Apprentissage par analyse linéaire discriminante des paramètres de fusion pour la recherche d'information multimédia texte-image

Christophe Moulin*, Christine Largeron*, Cécile Barat*,
Mathias Géry*, Christophe Ducottet*

*Université de Lyon, F-42023, Saint-Étienne, France ;
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42023, Saint-Étienne, France ;
Université de Saint-Étienne, Jean-Monnet, F-42023, Saint-Étienne, France.

Résumé. Avec le développement du numérique, des quantités très importantes de documents composés de texte et d'images sont échangés, ce qui nécessite le développement de modèles permettant d'exploiter efficacement ces informations multimédias. Dans le contexte de la recherche d'information, un modèle possible consiste à représenter séparément les informations textuelles et visuelles et à combiner linéairement les scores issus de chaque représentation. Cette approche nécessite le paramétrage de poids afin d'équilibrer la contribution de chaque modalité. Le but de cet article est de présenter une nouvelle méthode permettant d'apprendre ces poids, basée sur l'analyse linéaire discriminante de Fisher (ALD). Des expérimentations réalisées sur la collection ImageCLEF montrent que l'apprentissage des poids grâce à l'ALD est pertinent et que la combinaison des scores correspondante améliore significativement les résultats par rapport à l'utilisation d'une seule modalité.

1 Introduction

L'exploitation automatique des documents multimédia est un sujet d'actualité. En effet, le volume de données multimédia numérisées ne cesse de croître. C'est le cas des données disponibles sur le Web, mais c'est aussi le cas de nombreux autres domaines comme par exemple le domaine médical, la vidéosurveillance, la géographie, les archives publiques et plus récemment le développement des livres numériques liés aux appareils tels que les ultraportables ou les tablettes tactiles. La plupart des systèmes actuels de traitement de données multimédia (tels que les moteurs de recherche), n'exploitent généralement qu'un seul type d'information et, le plus souvent, il s'agit de l'information textuelle.

Le traitement de cette information textuelle s'appuie sur des techniques performantes et efficaces, mises au point et perfectionnées depuis des décennies, et qui ont déjà fait leurs preuves (van Rijsbergen (1979); Manning et al. (2008)). Notamment, le modèle vectoriel de Salton (Salton et al. (1975)) est à la base de nombreux moteurs de recherche. Depuis les années 2000, des progrès significatifs ont été faits pour le traitement des informations visuelles. En particulier, le modèle des sacs de mots visuels (Csurka et al. (2004)), inspiré de l'approche textuelle,

s'est montré performant pour la recherche ou la classification d'images (Nowak et al. (2006); Tirilly et al. (2008)). Ce modèle repose premièrement sur la construction d'un vocabulaire visuel décrivant les motifs caractéristiques d'une base d'images et deuxièmement sur la représentation des images sous la forme d'histogrammes d'occurrences de mots visuels de ce vocabulaire.

Un enjeu scientifique actuel se situe clairement dans la combinaison d'informations de différentes natures. En effet, de nombreux travaux ont montré que le traitement multimodal conduisait à de meilleurs résultats que l'exploitation de médias séparément (Snoek et al. (2005); Ayache et al. (2007)). C'est notamment le cas lorsque l'on considère la combinaison d'informations textuelles et d'informations image (Tollari et al. (2009); Moulin et al. (2010)). Une problématique fondamentale réside alors dans la recherche de l'équilibre dans l'importance à accorder à chacun des deux types d'informations.

L'objectif de cet article est de présenter une nouvelle méthode basée sur l'analyse linéaire discriminante (ALD) de Fisher, qui permet d'apprendre automatiquement les coefficients de pondération entre les informations textuelles et visuelles dans un modèle de combinaison linéaire de ces deux types d'information. L'article est organisé de la manière suivante. Après avoir présenté notre modèle de recherche d'information multimédia dans la section 2, nous introduirons une méthode d'apprentissage des paramètres de combinaison dans la section 3. Enfin les expérimentations seront décrites et analysées dans la section 4. La dernière section sera consacrée à la conclusion et aux perspectives de ces travaux.

