Analyse statistique de similarité dans une collection d'images

Madenda Sarifuddin**, Rokia Missaoui*, Jean Vaillancourt*
Youssef Hamouda** et Marek Zaremba*

* Département d'informatique et d'ingénierie, Université du Québec en Outaouais
e-mail: rokia.missaoui/jean.vaillancourt/marek.zaremba@uqo.ca
** Département d'informatique, Université du Québec à Montréal
C.P. 8888, succursale Centre-Ville, Montréal, Québec, Canada, H3C 3P8
e-mail: madenda@info.uqam.ca ; hamouda.youssef@courrier.uqam.ca

Résumé. Dans le cadre du développement d'un système de recherche d'images par le contenu, nous avons défini deux nouvelles mesures : la distance de dissimilitude DS*, et la distance de similarité E. La seconde est intégrée à la formule de la distribution de Gibbs et de celle de la mixture de Dirichlet généralisée, alors que la première est comparée à trois autres variantes d'estimation de la similarité : la distance Euclidienne ainsi que les distributions de Gibbs et Dirichlet intégrant la distance de similarité E. L'analyse empirique des quatre mesures de similarité porte sur les histogrammes de couleurs d'une collection d'images et montre que l'efficacité de la recherche, mesurée par le rappel et la précision, est la plus importante pour la distribution de Dirichlet modifiée et la plus faible pour la distance Euclidienne.

Mots-clés. Recherche d'images par le contenu, mesures de similarité, distribution de Dirichlet, distribution de Gibbs.

1 Introduction

La caractéristique visuelle la plus fréquemment utilisée dans la recherche d'images par le contenu est la couleur. Cette dernière est relativement robuste et indépendante de la taille et de la transformation géométrique de l'image. Les valeurs de la couleur sont d'habitude calculées dans les espaces de couleur correspondant à la perception humaine tels que HSV, L*a*b* ou L*u*b*, soit d'une manière approximative par identification d'un éventail de couleurs présentes dans l'image, soit d'une manière plus précise sous forme d'histogrammes.

Dans (Missaoui et al. 2003), nous proposons une approche de recherche d'images par le contenu (Rui et al. 1999) qui effectue deux types d'analyse de similarité : une analyse approximative (*coarse-grain*) faisant appel à une technique de regroupement conceptuel appliquée aux caractéristiques visuelles (couleurs et formes) exprimées sous forme d'une relation binaire entre une collection d'images et leurs propriétés, et une analyse fine (*fine-grain*) qui effectue un calcul de similarité sur des histogrammes de couleurs. Dans les deux types d'analyse, les caractéristiques visuelles sont automatiquement extraites à partir des images. La première analyse vise à faire de la fouille de données (*data mining*) sur une collection d'images en identifiant des concepts (ex. les images 2 et 6 possèdent le bleu, le vert et des lignes) et des règles d'association (ex. si des images comportent des courbes, alors il y a 60% de chances pour qu'elles puissent aussi contenir la couleur verte).

Dans cet article, nous nous concentrons sur le deuxième type d'analyse fréquemment appliqué en traitement d'images (Faloutsos et al 1994) en proposant deux nouvelles mesures : la distance de dissimilitude DS*, et la distance de similarité E. La seconde sera intégrée à la formule de la distribution de Gibbs et de celle de la mixture de Dirichlet généralisée pour obtenir une précision plus grande, alors que la première est une composante de E et sera comparée à trois autres variantes d'estimation de la similarité : la distance Euclidienne et les distributions de Gibbs et Dirichlet intégrant la distance de similarité E.

L'analyse empirique de ces mesures est effectuée sur une collection de plus de deux mille images de diverses catégories et vise à identifier la méthode possédant la capacité de discrimination la plus forte.

L'article est structuré comme suit. La section 2 donne un bref aperçu du processus d'extraction des couleurs. Dans la section 3, on décrit quelques mesures de similarité, alors que la section 4 présente l'analyse empirique de quatre mesures de similarité et la section 5 fournit une conclusion.

