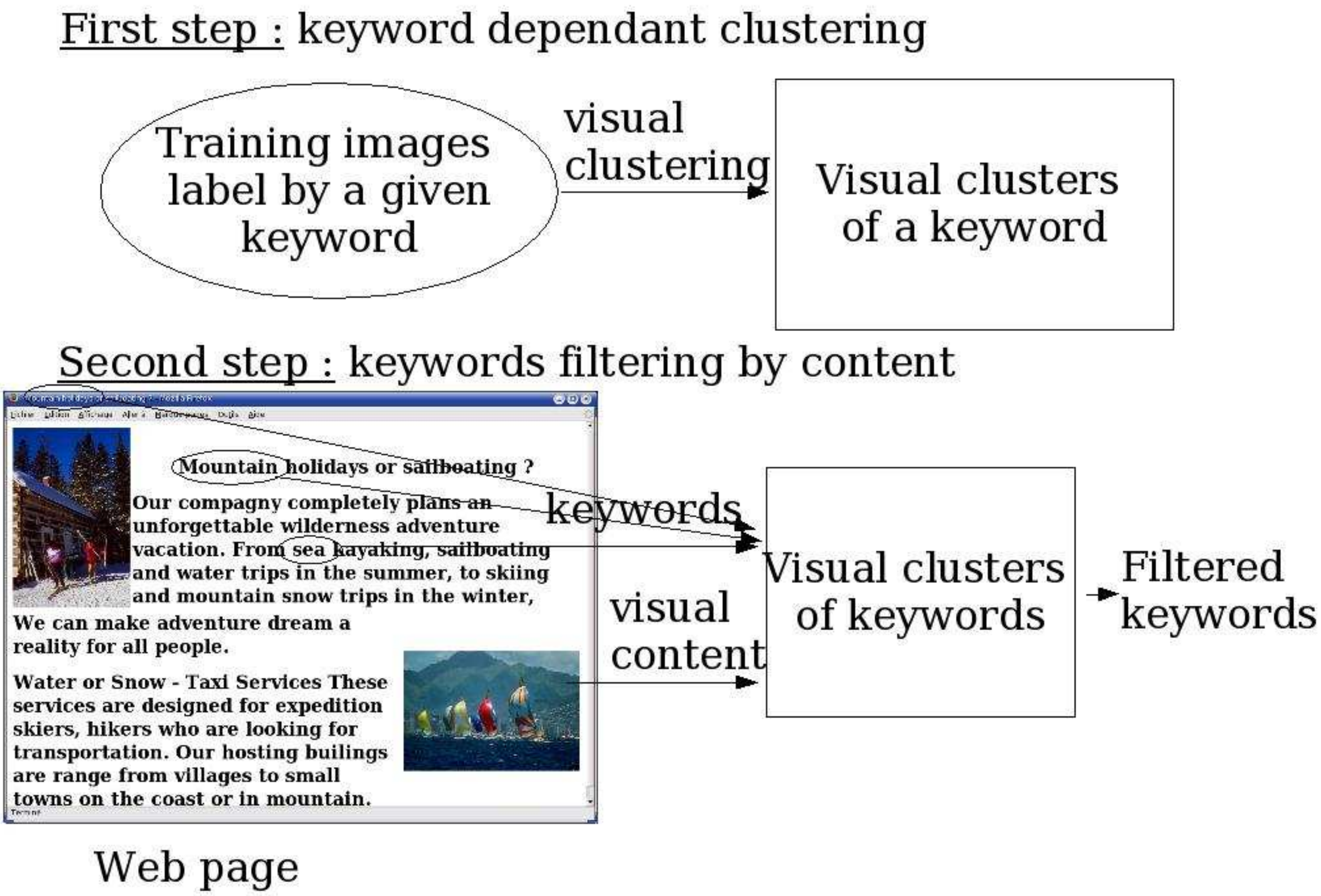
The feature selection results for 2500 TEST images and 52 keywords are shown in Fig. 6 and Tab. 1. ALDA feature selection (methods NADAPT and NBESTτ) reduces space dimension and increases score classification with U features. In Fig. 5(a), we compare, for each word, the normalized scores of 40DIMU (classification on U without feature selection) and the best NADAPT method (NADAPT0.30 U): most of the words are better discriminated with feature selection. In average (Fig. 6), classification on heterogeneity features give worse results than usual ones. In addition, feature selection on H (NADAPT H) give worse results than 40DIMH (classification on H without feature selection), it might be due to the lack of samples with H (only one vector by image). However, as shown in Fig. 5(b), some words are better discriminated with H than with U, so H could be useful to improve image classification of some concepts. Thus we propose, for Late Fusion, to learn for each word on DEV set, which one of U or H spaces maximizes NS. The Late Fusion curve in Fig. 6 show a significant improvement. We propose then to learn on DEV set for each word which τ gives best results (methods called Bestτ in Tab. 1). Results on TEST set show an improvement of +69% compared to 40DIMU.