2 Modèle de recherche d'information visuelle et textuelle

Le modèle de recherche d'information multimodale que nous avons développé, est basé sur une fusion tardive qui combine linéairement des scores textuels et visuels. Ce modèle comporte plusieurs modules, comme l'indique la figure 1. Un premier module, noté (a), est consacré à l'indexation des documents de la collection D, un deuxième à la représentation des requêtes Q (b). Les documents, comme les requêtes, composés à la fois de texte et d'images, sont représentés sous forme de sacs de mots. Dans ces modules, les textes et les images sont traités séparément. Le troisième module (c) vise à déterminer, pour une requête fournie par un utilisateur, un score par document et par type de contenu (texte ou image). Ce score sera d'autant plus élevé que le contenu du document, relativement au type d'information considéré (visuel ou textuel), correspondra à celui de la requête. Enfin, le dernier module (d) vise à combiner linéairement les scores obtenus pour chaque type d'information de façon à identifier les documents répondant le mieux à la requête. Dans ce modèle, des paramètres permettent de fixer les poids accordés à chaque type d'information. Dans cette section, après une présentation brève du modèle permettant d'introduire les notations, nous reviendrons sur le choix de ces paramètres.

2.1 Représentation du contenu textuel

Étant donnée D, une collection de documents et $T = \{t_1, ..., t_j, ..., t_{|T|}\}$, l'ensemble des termes présents dans les documents de D, chaque document d_i est représenté comme un vecteur de poids $\vec{d_i} = (w_{i,1}, ..., w_{i,j}, ..., w_{i,|T|})$ suivant le modèle introduit par Salton (Salton et al. (1975)). Dans ce modèle, l'importance du terme t_j dans le document d_i est mesurée par

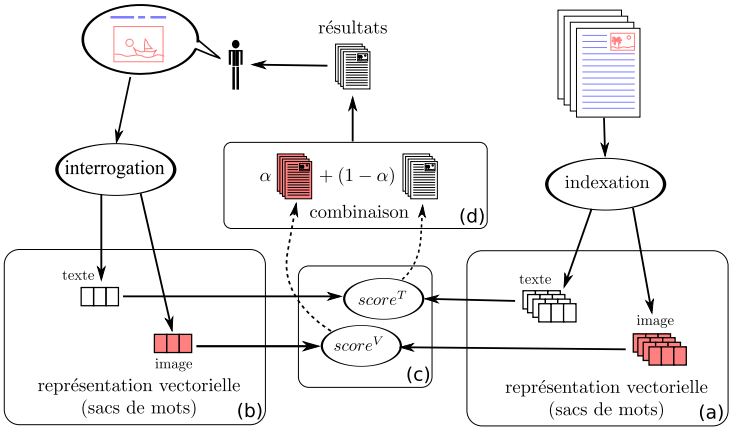

FIG. 1 – *Architecture globale du modèle de recherche d'information multimodale.*

sa fréquence (*term frequency*) $tf_{i,j}$ tandis que son importance sur la collection est évaluée par la fréquence inverse du document (*inverse document frequency*) idf_j. Le poids $w_{i,j}$ du terme t_j dans le document d_i est égal au produit $tf.idf$ dans lequel $tf_{i,j}$ et idf_j peuvent être calculés suivant la formule Okapi (Robertson et al. (1994)) implantée dans Lemur (Zhai (2001)) :

$$tf_{i,j} = \frac{k_1 n_{i,j}}{n_{i,j} + k_1(1 - b + b\frac{|d_i|}{d_{avg}})} \quad (1)$$

où b et k_1 sont des constantes, $n_{i,j}$ est le nombre d'occurrences du terme t_j dans le document d_i, $|d_i|$ est la taille du document et d_{avg} est la taille moyenne des documents de D en considérant comme taille d'un document le nombre de termes dans ce document,

$$idf_j = \log \frac{|D| + 1}{|D_j| + 0,5}$$

où $|D|$ est le nombre de documents dans la collection et $|D_j|$ est le nombre de documents de D dans lequel le terme t_j apparaît au moins une fois.

Une requête q_k étant assimilée à un texte court, elle est aussi représentée comme un vecteur de poids. Afin de trier les documents de D par ordre de pertinence pour la requête q_k, un score peut être calculé pour chaque document d_i par (Zhai (2001)) :

$$score_T(q_k, d_i) = \sum_{t_j \in q_k} tf_{k,j} idf_j tf_{i,j} idf_j$$

2.2 Représentation du contenu visuel

Le contenu visuel d'un document est représenté, comme son contenu textuel, sous forme vectorielle. Ceci nécessite la définition d'un vocabulaire visuel $V = \{v_1, ..., v_j, ..., v_{|V|}\}$ utilisé, comme pour la partie textuelle, pour associer à chaque image un vecteur de poids comportant autant de composantes qu'il y a de mots dans V.