2 Extraction des caractéristiques visuelles

Les principaux attributs de l'image sont la couleur, la forme et la texture. Ces attributs peuvent être utilisés individuellement ou en combinaison avec d'autres.

Dans ce qui suit, nous nous limitons à l'extraction des couleurs dans l'espace des couleurs L*a*b* où la luminosité L détermine l'éclairage de l'image ainsi que les proportions du blanc et du noir, la composante *a** fait varier la couleur entre le vert et le rouge, et la composante *b** représente les couleurs entre le bleu et le jaune. Toutes les autres couleurs sont obtenues par une combinaison de ces trois composantes. La figure 1.b représente la distribution des pixels selon les couleurs visibles (la chromaticité) de l'image dans les coordonnées 2D a* et b* au niveau blanc nominal de la luminance L, alors que la figure 2 exprime cette même distribution dans les coordonnées 3D cylindrique L*a*b*. On voit ainsi que la distribution de chaque couleur est bien concentrée sur un certain intervalle d'angle de teinte θ (voir équation 2), ce qui nous permet de distinguer et de vérifier la présence et l'absence de certaines couleurs dans l'image.

(a) 

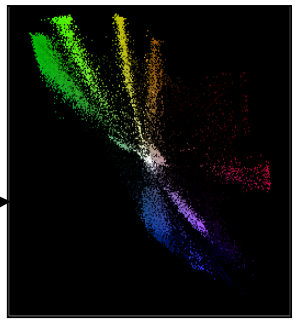 *(b)*

FIG. 1 - *Distribution des couleurs dans l'espace L*a*b**

Cette information de présence et d'absence des couleurs est exprimée par un vecteur de tons de couleurs pouvant être utilisé pour l'analyse approximative de similarité. Au sein d'un

même groupe d'images approximativement similaires, il est possible de faire une analyse fine de similarité sur des histogrammes 3D de forme cylindrique.

3 Mesures de similarité

La similarité (Santini *et al*., 1999) peut être calculée par le rapport entre la similitude et la dissimilitude de deux images. La similitude mesure le nombre de correspondances tandis que la dissimilitude indique le nombre de différences. Deux objets ou images sont parfaitement similaires si leur similitude est très grande et leur dissimilitude tend vers zéro, ou encore lorsque le rapport entre les deux mesures tend vers l'infini. Les mesures de similarité les plus souvent utilisées pour la recherche d'images par le contenu sont la distance L_1 (norme) et L_2 (Euclidienne*)* lesquelles se basent sur la dissimilitude.

Dans cette partie, nous illustrons d'abord les limites de la distance Euclidienne pour la recherche d'images multi-couleurs. Ensuite, nous proposons une nouvelle mesure de similarité pour les données à variables multiples ou des images multi-couleurs. Cette mesure, appelée *distance de similarité*, sera incorporée à chacune des deux distributions : Gibbs et Dirichlet en vue de rendre plus précis le processus de détermination des images les plus similaires à l'image soumise par l'utilisateur.

3.1 Calcul de dissimilitude et de similitude

Comme indiqué dans la section 2, la distribution des couleurs de l'image est représentée dans le plan 3D L*a*b* de forme cylindrique et exprimée par l'histogramme $V(L,t,\theta)$ où t est le chroma métrique et θ est l'angle métrique de teinte :

$$t = \left(a^{*2} + b^{*2}\right)^2 \qquad (1)$$

$$\theta = ar\tan(b^* / a^*) \qquad (2)$$

Ainsi, la comparaison de deux images peut se faire sur la base de leurs histogrammes. Cela revient à comparer la probabilité d'apparition des pixels ayant la même couleur à chacune des positions des deux histogrammes. Le temps de calcul relié à la comparaison d'histogrammes peut cependant être prohibitif. Pour remédier à cette situation, il est possible de faire un découpage (ou quantification d'histogramme) des axes en intervalles.

