# Sélection adaptative des descripteurs visuels et dérivation de métadescripteurs contextuels dépendant du mot-clé pour l'indexation automatique d'images

**Sabrina Tollari— Hervé Glotin**

*Laboratoire LSIS - Equipe INCOD*
*83957 La Garde cedex*
*{tollari,glotin}@univ-tln.fr*

RÉSUMÉ. *Pour améliorer les performances des systèmes automatiques d'indexation d'images à partir de données réalistes, nous proposons une méthode permettant de sélectionner les traits visuels les plus discriminants pour un mot-clé donné. Nous filtrons ces traits en approximant l'analyse factorielle discriminante (AFD) sur les images mal indexées d'une base d'images généralistes. Puis, nous construisons à l'aide d'un algorithme non-supervisé des classes visuelles dans l'espace visuel complet ou dans plusieurs sous-espaces composés des traits les plus discriminants de chaque mot-clé. Nous montrons sur la base COREL que notre filtrage de traits améliore jusqu'à 37% l'association d'un mot-clé avec ses classes visuelles, tout en réduisant le nombre de dimensions de 90%, et que l'espace visuel peut être enrichit par l'adjonction de méta-traits visuels qui mesure l'hétérogénéité de chaque trait visuel dans chaque image.*

ABSTRACT. *In order to enhance real automatic image indexing we propose a method reducing features space. We estimate the most discriminant visual features for a given keyword, by approximating Fisher discriminant analysis on the real not well labelled image databases (e.g. where there is many to many relations between visual concept and keyword). Then, we use a non-supervised clustering algorithm to build visual clusters: using all the features of the visual space, or several subspaces made up of the most discriminant features depending of each keyword. Comparisons of indexing scores on COREL show an indexing enhancement going up to 37%, while reducing the number of dimensions of 90%, and show how a meta visual feature called heterogeneity could improve indexation systems.*

MOTS-CLÉS : *Recherche d'images, indexation multi-dimensionnelle, CAH, AFD, hétérogénéité*

KEYWORDS: *Images retrieval, multi-dimensional indexing, clustering, DFA, heterogeneity*

## 1. Introduction

Les systèmes d'indexation d'images (ou plus largement audio-visuels) se séparent en deux grandes catégories : (1) ceux qui sont dédiés à la détection de concepts spécifiques et (2) ceux qui sont construits pour des données traitant de sujets généralistes. Dans le premier cas, les systèmes sont spécialisés pour le traitement d'images dans le but d'en tirer une information précise, comme la présence de certains objets (voitures, armes, visages...) ce qui permet d'optimiser les procédures de traitements visuels. Dans le second cas, les systèmes sont construits pour donner les meilleurs résultats pour tous les concepts en moyenne, mais pas pour chaque concept pris indépendamment. Dans la pratique, les utilisateurs sont plutôt consommateurs de systèmes généralistes de type (2) (agences de presse, web...) qui tireraient profit d'une plus forte adaptabilité de leurs traitements visuels afin de se rapprocher des performances des systèmes de type (1). Ces systèmes possèdent souvent des données de structures complexes. Un système de recherche d'images efficace doit pouvoir tirer profit de toutes ces informations en sélectionnant les traits les plus pertinents suivant le contexte, ainsi que des métadescripteurs les accompagnant comme l'hétérogénéité des surfaces de l'image (Ounis, 1999, Martinet, 2004, Martinet *et al.*, 2003). La figure 1 donne un exemple d'image segmentée en 8 régions (appelées dans cet article « blobs ») et annotée par 4 mots-clés. Pour le mot « water », les traits les plus discriminants serait la couleur bleu, la texture, mais pas la forme, au contraire du mot « boat ». Dans la pratique, nous constatons que peu de systèmes modélisent efficacement le couplage visuo-textuel pour permettre une indexation et une recherche d'informations réellement adaptées aux images. Cette pénurie peut être due à difficulté de modéliser les relations entre les attributs visuels et les mots d'une requête sémantique de l'utilisateur (paradigme dit du « fossé sémantique »), qui est d'autant plus délicate à aborder que les données pour la construction des modèles ne semblent pas encore adaptées au problème. Dans (Berrani *et al.*, 2002) est posé le problème de la sélection des descripteurs efficaces qui contiennent un nombre minimum de dimensions pour permettre des calculs et une recherche rapide. (Jebara *et al.*, 2000) montre qu'il est possible par le critère du Maximum Entropy Discrimination de sélectionner les traits au sein d'un processus de classification linéaire ou de régression. D'autres travaux (Haddad *et al.*, 2001) montre par une méthode de fouilles d'images que la réduction du nombre de dimensions de l'espace visuel n'affecterait par forcement l'efficacité des associations signal/symbole, mais sur une base de données de quelques centaines d'images.