La construction du vocabulaire V, réalisée en utilisant une approche par sac de mots visuels (Csurka et al. (2004)), s'effectue en deux étapes. Dans la première, chaque image de la collection D est découpée en imagettes, décrites à l'aide d'un descripteur visuel. Le découpage utilisé est un découpage régulier en 16×16 imagettes. Les découpages réguliers ou réguliers multi-échelles sont des méthodes simples mais efficaces qui sont bien adaptées pour manipuler de grandes collections (Fei-Fei et Perona (2005); Nowak et al. (2006)). Parmi les descripteurs susceptibles d'être retenus pour décrire chaque imagette, on peut citer le descripteur $SIFT$ (Lowe (2004)) qui décrit la distribution des orientations du gradient dans la cellule avec un vecteur à 128 dimensions. On pourrait également envisager l'utilisation d'autres descripteurs, basés par exemple sur la couleur. L'étape suivante consiste à classer automatiquement, à l'aide de l'algorithme des k-means, les vecteurs visuels associés aux imagettes figurant dans les documents de la collection. Les centres des classes sont retenus comme mots du vocabulaire visuel. On peut remarquer que cette étape est assimilable à la constitution de l'index sur la partie textuelle. Sur la partie visuelle, elle permet de passer d'un descripteur visuel de bas niveau (couleur ou texture par exemple) à un niveau de représentation plus général sous forme de vocabulaire visuel comportant autant de mots que le nombre de classes. On doit cependant noter que ces mots visuels, caractéristiques par exemple d'une texture ou d'une couleur, n'ont malgré tout pas un pouvoir d'expression comparable à celui du vocabulaire textuel.

Par analogie avec l'index textuel, ce vocabulaire visuel est utilisé pour représenter une image, associée à un document ou à une requête, sous forme de vecteur. L'image choisie est découpée en 16×16 imagettes, puis chaque imagette est décrite à l'aide du descripteur choisi pour construire le vocabulaire, et cette description est associée ensuite au mot visuel de V le plus proche en terme de distance euclidienne. Ainsi, de même qu'il est possible de déterminer le nombre d'occurrences $t_{i,j}$ d'un mot textuel t_j dans un document d_i, il est possible de dénombrer le nombre d'occurrence $v_{i,j}$ d'un mot visuel v_j dans une image : il s'agira du nombre d'imagettes de l'image qui sont les plus proches du mot v_j. Comme pour le texte, l'image peut ainsi être représentée à l'aide d'un vecteur de poids calculé avec des formules $tf.idf$. Finalement, il est possible de calculer un score de pertinence $score_V(q_k, d_i)$ d'un document d_i pour une requête image donnée q_k en considérant les termes visuels v_j figurant dans cette requête et dans les images du document par :

$$score_V(q_k, d_i) = \sum_{v_j \in q_k} tf_{k,j} idf_j tf_{i,j} idf_j$$

2.3 Combinaison des informations visuelles et textuelles

En considérant deux vocabulaires, l'un textuel T et l'autre visuel V, un score global peut être calculé pour un document d_i et pour une requête donnée q_k en combinant linéairement les scores obtenus respectivement sur chacune des modalités :

$$score(q_k, d_i) = \alpha score_V(q_k, d_i) + (1 - \alpha) score_T(q_k, d_i)$$

Dans ce score, le paramètre α permet d'accorder plus ou moins d'importance à l'information visuelle par rapport à l'information textuelle dans le classement des documents.

Dans le cas où plusieurs vocabulaires visuels seraient considérés, de façon à tenir compte par exemple de la texture et de la couleur, la formule précédente se généralise simplement de la façon suivante :

$$score(q_k, d_i) = (\sum_{j \in M} \alpha_j \; score_j(q_k, d_i)) + (1 - \sum_{j \in M} \alpha_j) score_T(q_k, d_i) \quad (2)$$

où α_j correspond au poids accordé au descripteur visuel j et M désigne l'ensemble des vocabulaires visuels considérés.

Il est clair que l'efficacité de ce modèle de recherche d'information multimodale dépend directement de son paramétrage. D'après des expérimentations menées dans le cadre de recherches antérieures (Tollari et Glotin (2007); Tollari et al. (2009); Moulin et al. (2009, 2010), il semble que les poids attribués respectivement à l'information textuelle et à l'information visuelle ne doivent pas être les mêmes dans la mesure où les résultats obtenus en considérant uniquement un descripteur textuel sont meilleurs que ceux obtenus uniquement avec un descripteur visuel. Cependant, il n'est pas facile, même pour un expert, de fixer les valeurs à attribuer à ces paramètres.