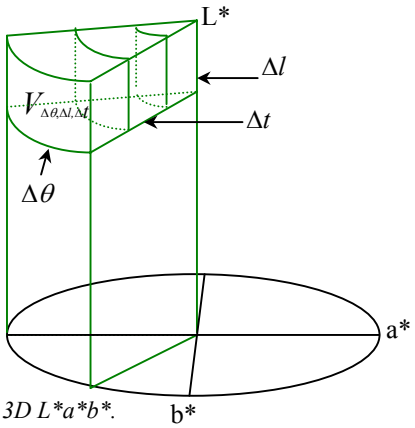

FIG. 2 - *Histogramme 3D L*a*b**.

Le choix du nombre de quantifications a un impact sur la précision et le coût de comparaison des histogrammes. Dans (Wand 1996), l'auteur propose une méthode de détermination d'un découpage optimal qui dépend de la taille des échantillons, ou dans notre cas, de la taille de l'image, et de la variance. En se basant sur cette étude, nous avons décomposé l'histogramme (figure 2) en 3468 sous-histogrammes de 17 coupes de couleur $\Delta\theta$, 12 coupes de chroma Δt et 17 coupes pour la luminance Δl. Le nombre de découpages peut varier selon la résolution d'image.

Finalement, chaque histogramme peut être représenté par un vecteur $\mathbf{V}=(V_{1,1,1}, \ldots, V_{k,m,n}, \ldots, V_{17,17,12})$, où la valeur $V_{k,m,n}$ correspond au pourcentage de pixels ayant la couleur *k*, le chroma *n* et la luminance *m* dans l'intervalle $k\Delta\theta$, $m\Delta l$, et $n\Delta t$.
La comparaison de deux images X et Y peut être faite sur la base d'une mesure de similarité telles que la distance L_1 et L_2 (distance Euclidienne) comme suit :

$$\textit{Distance } L_1 : \qquad L_1(X,Y) = \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{n=1}^{N}\left|V_{k,m,n}^{X} - V_{k,m,n}^{Y}\right| \tag{3}$$

$$\textit{Distance } \ L_2 : \qquad L_2(X,Y) = \sum_{k=1}^{K}\left(A_c \sum_{m=1}^{M} B_c \sum_{n=1}^{N}\left(V_{k,m,n}^{X} - V_{k,m,n}^{Y}\right)^2\right)^{1/2} \tag{4}$$

avec $\left|\sum_{m=1}^{M}\sum_{n}^{N} V_{k,m,n}^{X} - \sum_{m=1}^{M}\sum_{n}^{N} V_{k,m,n}^{Y}\right| \le A_c \le 1$ et $\left|\sum_{n}^{N} V_{k,m,n}^{X} - \sum_{n}^{N} V_{k,m,n}^{Y}\right| \le B_c \le 1$.

$k=\{1, 2, \ldots, 17\}$, $m=\{1, 2, \ldots, 17\}$, $n=\{1, 2, \ldots, 12\}$.

B_c correspond à la distance Euclidienne de chaque couleur ayant le même intervalle de luminance et A_c représente la distance Euclidienne de chaque couleur quelque soit sa luminance. Ces deux valeurs peuvent être choisies par l'utilisateur comme le poids indiquant l'importance de la couleur et de la luminance pour la recherche d'image. Les images les plus similaires à l'image requise sont donc celles qui ont une valeur L_1 ou L_2 proche de zéro.

Le résultat de la recherche d'images par calcul des distances L_1 et L_2 est relativement le même. La figure 3.a illustre le résultat obtenu par la distance Euclidienne avec $A_c = B_c = 1$. L'image sur le coin en haut à gauche est l'image requête, alors que les images cibles (résultant de la requête) sont ordonnées par similarité et placées de gauche à droite et du haut vers le bas. Nous constatons que bien que l'image à l'intersection de la ligne 2 et la colonne 4 soit loin d'être similaire à l'image requête, elle est classée 7ème. Par contre, l'image à l'intersection de la ligne 3 et la colonne 3 est classée au 10ème rang alors que visiblement, elle est plus similaire à l'image requête.

