# Approximation of Linear Discriminant Analysis for Word Dependent Visual Features Selection

Hervé Glotin Sabrina Tollari Pascale Giraudet

Université du Sud Toulon-Var
UMR CNRS 6168 LSIS

Research report LSIS.RR.2005.002

2005

## Contents

## Abstract

To automatically determine a set of keywords that describes the content of a given image is a difficult problem, because of (i) the huge dimension number of the visual space and (ii) the unsolved object segmentation problem. Therefore, in order to solve matter (i), we present a novel method based on an Approximation of Linear Discriminant Analysis (ALDA) from the theoretical and practical point of view. Application of ALDA is more generic than usual LDA because it doesn't require explicit class labelling of each training sample, and however allows efficient estimation of the visual features discrimination power. This is particularly interesting because of (ii) and the expensive manually object segmentation and labelling tasks on large visual database. In first step of ALDA, for each word $w_k$, the train set is split in two, according if images are labelled or not by $w_k$. Then, under weak assumptions, we show theoretically that Between and Within variances of these two sets are giving good estimates of the best discriminative features for $w_k$. Experimentations are conducted on COREL database, showing an efficient word adaptive feature selection, and a great enhancement (+37%) of an image Hierarchical Ascendant Classification (HAC) for which ALDA saves also computational cost reducing by 90% the visual features space.

## Keywords

Feature selection, Fisher LDA, visual segmentation, image auto-annotation, high dimension problem, word prediction, CBIR, HAC, COREL database, PCA.

# 1 Introduction

The need for efficient content-based image retrieval has increased in many application areas such as biomedicine, military, and Web image classification and searching.

Many approaches have been devised and discussed over more than a decade. While the technology to search text has been available for some time, the one to search images (or videos) is much more challenging. Most of image content based retrieval systems require the user to give a query based on image concepts, but in general he asks semantic queries using textual descriptions. Some systems aim to enhance image word research using visual information [13]. Anyway, one needs a fast system that robustly auto-annotates large un-annotated image databases. The general idea of image auto-annotation systems is to associate a class of 'similar' images with semantic keywords, e.g. to index by few keywords a new image according to a reference train set. This problem has been pursued in various approaches, such as neural networks, statistical classification, etc. One major issue in these models is the huge dimension number of visual space, and "it remains an interesting open question to construct feature sets that (...) offer very good performance for a particular vision task" [1].

Some recent works consider user feedback to estimate the most discriminant features. This exploration process before or during classification, like in Active Learning, requiers a lot of manual interactions, many hundreds for only 10 words [6]. Therefore these methods can't be applied to large image databases or large lexicons. In this paper we propose to answer to the previous question by automatically reducing the high dimensional visual space to the most efficient usual features for a considered word. The most famous method of dimensionality reduction is Principal Components Analysis (PCA). But PCA does not include label information of the data. 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. But where PCA seeks direction that are efficient for representation, Fisher Linear Discriminant Analysis (LDA) seeks ones that are efficient for discrimination ([3] pp 117).

Indeed recent works in audio-visual classification show that LDA is efficient under well labelled databases to determine the most discriminant features, reducing the visual space [4, 10, 7]. Unfortunately, most of the large image databases are not correctly labelled, and do not provide a one-to-one relation between keywords and image segments (see COREL image sample with their caption in Fig. 1). Consequently usual LDA can't be applied on real image databases.

Moreover because of the unsolved visual scene segmentation problem (see Fig. 1), real applications or training of image auto-annotation systems from web pages, would require a robust visual features selection method from uncertain data. Therefore, we present a novel Approximation of LDA (ALDA), in a theoretical and practical analysis. ALDA is simpler than usual LDA, because it doesn't need explicit labelling of the training samples for generating a good estimation of the most discriminant features. ALDA first stage consists, for each word $w_k$, to split train set in two, according if images are labelled by $w_k$ or not. Then, under weak assumption, we show that for a given $w_k$, Between and Within variances, between these two sets, are giving good estimates of the best discriminative features. Experimentations are illustrating features dependency to each word, and significant classification enhancements.