3 Apprentissage des paramètres de combinaison par ALD

Nous proposons de déterminer les valeurs des paramètres du modèle de recherche d'information multimodale par apprentissage automatique à l'aide d'un échantillon de base composé d'un ensemble de requêtes et pour chacune d'elles de la liste des documents de la collection qui répondent le mieux à cette requête. Cet ensemble est divisé en un échantillon d'apprentissage, utilisé pour calculer les valeurs des paramètres et en un échantillon test, distinct de l'échantillon d'apprentissage, qui permet d'évaluer le modèle après paramétrage. Lors de la phase d'apprentissage, le problème d'estimation des paramètres du modèle est assimilé à un problème de réduction de dimension dans un contexte de classement binaire : la recherche de la combinaison linéaire qui sépare au mieux les documents pertinents des documents non pertinents pour l'ensemble des requêtes. Ce problème est résolu analytiquement à l'aide de l'analyse linéaire discriminante (ALD) de Fisher.

3.1 Formalisation du problème de classement

Considérons le problème de classement binaire dans lequel chaque document est considéré comme pertinent ou non pertinent pour une requête donnée. Dans ce problème, les éléments à classer correspondent donc aux couples document-requête et ils appartiennent à l'une des deux classes "pertinent" ou "non pertinent". De plus, chaque élément peut être décrit par un vecteur de variables correspondant aux scores calculés pour l'ensemble des vocabulaires considérés.

Plus formellement, étant donné $\mathcal{D}$, l'ensemble des documents et $\mathcal{Q}$ un ensemble de requêtes d'apprentissage, l'ensemble des éléments à classer $\mathcal{X}$ est défini comme l'ensemble $\mathcal{D} \times \mathcal{Q}$ de tous les couples document-requête x_ℓ :

$$\mathcal{X} = \{(x_\ell)_{\ell=1,\ldots,|\mathcal{X}|} \; , \; x_\ell = (d_i, q_k)_{i=1,\ldots,|\mathcal{D}|, \, k=1,\ldots,|\mathcal{Q}|}\}$$

Chaque élément appartient à la classe $\mathcal{X}_R$ des éléments pertinents ou à la classe $\mathcal{X}_{\overline{R}}$ des éléments non pertinents selon que le document répond ou non à la requête :

$$\begin{aligned} \mathcal{X}_R &= \{x_\ell \in \mathcal{X} \mid d_i \text{ est pertinent pour } q_k\} \\ \mathcal{X}_{\overline{R}} &= \{x_\ell \in \mathcal{X} \mid d_i \text{ n'est pas pertinent pour } q_k\} \end{aligned}$$

De plus, dans le contexte de la recherche d'information multimodale, chaque élément x_ℓ est représenté par un vecteur de variables $\mathbf{x}_\ell$ dont les composantes correspondent aux scores, calculés pour chaque descripteur j de M entre le document d_i et la requête q_k :

$$\mathbf{x}_\ell = (x_{\ell,j})_{j\in M} = (score_j(d_i, q_k))_{j\in M}$$

où M désigne l'ensemble des vocabulaires considérés, généralement un vocabulaire textuel et plusieurs vocabulaires visuels, comme par exemple un vocabulaire basé sur la texture et un autre sur la couleur.

3.2 Détermination analytique des paramètres du modèle

Chaque élément x de $\mathcal{X}$ peut appartenir à l'une des deux classes et est représenté par un vecteur $\mathbf{x}$ de variables correspondant aux scores textuel et visuels. On se propose donc de déterminer une combinaison linéaire de ces scores qui sépare le mieux possible les deux classes d'éléments. Ce problème revient à rechercher un axe factoriel qui sépare au mieux les deux populations, en tenant compte de leur appartenance aux classes. L'analyse linéaire discriminante fournit une solution à ce problème à partir de la minimisation du discriminant linéaire de Fisher (Fisher (1936); Mahalanobis (1936); Klecka (1980); Mika et al. (1999)). Notons que cette analyse ne requiert aucune hypothèse sur les distributions conditionnelles des variables. En revanche, lorsqu'elle est utilisée pour définir un classifieur, on peut montrer que ce classifieur est optimal au sens de Bayes si les variables suivent une loi normale et qu'elle ont la même matrice de variance-covariance (homoscédasticité) (Klecka (1980)). Des études ont également montré que même si les hypothèses précédentes ne sont pas entièrement satisfaites, le classifieur reste performant (Bouveyron et al. (2005)). En conclusion, dans notre cas, aucune hypothèse n'est requise puisque l'on s'intéresse uniquement à la définition d'un axe factoriel.