(a)   (b)

FIG. 3 - *Résultats de recherche : (a) par la distance Euclidienne, (b) par la dissimilitude DS*.*

L'exemple précédent illustre la limite de la distance Euclidienne pour la recherche d'images. Le phénomène peut être expliqué par l'exemple du tableau 1. Les colonnes 1 et 2 représentent respectivement la couleur (bleu, rouge, violet, jaune et vert) dans l'image requête et son pourcentage d'apparition, tandis que les colonnes subséquentes décrivent le pourcentage d'apparition d'une couleur dans les images de la base de données. Les lignes L_1 et L_2 fournissent des distances entre l'image requête et chaque image de la base de données.

Elles indiquent que l'image 3 est la plus similaire à l'image requête bien que l'image 3 ne comporte pas la couleur rouge mais inclut la couleur jaune. Par contre, les images 1 et 2 ont les mêmes couleurs que l'image requête et sont visiblement plus similaires à cette dernière, mais sont classées après l'image 3.

Couleur	Image requête	Base d'images				
		Image 1	Image 2	Image 3	. . .	Image N
Bleu	0.50	0.50	0.70	0.50		0.45
Rouge	0.10	0.20	0.15			0.05
Violet	0.30	0.15	0.10	0.30		0.25
Jaune				0.10		0.20
Vert	0.10	0.15	0.05	0.10		0.05
L_1	0	0.30	0.50	0.20		0.40
L_2	0	0.187	0.235	0.141		0.224
$DS^*_{q=1}$	0	0.30	0.50	0.40		0.60
$DS^*_{q=2}$	0	0.187	0.235	0.283		0.412

TAB. 1.

Ceci nous incite à proposer une nouvelle mesure de similarité, appelée distance de similarité, qui serait plus précise que les distances L_1 et L_2 et qui tiendrait compte à la fois de la similitude et de la dissimilitude.

La similitude S_c entre deux images *X* et *Y* au niveau de la couleur *c* mesure la proportion de la couleur *c* commune aux deux images (Cf. formule 8).

Nous définissons la *distance de dissimilitud*e DS_c^* de la couleur *c* comme étant :

$$DS_c^* = L_c + D_c \tag{5}$$

où L_c représente une distance (ex. L_1 ou L_2) entre les probabilités $P_X(c)$ et $P_Y(c)$ de présence de la couleur *c* dans les images *X* et *Y* respectivement.

Ainsi, la distance DS_c^* incorpore la différence entre deux images au niveau de la couleur *c* par la distance L_1 ou L_2, et la valeur de D_c indiquant si la couleur *c* est présente ou absente dans les deux images (valeur nulle) ou dans l'une d'elles seulement (valeur non nulle) :

$$\begin{aligned} &D_c = max\left(P_x(c), P_y(c)\right) \;\; si \;\; P_x(c) = 0 \;\; ou \;\; P_y(c) = 0 \\ &D_c = 0, \text{ sinon} \end{aligned} \tag{6}$$

D_c sera positive chaque fois que la couleur *c* se trouve seulement dans l'une des deux images X et Y, et nulle autrement.

Par l'insertion de l'équation (6) dans (5) et en tenant compte de l'ensemble des couleurs, nous déterminons la *distance de dissimilitud*e de deux images *X* et *Y*, appelée *DS**, de manière générale comme suit :

$$DS^* = \left(\sum_c DS_c^{*q}\right)^{1/q} = \left[\sum_c \left(\left|P_x(c) - P_y(c)\right| + max\left(P_x(c), P_y(c)\right)_{si\ P_X(c)=0\ ou\ P_Y(c)=0}\right)^q\right]^{1/q} \quad (7)$$

La mesure DS* est une distance semi-métrique puisqu'il s'agit d'une application de $E \times E \rightarrow R_+$ où E représente l'espace des vecteurs de caractéristiques de couleur de l'image, et tel que $\forall X, Y \in E$:

- $DS^*(X, Y) \geq 0$;
- $DS^*(X, Y) = 0 \Rightarrow X = Y$;
- $DS^*(X, Y) = DS^*(Y, X)$.