7 Keyword Filtering

On the Web, images are associated to keywords among which some are relevant for the image and others not. In [5], they study the importance of title web page, HTML tags, surrounding text passages to retrieve efficiently images, but they couldn't check the visual relevance of words to images. We wish to filter words according to their visual relevance to the image. Unfortunately, we do not have image databases allowing to validate this filtering. Moreover, one would need validation of filtering by users. We thus propose another method: we suppose that reference words of an image are the relevant words for this image and others words of lexicon are the not relevant ones. We then use the visual clusters obtained previously by HAC of training blobs and optimised with DEV set, for filtering all the words of the lexicon for TEST image (see Fig. 8). This method could be used for image annotation [7], but our system is build for filtering existing keywords, and not for predicting keywords.

For each image of TEST, we calculate the NS score (described part 5.3) by taking *right* as the number of reference words associated to the image by the system, and *wrong* the number of words which were not reference words and were associated nevertheless to the image by the system. We make then the average of NS scores obtained for all images of TEST (see FILTERING Tab. 1).

The magnitude orders of results for classification and for filtering are almost the same ones except for NADAPTBestτ H where the results of filtering are better. It might be due to the lack of samples with H, and so each words is better discriminated for a different value of τ. This results are encouraging to use H features. Finally, we measure a filtering improve-


First step : keyword dependant clustering
Training images label by a given keyword
visual clustering
Visual clusters of a keyword
Second step : keywords filtering by content
keywords
visual content
Visual clusters of keywords
Filtered keywords
Web page

Mountain holidays or sailboating ?

Our compagny completely plans an unforgettable wilderness adventure vacation. From sea kayaking, sailboating and water trips in the summer, to skiing and mountain snow trips in the winter, We can make adventure dream a reality for all people.

Water or Snow - Taxi Services These services are designed for expedition skiers, hikers who are looking for transportation. Our hosting builings are range from villages to small towns on the coast or in mountain.

Figure 7: Schema of the keyword filtering system

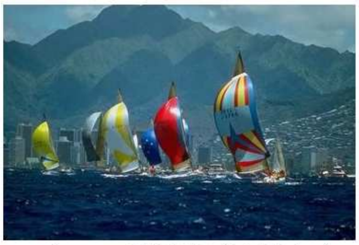


Image 172052 (10 blobs)
Label (3 on 3)
water(OK) mountain(OK) coast(OK)
sensi=1.00 specif=0.65
preci=0.15 NS=0.65

20 words associated by the system
desert(7) water(6) sky(6) wave(6) hills(6) closeup(6) mountain(6) coast(6) tree(6) beach(6) boat(5) branch(5) temple(5) sand(4) forest(4) cloud(4) people(4) fish(4) horizon(3) valley(3)

32 words non associated by the system
snow(2) statue(2) flower(1) head(1) building(1) window(1) woman(1) street(1) plants(1) field(1) vegetable(1) rock(1) bird(1) wall(1) baby cat cougar food fungus garden grass ground horse house ice leaf mushroom ocean pattern ruins stone texture

Total: 52 words

Figure 8: Example of keyword filtering of an image with the visual clusters build by NADAPT0.30 U ALDA method

ment up to 39% compare to filtering without feature selection.

8 Discussion and conclusion

Large image retrieval systems require, on the one hand, to be able to index images with few visual features and the most relevant words and, on the other hand, to quickly search a concept with some specific visual characteristics. We demonstrate in this article that ALDA and AMMD methods can determine for each word which are the most relevant visual features, and thus attenuate significantly the number of visual features necessary to nearest neighbor searches, and consequently reduce the high dimensional problem. Moreover, heterogeneity features (H) which is a fast indexing can discriminate some concepts better than usual ones (U). The late fusion of U and H allows a reduction of the visual space to the 3 most discriminating features, while improving scoring up to +59%. All parameters are learned only once, so the indexing and research of new images on the visual spaces obtained by fusion is very fast. We wish thereafter to work on Web images to build an image search engine dealing with texts and images, however Web images are badly annotated. Anyway we have just shown that ALDA and AMMD methods deal with mislabeled data [9]. These methods will thus permit us to learn which are the most relevant features to associate keywords to Web images, then to effectively filter these keywords according to the visual contents of the images.