Si $\mathbf{x} = (x_j)_{j\in M}$ désigne un vecteur de scores, et $\mathbf{z} = (\alpha_j)_{j\in M}$ le vecteur des coefficients de la combinaison linéaire, alors l'axe factoriel z recherché est donné par :

$$z = {}^t\mathbf{z}\mathbf{x} = \sum_{j\in M} \alpha_j x_j$$

On peut vérifier que la variance $V(z)$ de la variable z est égale à $V(z) = {}^t\mathbf{z}\mathbf{T}\mathbf{z}$ où $\mathbf{T}$ désigne la matrice de covariance totale associée aux scores. Or, d'après le théorème de Huygens, la matrice de covariance totale, qui est constante, se décompose en matrice de covariance intraclasse $\mathbf{W}$ et matrice de covariance interclasse $\mathbf{B}$. La recherche des paramètres du modèle revient alors à déterminer la combinaison des scores qui maximise la variance interclasse et qui minimise la variance intraclasse, ce qui revient à maximiser le discriminant de Fisher $F(\mathbf{z})$ défini par :

$$F(\mathbf{z}) = \frac{{}^t\mathbf{z}\mathbf{B}\mathbf{z}}{{}^t\mathbf{z}\mathbf{T}\mathbf{z}}$$

On démontre que le maximum est atteint pour le vecteur propre associé à la plus grande valeur propre de la matrice $\mathbf{T}^{-1}\mathbf{B}$ (Lebart et al. (1986)).

Dans le cas particulier d'un problème à deux classes, ce vecteur propre, et par conséquent les paramètres du modèle de recherche d'information multimodale, peut être calculé, à un facteur multiplicatif près, par :

$$\mathbf{z} = \mathbf{T}^{-1}(\mu_{\mathbf{R}} - \mu_{\overline{\mathbf{R}}}) \tag{3}$$

où $\mu_{\mathbf{R}}$ (resp. $\mu_{\overline{\mathbf{R}}}$) est le vecteur des moyennes associées aux scores des éléments pertinents (resp. non pertinents) défini par :

$$\mu_{\mathbf{R}} = \frac{1}{|\mathcal{X}_R|} \sum_{x_\ell \in \mathcal{X}_R} \mathbf{x}_\ell \quad \text{et} \quad \mu_{\overline{\mathbf{R}}} = \frac{1}{|\mathcal{X}_{\overline{R}}|} \sum_{x_\ell \in \mathcal{X}_{\overline{R}}} \mathbf{x}_\ell$$

4 Expérimentations

Afin d'évaluer notre méthode d'apprentissage des paramètres du modèle de représentation de documents multimédia, nous avons effectué plusieurs expérimentations sur la collection ImageCLEF (Tsikrika et Kludas (2008, 2009)). Le but est d'étudier l'impact de la prise en compte de l'information visuelle dans un système de recherche d'information multimodale pour lequel on a appris les paramètres de combinaison, puis d'évaluer la qualité de l'apprentissage de ces paramètres.

4.1 Description de la collection

La collection ImageCLEF contient 151 519 documents XML multimédia provenant de l'encyclopédie Wikipedia. Les documents sont composés d'une seule image accompagnée d'un texte court. Les images sont de tailles hétérogènes aux formats JPEG et PNG. Elles peuvent correspondre aussi bien à des photos, qu'à des dessins ou des peintures. Le texte court décrit généralement l'image, mais peut également contenir des informations relatives à l'utilisateur qui a fourni l'image ou relatives aux droits d'utilisation de cette dernière. Les principales caractéristiques des collections utilisées dans le cadre des compétitions ImageCLEF 2008 et 2009 (Tsikrika et Kludas (2008, 2009)) sont présentées dans la table 1.

	2008	2009
Nombre de documents	151 519	
Nombre moyen de mots textuels par document	33	
Nombre de requêtes	75	45
Nombre moyen d'images par requête	1,97	1,84
Nombre moyen de mots textuels par requête	2,64	2,93

TAB. 1 – *Collection ImageCLEF 2008 et 2009*

Chaque année, un ensemble différent de requêtes a été fourni :

– En 2008, il s'agissait d'un ensemble de 75 requêtes, possédant chacune une partie textuelle composée de quelques mots, mais ne comportant pas forcément une partie visuelle. Afin de disposer d'une information visuelle équivalente pour toutes ces requêtes, nous avons ajouté à chacune les 2 premières images pertinentes retournées en interrogeant la collection à l'aide uniquement de la partie textuelle de la requête. Cette procédure est assimilable à un retour de pertinence utilisateur. Ce premier jeu de requêtes est utilisé comme un échantillon d'apprentissage pour calculer les paramètres du modèle.

– En 2009, il s'agissait d'un ensemble de 45 requêtes, chacune composée d'une partie textuelle et d'une partie visuelle avec un nombre moyen d'images par requête égal à 1,84. Ce second ensemble est utilisé comme échantillon test pour évaluer le modèle après paramétrage.