# 2 LDA approximation and adaptive visual features

Major databases are not manually segmented and segment-labelled, thus given a set of training images $\Phi = \{\phi_j\}_{j\in\{1,..,J\}}$ and a lexicon $\lambda = \{w_k\}_{k\in\{1,..,K\}}$, each image $\phi_j$ is labelled with some words of $\lambda$ (e.g. $\phi_j$ has a global legend constructed with $\lambda$ as shown in Fig. 1). In order to extract visual features of each object included in each $\phi_j$, one can automatically segment each image in many areas called blobs. Unfortunately, blobs generally do not match with the shape of each object.

Even if they do, there is no way to relate each blob to the corresponding word.

Nevertheless, we show below that despite the fact that each word class $w_k$ is not associated to a unique blob, and vice-versa, one can estimate for each $w_k$ which are the most discriminant visual features.

To this purpose we need to define four sets: $S$, $T$, $T_G$ and $G$. Let be $S$ the theoretical set of values of one feature $x$, calculated on all the blobs that are *exactly* representing the word $w_k$. We note for any feature set E, $c_E$ its cardinal, $\mu_E$ the average of all $x_i$ values of $x \in E$, $v_E$ their variance. Let be $T$ the set of $x$ values *of all blobs included in all images labelled* by $w_k$ (of course $T$ includes $S$). Let be $T_G$ such that $T = T_G U S$, with empty intersection between $T_G$ and $S$. We assume $c_{T_G} \neq 0$ (otherwise each image labeled by $w_k$ contains only the corresponding blobs).

Let be $G$ the set containing all values of $x$ from all blobs contained in images that are not labelled by $w_k$. In the following, we only assume the weak assumption (hyp. 1) $\mu_{T_G} = \mu_G$ *and* $v_{T_G} = v_G$, which is related to the simple assumption of context independency provided by any large enougth image database. We note $B_{DE}$ (resp. $W_{DE}$) the Between variance (resp. the Within variance) between any sets D and E. The usual LDA is based on the calculation, for each feature $x$ of the theoretical discrimination power $F(x; w_k) = \frac{1}{1+V(x;w_k)}$ where $V(x; w_k) = \frac{W_{SG}}{B_{SG}}$. We show below that $\hat{V}(x; w_k) = \frac{W_{TG}}{B_{TG}}$ is a good approximation of $V(x; w_k)$, and that if one apply $V$ to ordinate all $x$ for a given word $w_k$, then this order is the same by applying $\hat{V}$, at least for the most discriminant features $x$. Therefore the selection of features whith higher theoretical discriminative powers $F$ can be carried out from the calculation of practical $\hat{F}(x; w_k) = \frac{1}{1+\hat{V}(x;w_k)}$ values.

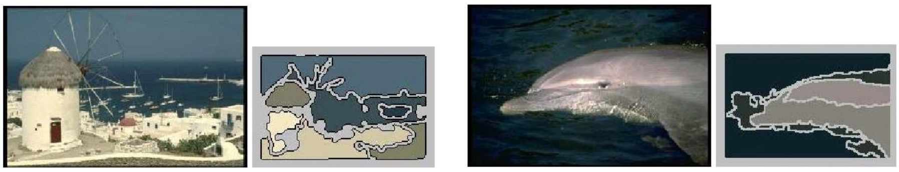

Figure 1: Examples of an automatic segmentation (Normalized Cuts algorithm [11]) of two COREL images [1]. Image caption are (left image) "Windmill Shore Water Harbor" and (right) "Dolphin Bottlenosed Closeup Water". Each blob of each image is labelled by all words of its image caption. Notice also that dolphin is split in two parts as many as other objects after the Normalized Cuts algorithm.

Let $p_S = \frac{c_S}{c_T}$ and $q_S = 1 - p_S = \frac{c_T - c_S}{c_T} = \frac{c_{T_G}}{c_T}$. We have $\mu_T = q_S.\mu_{T_G} + p_S.\mu_S$. Therefore:

$$\mu_T = q_S.\mu_G + p_S.\mu_S. \tag{1}$$

Let derive $v_T$ with $v_S$, $v_G$, and for any $x \in T$, the probability $p_i$ of event '$x = x_i$':