La contribution de cet article est double : (i) nous proposons un usage « à la limite » d'une analyse factorielle discriminante (AFD) sur la base de données COREL (Wang, 2004, Wang *et al.*, 2001) ne possédant que des relations ambigües entre mot-clé et objet d'une image afin de déterminer automatiquement quels sont les traits visuels discriminants à conserver, et (ii) nous montrons que l'utilisation des valeurs d'hétérogénéité des traits visuels d'une image permet de construire des métadescripteurs de l'image qui apportent une nouvelle information.

Dans partie suivante, nous présentons une méthode statistique simple qui permet de réduire l'espace visuel aux traits les plus pertinents pour un mot-clé donné. Dans

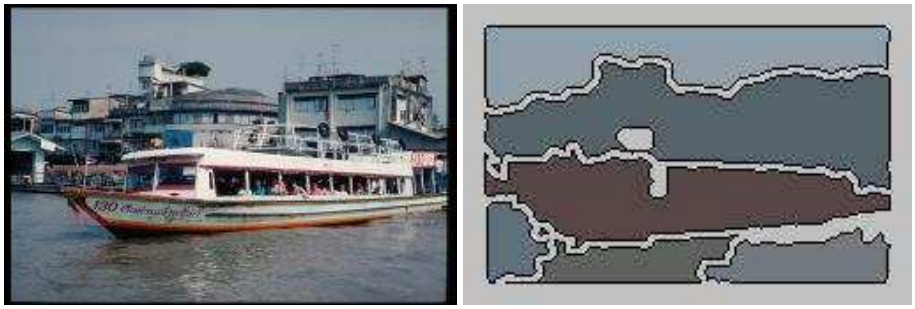

**Figure 1.** *Exemple de segmentation d'une image. Chaque segment est appelé « blob ». Chaque image de la base COREL est annotée manuellement par des mots de référence (pour cette image « water », « boat », « harbor » et « building »).*

la troisième partie, nous présentons une méthode permettant de comparer différentes expériences de réduction de l'espace visuel. Dans la quatrième partie, nous donnons les résultats d'expériences réalisées sous différents sous-espaces vectoriels, dont l'un est construit à l'aide de la métadonnée hétérogénéité. Enfin, dans la dernière partie, nous discutons nos résultats et nous concluons.