La première propriété est vraie car sur la base des formules 6 et 7, DS*(X, Y) est toujours positive ou nulle. Comme L_c est une distance entrant dans la composition de DS_c^*, la vérification de la deuxième propriété nécessite de démontrer que lorsque D_c est nulle pour une paire (X, Y), alors X = Y. Ceci est vrai par la définition même de D_c puisque qu'une valeur nulle de cette dernière indique que la couleur *c* est soit présente soit absente simultanément dans les deux images X et Y. La dernière propriété est évidente.

En consultant la ligne DS* du tableau 1, on note que le classement des images les plus similaires devient : 1, 3, 2, N pour q=1 et 1, 2, 3, et N pour q=2.

La figure 3.b montre le résultat de la recherche d'images obtenu par la dissimilitude *DS** pour q=2. Nous pouvons voir qu'il y a trois images de la figure 3.a qui sont filtrées et n'apparaissent pas dans la figure 3.b et sont remplacées par trois autres images qui sont davantage similaires à l'image requête. Cela nous incite à croire que la distance *DS** est probablement plus précise que la distance L_1 et L_2.

La question qui se pose maintenant est la suivante : quelle est la meilleure mesure de similarité ? Est-ce la dissimilitude seule, la similitude seule, ou une combinaison des deux ?

Supposons que S_c représente le nombre de similitudes de la couleur *c* dans deux images *X* et *Y,* alors elle peut être exprimée en fonction de la distance L_c (ex. L_1 ou L_2) comme suit :

$$S_c = max\left(P_x(c), P_y(c)\right) - L_c \quad (8)$$

Une formule plus générale de S_c est exprimée par l'équation 9 qui indique que si la couleur *c* est présente dans deux images X et Y (c-à-d $P_X(c)>0$ *et* $P_Y(c)>0$) et si la distance L_c tend vers zéro, alors la valeur $S_c = P_X(c) = P_Y(c)$ et les deux images sont parfaitement similaires au niveau de la couleur *c*. Inversement si la couleur *c* est absente dans une de deux images, la valeur S_c est égale à zéro et donc les deux images sont parfaitement différentes au niveau de la couleur *c*.

$$S_c = \left(max\left(P_X(c), P_Y(c)\right) - \left|P_X(c) - P_Y(c)\right|\right)^q \quad (9)$$

Le rapport entre la dissimilitude et la similitude de la couleur *c* est décrit par l'équation :

$$R_c = DS_c^* / S_c \quad (10)$$

Nous voyons que si la valeur R_c tend vers zéro ($DS_c^*=0$ *et* $S_c=1$), les deux couleurs sont parfaitement similaires alors qu'elles sont complètement différentes lorsque R_c tend vers l'infini ($DS_c^*=1$ *et* $S_c=0$). Donc pour le calcul de similarité basé sur la valeur R_c, on doit se contraindre au cas de $S_c > DS_c^*$.

La nouvelle mesure E_c est une distance de similarité entre deux images pour une couleur donnée *c*. La formule 11 indique que si la valeur de similitude S_c est inférieure ou égale à la dissimilitude DS_c^*, la deuxième partie de cette équation est négligée et la formule est égale à l'équation 5. Par contre, si S_c est plus importante que DS_c^*, la distance de similarité E_c va décroître selon la croissance de la valeur S_c. Nous remarquons aussi que lorsque $P_X(c) = P_Y(c)=0$, nous avons $DS_c = S_c = 0$ *et* $E_c = 0$. *Par contre,* si $P_X(c) >0$, $P_Y(c)>0$ *et* $DS_c \geq S_c$, nous avons $E_c = L_1$ (pour q=1) ou L_2 (pour q=2).

$$E_c = DS_c^*\left(1 + log_{10}(R_c)\right) \quad \text{ou} \quad E_c = DS_c^* + DS_c^*.log_{10}\left(DS_c^*/S_c\right)_{S_c > DS_c^* > 0} \quad \text{et}$$

$$E = \left(\sum_c E_c\right)^{1/q}{}_{E_c>0} \tag{11}$$

En suivant le même raisonnement que pour DS_c^*, il est possible de démontrer que E_c est également une distance semi-métrique.