4.2 Paramétrage du modèle vectoriel

Le logiciel Lemur a été utilisé avec les paramètres par défaut (Zhai (2001)). La valeur par défaut du paramètre k_1 de la formule Okapi (cf. équation 1) est fixée à 1. Étant donné que $|d_k|$ et d_{avg} ne sont pas définis pour une requête q_k, pour le calcul de $tf_{k,j}$, le paramètre b est égal à 0. Pour le $tf_{i,j}$ d'un document d_i et d'un terme t_j, ce paramètre b est fixé à 0,5. Par ailleurs, aucun anti-dictionnaire n'a été utilisé, et une lemmatisation des mots est réalisée à l'aide de l'algorithme de Porter (Porter (1980)). Le descripteur SIFT a été utilisé afin de construire le vocabulaire de mots visuels, dont la taille, correspondant au paramètre k dans l'algorithme des k-means, a été fixée à $10\,000$.

4.3 Evaluation des résultats

Plusieurs expérimentations ont été réalisées sur l'échantillon test ImageCLEF 2009, en considérant différents vocabulaires. Afin d'obtenir une base de comparaison, des premières expérimentations ont été menées en considérant un seul vocabulaire, textuel ou visuel. Ces expérimentations notées respectivement T et V_{sift} serviront de référence et seront comparées à celles obtenues en utilisant le modèle de recherche d'information multimodale.

Pour évaluer les performances de notre système, nous avons utilisé la mesure classique MAP (*Mean Average Precision*, cf. Kamps et al. (2008)) et le rappel R correspondant au nombre de documents pertinents retrouvés par le système sur le nombre de documents pertinents à retrouver dans la collection. Plus les valeurs de ces mesures sont élevées, meilleurs sont les résultats. Pour une requête donnée, ces mesures sont calculées à l'aide des $1\,000$ premiers documents retournés pas le système.

La mesure MAP est globale et permet d'évaluer les performances du système. Elle est notamment utilisée pour classer les participants de la compétition ImageCLEF (Tsikrika et Kludas (2008, 2009)).

5 Résultats

5.1 Vocabulaire textuel ou visuel

La table 2 résume les différents résultats obtenus pour le MAP en fonction des modalités (texte/image) utilisées.

Selon la mesure MAP, l'utilisation de la modalité visuelle seule conduit sans surprise, à de moins bons résultats (MAP de 0,0083) par rapport à ceux obtenus en utilisant seulement le texte (MAP de 0,1661). La mesure du rappel confirme ce résultat et montre clairement que la modalité textuelle permet de retrouver la majorité des documents pertinents (73%) alors que la modalité visuelle ne permet d'en retrouver que 11%. Dans la suite, l'expérimentation

Expérimentations	MAP	Rappel
T	0,1661	0,7336
V_{sift}	0,0083	0,1078

TAB. 2 – *Résultats de référence obtenus sur la collection ImageCLEF 2009 en considérant la mesure MAP et le rappel.*

n'exploitant que l'information textuelle nous servira de référence pour comparer les résultats qui combinent les deux modalités textuelles et visuelles.

5.2 Apprentissage par analyse linéaire discriminante

Les résultats obtenus sur la collection ImageCLEF 2009 en utilisant les paramètres de la combinaison linéaire calculés grâce à l'ALD sur la collection ImageCLEF 2008 sont présentés dans la table 3. Notons que comme il n'y a qu'une seule description visuelle, la combinaison linéaire des scores définie dans notre modèle ne dépend que du coefficient $\alpha_{V_{sift}}$ de la modalité visuelle. Le coefficient de la modalité textuelle est $1 - \alpha_{V_{sift}}$.

Expérimentations	MAP	Rappel
T	0,1661	0,7336
T et V_{sift} : $\alpha_{V_{sift}} : 0,059098$	0,1795	0,7515

TAB. 3 – *Résultats obtenus sur la collection 2009 en utilisant les coefficients calculés par ALD pour deux modalités à partir de la collection 2008.*

Comme nous pouvons le constater sur la table 3, même si l'utilisation de l'information visuelle seule ne conduit pas à de bons résultats, en la combinant avec l'information textuelle, elle permet d'améliorer les résultats de référence. En effet, la combinaison de la modalité textuelle à celle du descripteur visuel V_{sift} permet d'augmenter le MAP de 0,1661 à 0,1795. De plus, cette combinaison permet de retrouver de nouveaux documents pertinents avec une amélioration du rappel de 73% à 75%.

La significativité de ces résultats a été contrôlée à l'aide de tests statistiques. Le test de Student apparié unilatéral a été réalisé sur les précisions moyennes des 45 requêtes de la collection 2009. Ce test montre que la probabilité critique (*p value*) est de 0,023290 pour l'expérimentation combinant la modalité textuelle et le descripteur $sift$, ce qui conduit en prenant un risque de 5% à refuser l'hypothèse d'égalité des moyennes (Saporta (2006)). Les améliorations des résultats combinant une modalité textuelle et une modalité visuelle peuvent donc être considérées comme significatives.