## 2. Sélection des traits visuels pertinents et réduction du nombre de dimensions

Déterminer quels sont les traits visuels les plus efficaces pour annoter une image avec un mot est un problème difficile car les données mises à disposition pour le traiter ne sont pas conformes aux exigences des méthodes statistiques classiques. En effet, si des travaux ont par exemple déjà montré qu'il est possible d'utiliser des méthodes simples comme la LDA pour discriminer des traits acoustiques et visuels afin d'améliorer les performances de systèmes de reconnaissance audio-visuels (Neti *et al.*, 2001), ces études ont été faites sur des corpus bien étiquetés, c'est-à-dire décrivant une relation univoque entre une classe conceptuelle et le signal. Les bases de données actuellement disponibles pour l'apprentissage de modèles d'auto-indexation d'images sont rares, et ne décrivent pas une relation univoque entre chaque objet d'une image et un mot-clé, ce qui serait le cadre idéal pour l'application d'une Analyse Factorielle Discrimimante (AFD) pour déterminer les traits visuels les plus discriminants pour un mot-clé donné. La difficulté principale pour une application efficace de l'AFD sur notre corpus d'images est que nous ne disposons pas d'une indexation textuelle pour chaque blob, mais d'un ensemble de mots pour une image. Nous faisons cependant l'hypothèse : *si la base d'images présente chaque concept avec une variété contextuelle assez large, alors les analyses de variances des traits visuels ne seront significatives que pour l'objet récurrent considéré.* Nous allons donc appliquer l'AFD « à la limite » (Glotin *et al.*, 2005) pour déterminer les traits les plus pertinents pour décrire chaque mot-clé de la base COREL, et nous vérifierons la validité des résultats obtenus en mesurant les gains d'une classification hiérarchique ascendante des concepts.

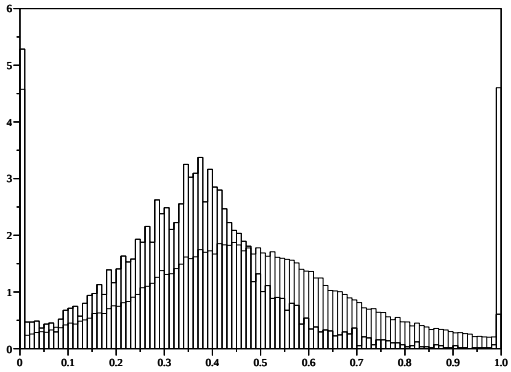


**Figure 2.** *Exemple d'histogrammes pour les classes MOT et NONMOT du mot-clé « snow » pour le trait visuel le plus discriminant : « B » de RGB.*

Pour chaque mot du lexique, nous construisons donc une partition en deux classes de la base d'apprentissage : l'ensemble des images qui sont indexées par ce mot (appelée classe MOT) et l'ensemble des images qui ne sont pas indexées par ce mot (appelée classe NONMOT). La figure 2 donne un exemple des distributions obtenues pour les classes MOT et NONMOT du mot « snow » pour l'un des traits visuels. Pour mesurer la dispersion entre les deux classes, nous calculons pour chaque mot $w_i$ et pour chaque trait visuel $v_j$, la variance inter-classe $B(v_j; w_i)$ (variance des moyennes de chaque classe) et la variance intra-classe $W(v_j; w_i)$ (moyenne des variances de chaque classe) entre les deux classes. Finalement, nous calculons pour chaque mot $w_i$ et pour chaque trait $v_j$ le pouvoir discriminant $F(v_j; w_i)$ défini par :

$$F(v_j; w_i) = \frac{B(v_j; w_i)}{B(v_j; w_i) + W(v_j; w_i)}. \quad [1]$$

Nous pouvons donc, dans un premier temps, choisir pour chaque mot $w_i$ les $N$ dimensions les plus discriminantes parmi les $\delta$ dimensions du vecteur visuel (méthode NBEST). Dans un deuxième temps, nous déterminons automatiquement pour chaque mot $w_i$ les $N$ dimensions les plus discriminantes (méthode NADAPT) telles que, après classement par ordre décroissant des $F(v_j; w_i)$, la somme de leurs pouvoirs discriminants cumule 50% de la somme totale des pouvoirs discriminants de tous les traits pour ce mot :

$$\sum_{j=1}^{N} F(v_j; w_i) = \frac{1}{2} \sum_{j=1}^{\delta} F(v_j; w_i). \quad [2]$$

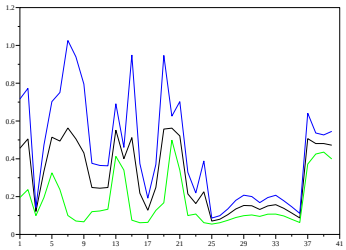

**Figure 3.** *Exemple de classe visuelle pour le mot « woman » représentée par le vecteur centroïde (courbe du milieu). Les deux autres courbes sont respectivement la somme du vecteur centroïde plus ou moins le vecteur d'écarts-types multiplié par une constante. En abscisse, les 40 dimensions visuelles (6 de formes, 18 de couleurs, 16 de textures). En ordonnée, la valeur de chaque trait comprise entre 0 et 1.*