$$\begin{aligned}
v_T = \sum_{x_i \in T} \Big(x_i - \mu_T\Big)^2 p_i \quad &= \sum_{x_i \in T} \Big(x_i - q_S.\mu_G - p_S.\mu_S\Big)^2 p_i \\
= \sum_{x_i \in T_G} \Big((x_i - \mu_G) + p_S(\mu_G - \mu_S)\Big)^2 p_i &+ \sum_{x_i \in S} \Big((x_i - \mu_S) + q_S(\mu_S - \mu_G)\Big)^2 p_i \\
= \sum_{x_i \in T_G} (x_i - \mu_{T_G})^2 p_i + 2p_S(\mu_G - \mu_S) &\sum_{x_i \in T_G} (x_i - \mu_G) p_i + p_S^2(\mu_G - \mu_S)^2 \sum_{x_i \in T_G} p_i \\
+ \sum_{x_i \in S} (x_i - \mu_S)^2 p_i + 2q_S(\mu_S - \mu_G) &\sum_{x_i \in S} (x_i - \mu_S) p_i + q_S^2(\mu_S - \mu_G)^2 \sum_{x_i \in S} p_i \\
= q_S.v_{T_G} + 2p_S(\mu_G - \mu_S)\Big(\sum_{x_i \in T_G} x_i.p_i &- \mu_G \sum_{x_i \in T_G} p_i\Big) + p_S^2(\mu_G - \mu_S)^2 q_S \\
+ p_S.v_S + 2q_S(\mu_S - \mu_G)\Big(\sum_{x_i \in S} x_i.p_i &- \mu_S \sum_{x_i \in S} p_i\Big) + q_S^2(\mu_S - \mu_G)^2 p_S \\
= q_S.v_G + 2p_S(\mu_G - \mu_S)(q_S.\mu_{T_G} &- \mu_G.q_S) + p_S^2.q_S(\mu_G - \mu_S)^2 \\
+ p_S.v_S + 2.q_S.(\mu_S - \mu_G).(p_S.\mu_S &- \mu_S.p_S) + q_S^2.p_S(\mu_S - \mu_G)^2
\end{aligned}$$

$$\text{then } v_T = q_S.v_G + p_S.v_S + p_S.q_S.(\mu_G - \mu_S)^2 \tag{2}$$

We are now able to derive and link $B_{TG}$ and $B_{SG}$:

$$\begin{aligned}
B_{TG} &= \frac{c_T}{c_T + c_G}\Big(\mu_T - \frac{c_T.\mu_T + c_G.\mu_G}{c_T + c_G}\Big)^2 + \frac{c_G}{c_T + c_G}\Big(\mu_G - \frac{c_T.\mu_T + c_G.\mu_G}{c_T + c_G}\Big)^2 \\
&= \frac{c_T}{c_T + c_G}\Big(\frac{c_G.\mu_T - c_G.\mu_G}{c_T + c_G}\Big)^2 + \frac{c_G}{c_T + c_G}\Big(\frac{c_T.\mu_G - c_T.\mu_T}{c_T + c_G}\Big)^2
\end{aligned}$$

$$B_{TG} = \frac{c_T.c_G(\mu_T - \mu_G)^2}{(c_T + c_G)^2} \tag{3}$$

$$= \frac{c_T.c_G.(q_S.\mu_G + p_S.\mu_S - \mu_G)^2}{(c_T + c_G)^2} = \frac{c_T.c_G.p_S^2(\mu_S - \mu_G)^2}{(c_T + c_G)^2} = \frac{c_G.c_S^2.(\mu_S - \mu_G)^2}{c_T.(c_T + c_G)^2}.$$

$$\text{Similary to Eq. (3) we have: } B_{SG} = \frac{c_S.c_G.(\mu_S - \mu_G)^2}{(c_S + c_G)^2}. \tag{4}$$

$$\text{Thus from Eq. (4) and (5): } B_{TG} = \frac{c_S.(c_S + c_G)^2}{c_T.(c_T + c_G)^2}.B_{SG}. \tag{5}$$