References

[1] T. D. Albright. Why do things look as they do?: Contextual influences on visual processing. *Journal of Vision*, 2(10):60–60, 12 2002.

[2] L. Amsaleg, P. Gros, and S.-A. Berrani. Robust object recognition in images and the related database problems. *Multimedia Tools and applications*, 23(3):221–235, 2004.

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

[4] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is “nearest neighbor” meaningful? In *Proc. of ICDT*, pages 217–235. Springer-Verlag, 1999.

[5] T. A. S. Coelho, P. P. Calado, L. V. Souza, B. Ribeiro-Neto, and R. Muntz. Image retrieval using multiple evidence ranking. In *IEEE Knowledge and Data Engineering*, pages 408–417, 2004.

[6] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley and Sons, Inc., 2000.

[7] H. Glotin and S. Tollari. Fast image auto-annotation with visual vector approximation clusters. In *Proc. of 4th Content-Based Multimedia Indexing (CBMI)*, Riga, June 2005.

[8] H. Glotin, S. Tollari, and P. Giraudet. Approximation of linear discriminant analysis for word dependent visual features selection. In *Proc. of Advanced Concepts for Intelligent Vision Systems (ACIVS2005), LNCS 3708-Springer*, pages 170–177, Sept. 2005.

[9] 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.

[10] P. H. Gosselin and M. Cord. A comparison of active classification methods for content-based image retrieval. In *ACM Workshop on Computer Vision Meets Databases (CVDB)*, pages 51–58, Paris, 2004.

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

[12] 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.

[13] J. Martinet. *A relational vectorial information retrieval model adapted to images.* PhD thesis, Université Joseph Fourier, Grenoble, 2004.

[14] J. Martinet, Y. Chiaramella, and P. Mulhem. A model for weighting image objects in home photographs. In *ACM CIKM’05*, pages 760–767, Bremen, Germany, 2005.

		DIMENSION		CLASSIFICATION				FILTERING			
		aver.	reduc.	aver.	aver.	aver.	gain	aver.	aver.	aver.	gain
method	τ	N	%	recall	precision	NS	%	recall	precision	NS	%
40DIMU (ref.)	1.00	40	-	0.630	0.061	**0.192**	-	0.667	0.085	**0.229**	-
40DIMH	1.00	40	+0	0.387	0.058	0.204	+6	0.715	0.098	0.239	+4
NADAPT H ALDA	Bestτ	9.4	-77	0.635	0.062	0.211	+9	0.704	0.103	**0.286**	**+25**
NADAPT H AMMD	Bestτ	6.7	-83	0.625	0.062	0.204	+6	0.677	0.104	0.262	+14
NADAPT U ALDA	0.30	3.1	-92	0.692	0.077	0.275	+43	0.711	0.094	0.297	+30
NADAPT U ALDA	Bestτ	7.8	-81	0.622	0.083	0.293	+52	0.631	0.100	**0.304**	**+33**
NADAPT U AMMD	0.10	1.9	-95	0.627	0.090	0.280	+45	0.631	0.094	0.286	+25
NADAPT U AMMD	Bestτ	3.6	-91	0.608	0.091	**0.305**	**+58**	0.588	0.102	0.287	+25
Late Fusion ALDA	0.30	3.3	-92	0.710	0.072	0.301	+56	0.723	0.095	**0.318**	**+39**
Late Fusion ALDA	Bestτ	8.4	-79	0.669	0.080	**0.325**	**+69**	0.659	0.099	**0.317**	**+38**
Late Fusion AMMD	0.40	10.7	-73	0.660	0.079	0.296	+54	0.670	0.100	0.310	+35
Late Fusion AMMD	Bestτ	4.5	-89	0.650	0.084	**0.323**	**+68**	0.635	0.099	0.310	+35

Table 1: Best results for each method compared to the reference classification U on usual feature without selection (40DIMU) for 52 words and 2500 TEST images.

[15] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *ACM Multimedia*, pages 275–278, 2003.

[16] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri. Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop. In *IEEE Workshop Multimedia Signal Processing*, pages 619–624, Cannes, 2001.

[17] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[18] 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.

[19] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *ACM Multimedia 2001*, 2001.

[20] N. Vasconcelos. Feature selection by maximum marginal diversity: optimality and implications for visual recognition. In *Proc. of IEEE ICIP*, 2003.

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

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