## 3. Construction et évaluation de classes visuelles

Une fois que nous avons déterminé pour chaque mot les traits les plus discriminants, nous réalisons plusieurs expériences sur divers sous-espaces visuels pour valider nos hypothèses. Pour cela, nous construisons d'abord sur un ensemble d'images d'apprentissage des classes visuelles, qui nous servent de classes de références, afin de modéliser les dépendances entre les traits visuels pour chaque mot. Puis, nous évaluons la qualité de cette classification, c'est-à-dire de l'association entre un mot et ses classes visuelles, en calculant le score obtenu par classification supervisée des images d'une base de test. La méthode utilisée est décrite plus en détails dans (Tollari, 2005), nous la résumons ci-après.

Pour chaque mot, nous construisons un sous-ensemble de la base d'apprentissage à partir des images possédant ce mot. Nous réalisons ensuite sur ce sous-ensemble une Classification Ascendante Hiérarchique (CAH) (Lance *et al.*, 1967) (construction non-supervisée de clusters) qui permet de regrouper progressivement les blobs proches dans l'espace au sein d'une même classe. Nous déterminons la valeur d'arrêt de la CAH en choisissant celle qui donne le meilleur score (le calcul du score est expliqué ci-après). Nous gardons alors seulement les classes qui contiennent un nombre significatif de blobs. Chaque classe (exemple figure 3) est représentée uniquement par un couple de vecteurs de même dimension : le vecteur centroïde et le vecteur des écarts-types de la classe visuelle de chaque dimension, ainsi que par une constante optimisée de manière empirique sur un ensemble de développement pour donner le meilleur score pour chaque mot (Tollari, 2005). Le fait que cette méthode de construction de classes visuelles puisse produire plusieurs classes visuelles pour un même mot est intéressant, car un mot peut avoir plusieurs sens, et que chacun de ces sens peut avoir plusieurs représentations visuelles. Par exemple, le mot anglais « plant » peut

correspondre à un végétal (une plante) ou bien à un bâtiment (une usine). De plus, la plupart des plantes sont vertes, mais elles peuvent être aussi de différentes couleurs.

Les classes visuelles que nous venons d'obtenir représentent les régions de l'espace visuel correspondant aux caractéristiques visuelles possibles d'un mot. C'est pourquoi lors de la phase de test, un mot sera associé à un blob d'une image de test si le vecteur visuel du blob appartient à l'une des classes visuelles du mot. Un mot sera associé à une image de test si au moins un blob de l'image a été associé avec ce mot. Chaque image de la base de test possède initialement un ensemble de mots de référence, nous pouvons donc faire un calcul de score. Nous utilisons le score « Normalized Score » (Barnard *et al.*, 2003, Monay *et al.*, 2003) : $NS = \frac{right}{n} - \frac{wrong}{N-n}$ où $right$ est le nombre d'images initialement annotées par le mot considéré qui possèdent au moins un blob dans l'une des classes visuelles de ce mot, $wrong$ est le nombre d'images qui n'étaient pas annotées par ce mot, mais qui ont au moins un blob dans l'une des classes visuelles de ce mot, $N$ est le nombre total d'images dans la base de test et $n$ est le nombre d'images dans la base de test annotées par ce mot [1].

## 4. Expérimentations

Les expérimentations consistent tout d'abord a réaliser une CAH sur l'ensemble des traits visuels et à calculer les scores NS pour l'ensemble des mots du lexique, puis à effectuer des CAH sur plusieurs sous-espaces visuels et enfin à comparer les scores NS obtenus avec la première expérience afin de voir si le choix d'utiliser le pouvoir discriminant des traits visuels par rapport aux mots permet d'améliorer les scores NS et donc de faire de meilleures associations entre classes visuelles et mot-clé.