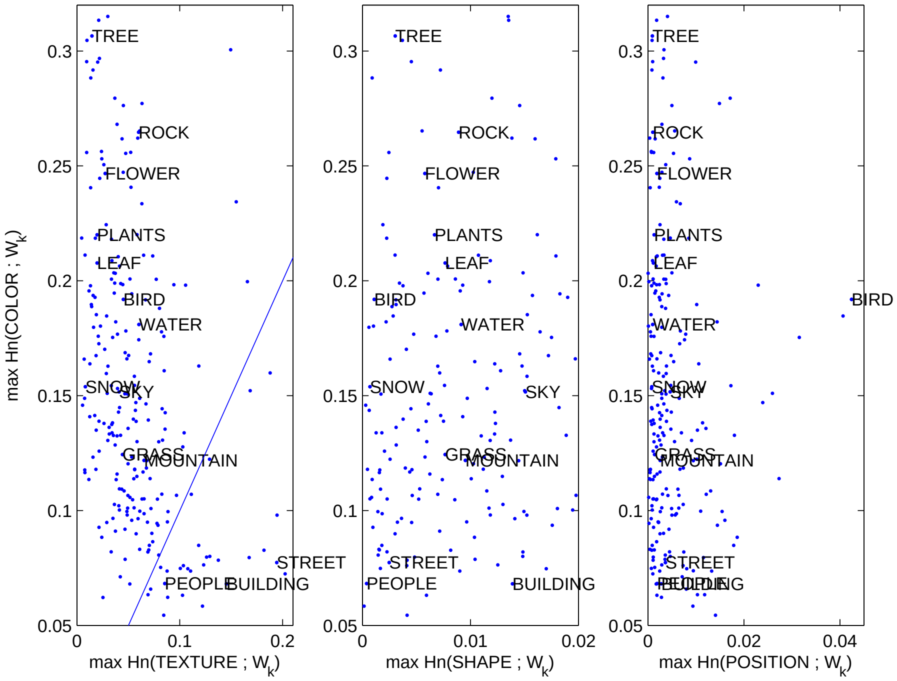

Figure 2: Maximum values of normalised estimated discrimination power $Hn(x; w_k) = \hat{F}(x; w_k)/\sum_x \hat{F}(x, w_k)$ for COLOR, TEXTURE, SHAPE, and POSITION features sets for the 14 most frequent words of the database (other words are represented by a simple dot). Results are intuitively correct: TREE, ROCK, FLOWER, PLANTS are mostly discriminated by color; while BUILDING and STREET are more discriminated by texture. SHAPE is in average not very competitive in comparison to COLOR, neither POSITION. BIRD is the word the most discriminated by POSITION, indeed most of COREL images with a bird represent a bird in the image center.

We also derive the Within variances $W_{TG}$ and $W_{SG}$:

$$W_{TG} = \frac{c_T.v_T + c_G.v_G}{c_T + c_G} \quad = \frac{c_T.(q_S.v_G + p_S.v_S + p_S.q_S.(\mu_G - \mu_S)^2) + c_G.v_G}{c_T + c_G}$$

$$= \frac{(q_S.c_T + c_G).v_G + p_S.c_T.v_S + p_S.q_S.c_T.(\mu_G - \mu_S)^2}{c_T + c_G}$$

$$\text{then } W_{TG} = \frac{(c_T - c_S + c_G).v_G + c_S.v_S + p_S.q_S.c_T.(\mu_G - \mu_S)^2}{c_T + c_G}. \tag{6}$$

$$\text{By definition } W_{SG} = \frac{c_S.v_S + c_G.v_G}{c_S + c_G}, \quad \text{so } v_G = \frac{c_S + c_G}{c_G}.W_{SG} - \frac{c_S.v_S}{c_G}.$$

$$W_{TG} = \frac{(c_T - c_S + c_G).\left(\frac{c_S + c_G}{c_G}.W_{SG} - \frac{c_S.v_S}{c_G}\right) + c_S.v_S + p_S.q_S.c_T.(\mu_G - \mu_S)^2}{c_T + c_G}$$

$$= \frac{(c_T - c_S + c_G).(c_S + c_G)}{c_G.(c_T + c_G)}.W_{SG} - \frac{c_S.(c_T - c_S)}{c_G.(c_T + c_G)}.v_S + \frac{c_S.(c_T - c_S)}{c_T.(c_T + c_G)}.(\mu_G - \mu_S)^2. \tag{7}$$

$$\hat{V}(x; w_k) = \frac{\frac{(c_T - c_S + c_G).(c_S + c_G)}{c_G.(c_T + c_G)}.W_{SG} - \frac{c_S.(c_T - c_S)}{c_G.(c_T + c_G)}.v_S + \frac{c_S.(c_T - c_S)}{c_T.(c_T + c_G)}.(\mu_G - \mu_S)^2}{\frac{c_S.(c_S + c_G)^2}{c_T.(c_T + c_G)^2}.B_{SG}}$$