3.2 Distribution de Gibbs

La distribution de Gibbs (*Gibbs random field*) est très populaire en physique statistique et bien répandue dans le domaine du traitement d'image en rehaussement d'image, analyse de texture, et comparaison d'images. Dans (Rémillard et al. 1999), on se sert de la distribution de Gibbs pour modéliser les images et les structures spatiales sous-jacentes, et on démontre le pouvoir de discrimination d'une telle approche dans la recherche d'images similaires. L'approche semble très bien appropriée pour les images en niveaux de gris et sans transformation géométrique importante.

Dans cette partie, nous présentons une méthode de calcul de similarité par la distribution de Gibbs en utilisant la mesure de similarité E, proposée dans la partie 3.1.

Tel indiqué précédemment, la recherche d'image peut être basée sur l'évaluation du rapport entre la similitude S_c et la dissimilitude DS_c^* pour chaque couleur *c* prise en considération. En outre, la variation de l'angle θ de l'histogramme 3D correspond à la variation de couleur. Autrement dit, une partition avec un grand intervalle d'angle θ est équivalente à une séparation de couleurs. Donc, une première hypothèse est l'indépendance de chaque partition de couleur selon l'angle θ. Par ailleurs, la variation lente de la luminance ou du chroma ne provoque pas de changement brusque de la vision de la couleur. Ceci nous amène à la deuxième hypothèse qui stipule que la similitude d'une couleur peut être aussi vérifiée par la luminance et le chroma voisins.

Sur la base de ces deux hypothèses et par généralisation des formules 7 à 9, nous déterminons alors les mesures de similitude et de dissimilitude pour chaque partition *k,m,n* en prenant $q = 2$. Pour ne pas avoir une grande différence de similarité entre la couleur requête et la couleur ciblée vis-à-vis de la luminance et du chroma, nous nous sommes limités au voisinage d'ordre 1.

$$S_{k,m,n} = \begin{cases} max\left(\left[max(V^X_{k,m,n}, V^Y_{k,j,i}) - |V^X_{k,m,n} - V^Y_{k,j,i}|\right], \left[max(V^X_{k,j,i}, V^Y_{k,m,n}) - |V^X_{k,j,i} - V^Y_{k,m,n}|\right]\right)^2 \\ \text{si } max(V^X_{k,m,n}, V^Y_{k,j,i}) > 0,\ j=(m-1,m,m+1),\ i=(n-1,n,n+1) \\ 0 \quad si \quad max(V^X_{k,m,n}, V^Y_{k,j,i}) = 0, \end{cases} \tag{12}$$

$$DS*_{k,m,n} = \begin{cases} DE_{k,m,n} + DC_{k,m,n} & si \quad V^X_{k,m,n} = 0 \ ou \ V^Y_{k,m,n} = 0 \\ DE_{k,m,n} & \text{sinon} \end{cases} \tag{13}$$

$$\text{où} \quad DE_{k,m,n} = (V^X_{k,m,n} - V^Y_{k,m,n})^2 \quad et \quad DC_{k,m,n} = max(V^X_{k,m,n}, V^Y_{k,m,n})^2$$

La partition pour chaque couleur k et chaque intervalle de luminance m est nommée $V_{k,m}$. Le rapport entre la dissimilitude et la similitude pour chaque partition $V_{k,m}$ est donné par :

$$RS_{k,m} = DS^*_{k,m} / S_{k,m} \tag{14}$$

où

$$S_{k,m} = \left[max\left(\sum_{n=1}^{N} V^X_{k,m,n}, \sum_{n=1}^{N} V^Y_{k,m,n}\right)^2 - \left(\sum_{n=1}^{N} V^X_{k,m,n} - \sum_{n=1}^{N} V^Y_{k,m,n}\right)^2 \right]^{1/2} + \left(\sum_{n=1}^{N} S_{k,m,n}\right)^{1/2} \tag{15}$$

$$DS*_{k,m} = \begin{cases} \left(DE_{k,m} + DC_{k,m}\right)^{1/2} + \left(\sum_n DS*_{k,m,n}\right)^{1/2} & si \ V^X_{k,m} = 0 \ ou \ V^Y_{k,m} = 0 \\ \left(DE_{k,m}\right)^{1/2} + \left(\sum_n DS*_{k,m,n}\right)^{1/2} & \text{sinon} \end{cases} \tag{16}$$

avec $V_{k,m} = \sum_n V_{k,m,n}$ et $DE_{k,m}$ fait référence à la distance Euclidienne pour chaque partition $V_{k,m}$.