La base d'images utilisée est un sous-ensemble de COREL (Wang, 2004, Wang *et al.*, 2001). Elle est composée de 10000 images qui possèdent de 1 à 5 mot-clés choisis manuellement parmi un ensemble de 250 mots environ. En moyenne, il y a 3,6 mots-clés par image. Les images ont été prétraitées par des chercheurs du Computer Vision Group de l'université de California (Berkeley) et du Computing Science Department de l'université d'Arizona (Barnard *et al.*, 2003). Chaque image a été segmentée en utilisant la méthode « normalized cuts » (Shi *et al.*, 2000) et les 10 plus grands blobs ainsi créés ont été conservés. En moyenne sur notre corpus, il y a 9,5 blobs par image. La figure 1 donne un exemple de segmentation par « normalized cuts » . Les auteurs ont choisi d'extraire des caractéristiques visuelles générales calculables sur tout type de segments. Chaque blob est décrit par un vecteur de 40 dimensions, composées de : 6 dimensions de formes (aire, périmètre sur aire, convexité, moment d'inertie, position en x et y du barycentre du segment), 18 dimensions de couleurs (RGB, RGS, LAB et leurs écarts types), 16 dimensions de textures. Nous avons ensuite normalisé dans

1. On remarque que $-1 \leq NS \leq 1$. $NS = 1$ quand on trouve les $n$ mots de références, et aucun des autre mots, -1 quand on ne trouve que les mots qui ne sont pas de référence, 0 quand on trouve tous les mots. Le rapport $right/n$ (aussi appelé rappel ou sensibilité) donne le taux de bonnes indexations, le rapport $wrong/(N-n)$ (équivalent à : un moins spécifi cité) donne le taux de mauvaises indexations. Le score NS mesure donc la différence entre ces deux rapports.

**Tableau 1.** *Comparaisons des valeurs de scores NS moyens, minimaux et maximaux pour les 35 mots les plus fréquents dans la base d'apprentissage pour les expériences 40DIM, 40DIMH, 5BEST, NADAPT0.5.*

| | 40DIM | 40DIMH | | 5BEST | | NADAPT0.5 | | |
|---|---|---|---|---|---|---|---|---|
| | NS | NS | Gain | NS | Gain | N | NS | Gain |
| Dimensions | 40 | 40 | | 5 | | 4.14 | | |
| Moyenne | 0.22 | 0.16 | -17% | 0.28 | +33% | 4.14 | 0.29 | +37% |
| Minimum | 0.04 | -0.01 | -109% | 0.02 | -84% | 1 | 0.04 | -54% |
| Maximum | 0.38 | 0.34 | +80% | 0.51 | +199% | 8 | 0.48 | +127% |

$[0, 1]$ les vecteurs visuels par estimation MLE de distribution Gamma. Le lexique a été réduit aux mots-clés ayant plus de 20 occurrences dans la base d'apprentissage : il est donc finalement composé d'un ensemble d'environ 160 mots-clés, dont 35 appartiennent à plus de 100 images de la base d'apprentissage. Le corpus est ensuite séparé aléatoirement en un ensemble d'apprentissage TRAIN de 5000 images, un ensemble de développement DEV de 2500 images et un ensemble de test TEST de 2500 images.

Nous avons effectué tout d'abord une CAH sur les 40 dimensions visuelles (expérience appelée 40DIM, tableau 1 et abscisse de la figure 4) comme dans (Tollari, 2005)[2]. Les mots qui ont un fort NS sont : *fi eld*, *grass*, *cloud*, *snow*, *cat*. Nous remarquons que dans l'ensemble ces mots ont une forte consistance visuelle (les champs et l'herbe c'est souvent vert, la neige et les nuages c'est souvent blanc). Les mots qui ont un faible NS sont : *people*, *flower*, *closeup*, *wall*, *sand*, *tree*, *plants*, *house*. Les faibles scores obtenus pour ces mots peuvent peut-être s'expliquer par le bruit ajouté par certaines dimensions. Par exemple, une fleur peut être de différentes couleurs, on ne peut donc pas associer une couleur particulière au mot fleur, les traits couleurs apportent certainement plus de bruit que d'information aux classes visuelles de ce mot.