La figure 2 montre les valeurs du MAP obtenues pour différentes valeurs de α sur la collection ImageCLEF 2009. Nous pouvons remarquer d'une part la forte influence du paramètre α sur le MAP et d'autre part, que la valeur idéale de α est de 0,084 alors que la valeur $\alpha_{V_{sift}}$ apprise sur la collection ImageCLEF 2008 est de 0,059. Ces deux valeurs sont relativement proches et les valeurs des MAP correspondants sont encore plus proches (0,1822 contre

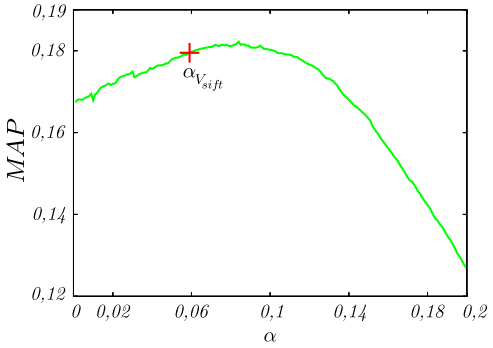

FIG. 2 – *MAP sur les requêtes 2009 pour une valeur du paramètre α comprise entre 0 et 0,2.*

0,1795). Ces résultats confirment l'importance de la combinaison et l'efficacité de la méthode d'apprentissage à l'aide de l'analyse linéaire discriminante de Fisher.

6 Conclusion

Dans le contexte de la recherche d'information multimédia, nous avons introduit une nouvelle méthode basée sur l'analyse linéaire discriminante de Fisher pour apprendre les valeurs des paramètres de combinaison d'un modèle de recherche d'information qui combine linéairement les deux informations textuelles et visuelles. Ces paramètres permettent de pondérer l'importance associée aux différents types d'informations.

Nos expérimentations ont été menées sur la collection ImageCLEF, extraite de l'encyclopédie Wikipédia, et composée de plus de 150 000 documents contenant du texte et une image. Elles montrent que notre méthode est efficace pour l'apprentissage du coefficient de pondération entre les informations textuelles et visuelles. D'une part les résultats avec fusion sont significativement meilleurs que ceux obtenus à l'aide de la modalité textuelle seule et d'autre part la valeur du coefficient apprise est proche du coefficient optimal.

L'intérêt de notre méthode est qu'elle propose une approche rigoureuse et algorithmiquement efficace pour déterminer des paramètres de combinaison qui sont souvent fixés arbitrairement ou déterminés à partir d'expérimentations coûteuses et pas toujours bien fondées. Elle peut s'appliquer à la combinaison d'un nombre arbitraire de modalités représentées à partir de vecteurs tf-idf, à condition de disposer d'un échantillon d'apprentissage représentatif et de taille suffisante. Chaque modalité peut être représentée par plusieurs sacs de mots issus de descripteurs différents. Par exemple, pour la modalité visuelle pour laquelle il n'existe pas de description universelle, on peut choisir de combiner plusieurs descripteurs de texture, de couleur, ou de forme.

En perspective, il serait justement intéressant d'évaluer les performances de l'apprentissage en combinant un descripteur textuel et plusieurs descripteurs visuels. On peut aussi envisager de multiplier les descripteurs textuels pour ajouter par exemple des informations de structure ou alors considérer d'autres modalités comme le son ou la vidéo en leur ajoutant des descripteurs adaptés. Sur le plan méthodologique, une autre approche à explorer serait d'adapter

les coefficients de combinaison en fonction des requêtes. En effet, une analyse plus fine de nos résultats montre que le bénéfice de la combinaison n'est pas identique en fonction des requêtes : la combinaison améliore significativement les résultats de certaines requêtes alors qu'elle est moins efficace voire néfaste pour d'autres. Ainsi, si l'on était capable de faire une pré-classification des requêtes suivant leur caractère visuel par exemple, on pourrait apprendre des paramètres de combinaison qui leur sont spécifiques.

Références

Ayache, S., G. Quénot, et J. Gensel (2007). Classifier fusion for SVM-based multimedia semantic indexing. In *ECIR'07 : 29th European Conference on Information Retrieval*, pp. 494–504.

Bouveyron, C., S. Girard, et C. Schmid (2005). Analyse discriminante de haute dimension. Technical Report 5470, INRIA.

Csurka, G., C. Dance, L. Fan, J. Willamowski, et C. Bray (2004). Visual categorization with bags of keypoints. In *ECCV'04 : 8th European Conference on Computer Vision : workshop on Statistical Learning in Computer Vision*, pp. 59–74.