$$= \frac{c_T(c_T - c_S + c_G)(c_T + c_G)}{c_G.c_S(c_S + c_G)} \frac{W_{SG}}{B_{SG}} + \frac{(c_T - c_S)(c_T + c_G)}{c_S.c_G}\left(1 - \frac{c_T}{c_G} \frac{v_S}{(\mu_G - \mu_S)^2}\right)$$

$$\text{thus } \hat{V}(x; w_k) = A(w_k).V(x; w_k) + B(w_k).\Big(1 - C(x; w_k)\Big) \tag{8}$$

where $A$ and $B$ are positive constants independent of $x$, only depending on number of blobs in sets $T$, $S$, $G$ (experimentations on COREL database show that for all words, A and B are close to 10). Therefore, for any given word $w_k$, $\hat{V}(x; w_k)$ is a linear function of $V(x; w_k)$ if $C(x; w_k)$ is negligible in front of 1. This is the case if (hyp. 2) $\frac{c_T}{c_G}$ is small, which is true in COREL database since it is close to 0.01 for most words, and never exceeds 0.2 (actually one can build any database such that $C_T << C_G$) and (hyp. 3) $v_S$ is tiny in front of $(\mu_G - \mu_S)^2$ which is the case when $x$ is a reasonably good feature to discriminate $G$ and $S$ (e.g. $w_k$ is represented by a rather stationnary feature value different from the mean contextual value). Then order of $\hat{V}$ and $V$ values are the same. Finally, for each word $w_k$, even without knowing which blob of the image it labels, one can estimate the most discriminant features by simply ranking $\hat{F}$ values.

Thereby, in order to estimate how many and which of the $\mathcal{X}_n, n \in \{1, .., \delta\}$ features are really discriminant for each word $w_k$, we simply sort by decreasing order all the $\hat{F}(\mathcal{X}_n; w_k)$, and calculate $N < \delta$ where $\delta$ is the dimension number of visual space and $N$ is defined by: $\sum_{n=1}^{N} \hat{F}(\mathcal{X}_n; w_k) = \frac{\sum_{n=1}^{\delta} \hat{F}(\mathcal{X}_n; w_k)}{2}$. Thus $\mathcal{X}_1, .., \mathcal{X}_N$ are considered as the $N$ best discriminative features for $w_k$.

# 3 Experimentations on COREL image database

To test the efficiency of ALDA, extensive experiments are done on the COREL images database [9] made of 10 000 images with approximately 100 000 segments preprocessed by K. Barnard and al. [1]. Each image is labelled by an average 3.6 words from a lexicon of 267 different words, and has an average of 10 visual segments ('blobs') from the Normalized Cuts algorithm [11], which somehow produces small ones. Each blob is described by a set of $\delta = 40$ features listed below by their dimension index. Firstly POSITION and SHAPE: (1,2) horizontal and vertical blob's position; (3) the proportion of the blob in its image; (4) ratio of bold's area to the perimeter squared; (5) moment of inertia; (6) ratio of the blob's area by its convex hull. COLOURS (7,..,24) are represented by the average and standard deviation of (R,G,B), (r,g,S) and (L,a,b). TEXTURES (25,..,40) are extracted by gaussian filters [1].

## 3.1 F estimation for COLOR, TEXTURE, SHAPE and POSITION

We run ALDA on 6 000 COREL images, and measure for each word the maximum value of $\hat{F}$ for SHAPE, COLOR or TEXTURE features sets. These values represented in Fig. 2 for the 14 most frequent words are intuitively correct and show the word dependence of ALDA. The repartition analysis, over words of all the 6 000 images of the train set, of selected $N$ best features are respectively 3% for POSITION, 8% for SHAPE features, 65% for COLOR features, 24% for TEXTURE features. COLOR features are confirmed to be the most discriminant ones (see also Fig. 2). The simple TEXTURE features (16 gaussian filters) are better than the SHAPE ones, certainly because blobs' segmentation are imprecise (see Fig. 1).