Ensuite, pour calculer et analyser les images les plus similaires à l'image requête, nous introduisons l'équation 11 dans la distribution de Gibbs comme suit :

$$P(X, Y_j) = \frac{exp(-E(j))}{\sum_{i=1}^{J} exp(-E(i))} \tag{17}$$

$$E(j) = \sum_{k=1}^{K} \sum_{m=1}^{M} e^{-\left(\alpha_{k,m} + DS^*_{k,m}\right)} \quad avec \quad \alpha_{k,m} = DS^*_{k,m}.log_{10}\left(DS^*_{k,m} / S_{k,m}\right)_{S_{k,m} > DS^*_{k,m} > 0} \tag{18}$$

où j correspond à la $j^{ème}$ image à comparer parmi l'ensemble J d'images. La figure 4.c montre le résultat de la recherche d'images par la distribution de Gibbs.

Il est important de noter que nous n'utilisons pas le plein sens physique d'une distribution de Gibbs, mais uniquement le pouvoir discriminant qu'apporte l'exponentiation des erreurs.

3.3 Distribution de Dirichlet

La distribution de Dirichlet est une généralisation multivariante de la distribution bêta offrant une grande flexibilité et une facilité d'application. En vue de remédier aux inconvénients de

la Gaussienne, on propose dans (Bouguila, 2002) une généralisation de la distribution de Dirichlet et on montre que la nouvelle formule aboutit à une meilleure classification des données que la mixture Gaussienne.
Dans cette partie, nous appliquons la distribution de Dirichlet à la recherche d'images similaires. La distribution Dirichlet, ayant les variables $\mathbf{X}=(X_1, \ldots, X_I)$ et les paramètres $\boldsymbol{\alpha}=(\alpha_1, \ldots, \alpha_I)$, est donnée par :

$$P(X_1,...,X_I) = \frac{\Gamma(|\vec{\alpha}|)}{A^{|\vec{\alpha}|}\prod_{i=1}^{I}\Gamma(\alpha_i)}\prod_{i=1}^{I} X_i^{\alpha_i - 1} \tag{19}$$

où $|\vec{\alpha}| = \sum_{i=1}^{I}\alpha_i, \quad \alpha_i > 0$ et $\sum_{i=1}^{I} X_i < 1, \quad 0 < X_i < A \quad \forall i = 1,...,I$

Dans ce cas, la variable X_i peut être considérée comme la variable de couleur à la partition $i=(m,k)$ et le paramètre α_i comme une fonction de la distance de similarité proposée dans l'équation (11).

$$X_i = max\left(V_{k,m}^{X}, V_{k,m}^{Y}\right) \tag{20}$$

$$\alpha_i = e^{-\left(\alpha_{k,m} + DS_{k,m}^{*}\right)} \quad avec \quad \alpha_{k,m} = DS_{k,m}^{*}.log_{10}\left(DS_{k,m}^{*}/S_{k,m}\right)_{S_{k,m}>DS_{k,m}^{*}>0} \tag{21}$$

où $i= 1, \ldots, M.K$=289, $DS^*_{k,m}$ et $S_{k,m}$ sont donnés par les équations 14 et 15 respectivement.

La figure 4.d montre le résultat de l'analyse de similarité basée sur la distribution de Dirichlet.

3.4 Analyse préliminaire des variantes

Dans cette partie, nous illustrons d'une manière *qualitative* l'intérêt de chacune des quatre méthodes d'analyse de similarité. Une étude plus détaillée sera fournie en section 4.2 et vise à illustrer d'une façon plus quantitative, via quelques mesures d'efficacité, le potentiel des méthodes mises en jeu.