Fei-Fei, L. et P. Perona (2005). A bayesian hierarchical model for learning natural scene categories. In *CVPR'05 : IEEE computer society conference on Computer Vision and Pattern Recognition*, pp. 524–531.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics 7*(2), 179–188.

Kamps, J., J. Pehcevski, G. Kazai, M. Lalmas, et S. Robertson (2008). INEX 2007 evaluation measures. In *INEX'07 : Focused access to XML documents, 6th Workshop of the Initiative for the Evaluation of XML Retrieval*, Berlin, Heidelberg, pp. 24–33.

Klecka, W. (1980). *Discriminant analysis*, Volume 19. Sage Publications, Inc.

Lebart, L., A. Morineau, et J. Fénelon (1986). *Traitement des Données Statistiques*. Dunod.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60*(2), 91–110.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India 2*(1), 49–55.

Manning, C., P. Raghavan, et H. Schtze (2008). *Introduction to information retrieval*. Cambridge University Press.

Mika, S., G. Ratsch, J. Weston, B. Scholkopf, et K. Mullers (1999). Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX. Proceedings of the IEEE Signal Processing Society Workshop*, pp. 41–48. IEEE.

Moulin, C., C. Barat, C. Lemaître, M. Géry, C. Ducottet, et C. Largeron (2009). Combining text/image in WikipediaMM task 2009. In *CLEF'09 : 10th workshop of the Cross-Language Evaluation Forum*, pp. 164–171.

Moulin, C., C. Largeron, et M. Géry (2010). Impact of visual information on text and content based image retrieval. In *S+SSPR'10 : 13th international workshop on Structural, Syntactic, and Statistical Pattern Recognition*, pp. 159–169.

Nowak, E., F. Jurie, et B. Triggs (2006). Sampling strategies for bag-of-features image classification. In *ECCV'06 : 9th European Conference on Computer Vision : workshop on Statistical Learning in Computer Vision*, pp. 490–503.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program 14*(3), 130–137.

Robertson, S. E., S. Walker, M. Hancock-Beaulieu, A. Gull, et M. Lau (1994). Okapi at TREC-3. In *Text REtrieval Conference*, pp. 21–30.

Salton, G., A. Wong, et C. S. Yang (1975). A vector space model for automatic indexing. *Communations of the ACM 18*(11), 613–620.

Saporta, G. (2006). *Probabilités, analyses des données et statistique* (2e édition révisée et augmentée ed.). Éditions Technip.

Snoek, C., M. Worring, et A. Smeulders (2005). Early versus late fusion in semantic video analysis. In *MM'05 : 13th ACM international conference on MultiMedia*, pp. 399–402.

Tirilly, P., V. Claveau, et P. Gros (2008). Language modeling for bag-of-visual words image categorization. In *CIVR'08 : international conference on Content-based Image and Video Retrieval*, pp. 249–258.

Tollari, S., M. Detyniecki, C. Marsala, A. Fakeri-Tabrizi, M.-R. Amini, et P. Gallinari (2009). Exploiting visual concepts to improve text-based image retrieval. In *ECIR'09 : Proceedings of European Conference on Information Retrieval*.

Tollari, S. et H. Glotin (2007). Web image retrieval on imageval : Evidences on visualness and textualness concept dependency in fusion model. In *CIVR'07 : ACM International Conference on Image and Video Retrieval*.

Tsikrika, T. et J. Kludas (2008). Overview of the wikipediaMM task at ImageCLEF 2008. In *CLEF'08 : Evaluating Systems for Multilingual and Multimodal Information Access,9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark.

Tsikrika, T. et J. Kludas (2009). Overview of the wikipediaMM task at ImageCLEF 2009. Technical report, 10th Workshop of the Cross-Language Evaluation Forum.

van Rijsbergen, C. K. (1979). *Information retrieval*. Butterworth-Heinemann ; 2nd edition.

Zhai, C. (2001). Notes on the lemur tfidf model. Technical report, Carnegie Mellon University.

Summary

Due to the expansion of digital processes, large amount of documents composed of text and images are now shared. It raises the need for the development of models allowing to efficiently exploit such multimedia data. In the context of information retrieval, a common strategy consists in representing textual and visual information separately and fusing the scores obtained with each single media using a linear combination. A main question in this approach is how to calculate weights that balance the importance given to each modality. The aim of this article is to present a new approach for learning these weights thanks to the linear discriminant analysis of Fisher (LDA). Experimentations performed on the ImageCLEF collection show that learning the weights with the LDA is relevant and leads to an improvement when combining visual and textual scores against those obtained with a single modality.