Nous souhaitons réduire le nombre de traits de l'espace tout en améliorant les scores obtenus pour chacun des mots, mais nous ne savons pas à priori quelles traits garder. Nous avons essayé tous d'abord des associations de traits visuels naïves. Par exemple, les expériences consistant à construire les classes visuelles sur : les 6 dimensions visuels de LAB (NS moyen : 0.05), les 18 dimensions de couleurs (NS moyen : 0.10), les 5 (resp. les 10) dimensions les plus discriminantes pour l'ensemble des mots (NS moyens : 0.08 et 0.10). Nous remarquons que ces méthodes donnent des scores moyens bien inférieurs à la méthode 40DIM ce qui montre que réduire l'espace visuel de manière naïve réduit sensiblement les scores NS, et que de plus utiliser les traits les plus discriminants pour l'ensemble des mots n'est pas en moyenne un bon choix.

Nous réalisons alors une autre expérience moins naïve consistant à prendre pour chaque mot, les 5 traits les plus discriminants (5BEST). Les résultats (tableau 1)

2. Dans (Tollari, 2005) la base d'apprentissage est de 7000 images au lieu de 5000. Nous obtenons cependant une moyenne des scores NS de 0.22 identique.

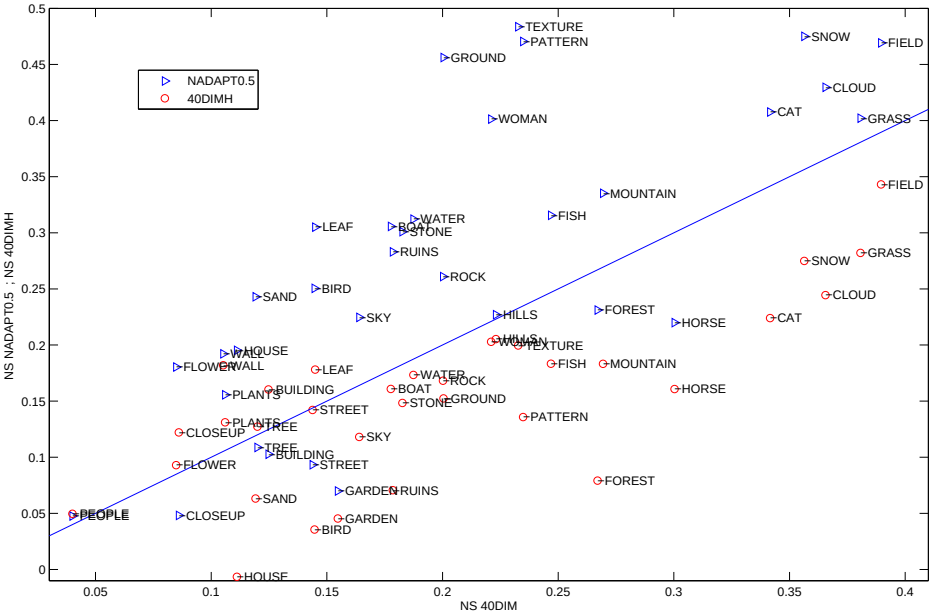


**Figure 4.** *Représentation des consistances visuelles (scores NS) des mots pour la méthode 40DIM en abscisse et pour les méthodes NADAPT0.5 et 40DIMH en ordonnée. Le gain de classifi cation peut se mesurer avec la distance à la diagonale. L'hétérogénéité représente le contexte visuel de l'objet dans la scène.*