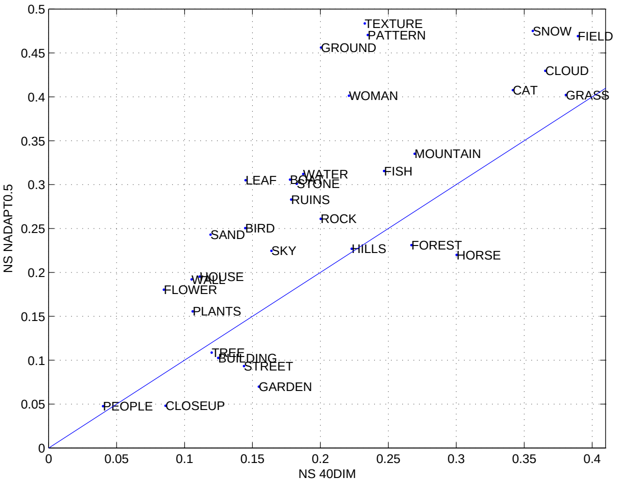

Figure 3: Word visual consistency representation for 40DIM method (in X-coordinate) and for NADAPT0.5 method (in Y-coordinate). NADAPT0.5 method gives better results than 40DIM except for *closeup, garden, street, forest, horse.*

## 3.2 Hierarchical Ascendant Classifications improved by ALDA

To demonstrate ALDA efficiency on a classification task, we now run on COREL a Hierarchical Ascendant Classifications (HAC) of visual features into word categories [12]. As in [2], we measure the system performance using the Normalised Score $NS = sensi. + specif - 1$ [1, 8]. Compared to the raw visual input space, good results have been obtained reducing HAC visual features inputs to ALDA $N$ best discriminant features as previously defined end of section 2 (method called NADAPT0.5). NS values for HAC on the 40 usual visual dimensions or word adaptive features are shown in Fig. 3. Classification of the 3 000 images of the test set shows a gain of +37% of NS, and simultaneously an average over all words of a dimension reduction from $\delta = 40$ to 4 best features (see [12] for more details on the HAC experiments).

# 4 Conclusion

In this paper we present ALDA based on an approximation of the Fisher LDA. We shown that, under weak assumptions (hyp. 1 to 3), ALDA estimates $N$ best features which enhance HAC task, while reducing by 10 the visual space dimension. The main contributions on this paper are summarized as follows: (a) For the first time a theoretical demonstration of ALDA is given in the first section. (b) We implement ALDA on a reference image database and we analyse word dependant features sets constructed using ALDA. (c) We integrate ALDA in a simple HAC model, leading to significant improvements. Further auto-annotation experiments are currently being done on COREL with a bayesian system (DIMATEX model [5]), yielding to promising first results.

# 5 Acknowledgments

We thank K. Barnard and J. Wang [14] for providing COREL image database.

# References

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

[2] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. Forsyth. The effects of segmentation and feature choice in a translation model of object recognition. *Computer Vision and Pattern Recognition*, pages 675–682, 2003.

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

[4] J. Luettin G. Potamianos and C. Neti. Hierarchical discriminant features for audio-visual LVCSR. In *Proc. of IEEE Int. Conf. ASSP*, 2001.

[5] Hervé Glotin and Sabrina Tollari. Image auto-annotation method using dichotomic visual clustering for CBIR. In *Proc. of IEEE EURASIP Fourth International Workshop on Content-Based Multimedia Indexing (CBMI2005)*, june 2005.

[6] Philippe H. Gosselin and Matthieu Cord. A comparison of active classification methods for content-based image retrieval. In *Proc. of the 1st Internationnal Workshop on Computer Vision Meets Databases (CVDB2004) in conjonction with ACM SIGMOD 2004*, pages 51–58, Paris, France, 2004.

[7] Q.S. Liu, R. Huang, H.Q. Lu, and S.D. Ma. Face recognition using kernel based Fisher discriminant analysis. In *Proc. of Int. Conf. Automatic Face and Gesture Recognition*, pages 197–201, may 2002.

[8] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 275–278, 2003.

[9] H. Muller, S. Marchand-Maillet, and T. Pun. The truth about corel - evaluation in image retrieval. In *The Challenge of Image and Video Retrieval (CIVR02)*, 2002.

[10] 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 *Proc. IEEE Work. Multimedia Signal Process.*, 2001.

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

[12] Sabrina Tollari and Hervé Glotin. Keyword dependant selection of visual features and their heterogeneity for image content-based interpretation. Technical Report LSIS.RR.2005.003, LSIS, 2005.

[13] Sabrina Tollari, Hervé Glotin, and Jacques 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.

[14] J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.