montrent un gain moyen par rapport à la méthode 40DIM de +33% alors que la réduction du nombre de dimensions est de 87% (5 dimensions au lieu de 40). Nous effectuons une autre expérience NADAPT0.5 consistant à prendre pour chaque mot les $N$ dimensions les plus discriminantes telles qu'elles cumulent 50% des valeurs de pouvoir discriminant de l'ensemble des traits visuels du mot considéré. Les résultats (tableau 1 et figure 4) montrent un score NS moyen de 0.29 soit un gain moyen par rapport à 40DIM de +37% et un gain moyen par rapport à 5BEST de +12% pour une réduction du nombre de dimensions de 90% par rapport à 40DIM (4.14 au lieu de 40) et de 17% par rapport à 5BEST (4.14 au lieu de 5). Cette dernière expérience valide notre hypothèse selon laquelle réduire l'espace visuel aux dimensions les plus discriminantes pour chaque mot indépendamment permet une amélioration du score NS par rapport à une méthode prenant en compte tous les traits, et donc permet une amélioration de l'association entre des traits visuels et un mot. De plus, choisir le nombre de dimensions en fonction de leurs capacités discriminantes cumulées semble être un choix approprié. Rappelons que nous travaillons sur une base associant un ensemble

de mots à une image, et non pas un mot par blob. Si nous avions une indexation textuelle par blob, les scores auraient certainement encore été meilleurs.

Nous réalisons une dernière expérience sur un nouvel espace visuel. Nous construisons des métadescripteurs en calculant pour chacun des 40 traits visuels, l'hétérogénéité[3] (Martinet, 2004) des valeurs des blobs de l'image pour ce trait. Nous lançons alors une CAH sur l'espace des 40 nouveaux traits (méthode 40DIMH). Les résultats (tableau 1 et figure 4) donnent un score moyen de 0.16. Sur la figure, nous remarquons que les mots qui ont un faible score NS par la méthode 40DIM obtiennent un meilleur score par la méthode 40DIMH (*closeup, plants, wall, building, leaf*), et inversement les mots qui ont un fort score par 40DIM ont un faible score par 40DIMH (*snow, grass, cloud, cat, fi eld*). On peut donc en déduire que l'hétérogénéité apporte une information non négligeable complémentaire de celle de 40DIM. D'autre part, d'après les expériences réalisées en psychovision par J. Martinet ((Martinet, 2004) page 117), le critère d'hétérogénéité appliqué aux surfaces a plus ou moins d'impact dans les descriptions visuelles d'objets. En particulier, il montre que l'objet « oiseau » est très mal discriminé avec le critère d'hétérogénéité, mais que par contre un portrait ou un bateau sont très bien discriminés par ce critère. De même, nous remarquons dans la figure 4 que « bird (0.14,0.04)) » est mal discriminé en 40DIMH, mais que « boat (0.17,0.16) » et le portrait « woman (0.22,0.20) » le sont beaucoup plus. Nous notons donc la même tendance qu'il faudrait confirmer sur des expériences en psychovision.

## 5. Discussions et conclusion

Nous avons montré qu'il est possible de réduire le nombre de dimensions d'un espace visuel tout en améliorant la qualité de la relation entre un mot et ses classes visuelles en choisissant pour chaque mot les traits visuels qui possèdent le plus fort pouvoir discriminant. En effet, l'utilisation de l'AFD pour déterminer les traits visuels les plus discriminants permet une amélioration de l'association d'un mot-clé avec ses classes visuelles de 37% pour une réduction du nombre de dimensions de 90%. Ces scores montrent que l'hypothèse posée au chapitre 2 est valide. Ces résultats peuvent paraître surprenant, car il est plutôt logique de penser que plus on prend des traits visuels et plus les résultats devraient être bons. C'est certainement vrai lorsque l'on recherche des régions d'images proches visuellement, mais les résultats montrent que lorsque l'on recherche des concepts dans des images, il vaut mieux choisir les traits visuels qui sont caractéristiques de ce concept. Par exemple, si un utilisateur recherche le concept « fleur », le système ne devrait pas prendre en compte les traits de couleurs, car une fleur peut être de différentes couleurs. Par contre, ils devraient prendre en compte les traits d'écarts types pour la couleur, car les fleurs sont souvent de couleurs constantes. D'autre part, nous avons montré que les valeurs d'hétérogénéité apportent une information non négligeable complémentaire de celle fournie par les traits visuels qui les ont construits. Si l'on combine efficacement les traits visuels classiques et les

3. La valeur de l'hétérogénéité de la dimension $p$ pour l'image $d$ qui contient le blob $b_j$ qui a pour valeur à la dimension $p$ $b_{j,p}$ est l'entropie : $H_p = -\sum_{b_j \in d} b_{j,p} \times \log_2(b_{j,p})$.

valeurs d'hétérogénéité, on doit pouvoir obtenir des classes visuelles plus pertinentes. Une façon de combiner ces valeurs est de réaliser la sélection des traits visuels ayant les plus forts pouvoirs discriminants parmi les 40 traits visuels classiques et les 40 traits visuels d'hétérogénéité. Nos prochains travaux poursuivront sur des méthodes adaptatives de combinaisons de traits et de méta-traits visuels.

## 6. Bibliographie


Barnard K., Duygulu P., de Freitas N., Forsyth D., Blei D., Jordan M. I., « Matching Words and Pictures », *Journal of Machine Learning Research*, vol. 3, p. 1107-1135, 2003.

Berrani S.-A., Amsaleg L., Gros P., « Recherche par similarités dans les bases de données multidimensionnelles : panorama des techniques d'indexation », *Ingénierie des systèmes d'information (RSTI série ISI-NIS)*, vol. 7, n° 5-6, p. 65-90, 2002.

Glotin H., Tollari S., Giraudet P., « Approximation of Linear Fisher Discriminant Analysis for Adaptive Word Dependent Visual Feature Sets Improving Image Auto-annotation », *Proc. of Advanced Concepts for Intelligent Vision Systems (ACIVS2005), submitted*, 2005.

Haddad H., Mulhem P., « Utilisation de la Fouille de Données Images pour l'Indexation Automatique des images », *Inforsid'2001*, 2001.

Jebara T., Jaakkola T., « Feature Selection and Dualities in Maximum Entropy Discrimination », *Proc. of the 16th Annual Conference on Uncertainty in Artificial Intelligence (UAI-00)*, Morgan Kaufmann Publishers, San Francisco, CA, p. 291-300, 2000.

Lance G., Williams W., « A general theory of classifi catory sorting strategies : I. Hierarchical systems », *Computer Journal*, vol. 9, p. 373-380, 1967.

Martinet J., Un modèle vectoriel relationnel de recherche d'information adapté aux images, Thèse de doctorat, Université Joseph Fourier, Grenoble I, 2004.

Martinet J., Ounis I., Chiaramella Y., Mulhem P., « A Weighting Scheme for Star-Graphs », *European Conference on IR Research (ECIR2003)*, p. 546-554, 2003.

Monay F., Gatica-Perez D., « On image auto-annotation with latent space models », *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, p. 275-278, 2003.

Neti C., Potamianos G., Luettin J., Matthews I., Glotin H., Vergyri D., « Large-vocabulary audio-visual speech recognition : A summary of the Johns Hopkins Summer 2000 Workshop », *Proc. IEEE Work. Multimedia Signal Process.*, 2001.

Ounis I., « A Flexible Weighting Scheme for Multimedia Documents », *DEXA'99 : Proc. of the 10th International Conference on Database and Expert Systems Applications*, 1999.

Shi J., Malik J., « Normalized Cuts and Image Segmentation », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n° 8, p. 888-905, 2000.

Tollari S., « Filtrage de l'indexation textuelle d'une image au moyen du contenu visuel pour un moteur de recherche d'images sur le web », *Actes d'ACM CORIA'05*, 2005.

Wang J., « http ://wang.ist.psu.edu/docs/home.shtml », 2004.

Wang J. Z., Li J., Wiederhold G., « SIMPLIcity : Semantics-Sensitive Integrated Matching for Picture Libraries », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, n° 9, p. 947-963, 2001.