La figure 4 présente les résultats d'analyse de similarité obtenus par la distance euclidienne DE (figure 4.a), la distance de similarité DS* (figure 4.b), la distribution de Gibbs incluant E (figure 4.c) et la distribution de Dirichlet incorporant E (figure 4.d). Elle nous permet d'observer que le meilleur résultat de recherche d'images est donné par l'analyse basée sur l'intégration de la mesure de similarité E dans la distribution de Dirichlet. Nous allons donc vérifier si ce résultat reste valable en faisant des expérimentations basées sur les mesures de rappel et de précision.

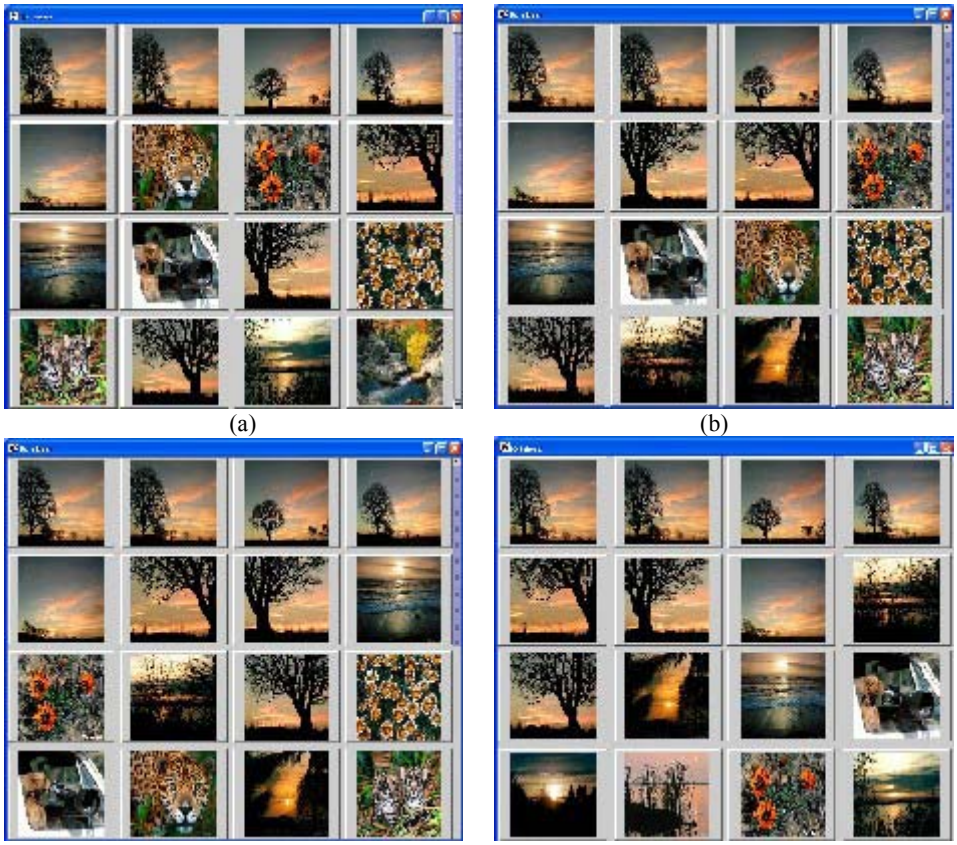

FIG. 4 - *Résultats de recherche d'images obtenus par : (a) la distance Euclidienne DE, (b) la distance de similarité DS*, (c) la distribution de Gibbs, (d) la distribution de Dirichlet.*

4 Expérimentations

Dans le but d'évaluer empiriquement la performance de notre système de recherche par le contenu, des mesures de rappel et précision (Raghavan et al 1989) sont calculées pour chacune des quatre variantes d'analyse de similarité. Une moyenne des valeurs pour l'ensemble des requêtes et l'ensemble des intervenants est ensuite déterminée.
Le rappel mesure la proportion des images pertinentes extraites par rapport au total des images pertinentes. La précision mesure quant à elle la proportion des images pertinentes extraites par rapport au total des images extraites. Pour pouvoir utiliser ces mesures, il faut d'abord identifier les images pertinentes pour chaque requête en déterminant des jugements de pertinence généralement fournis par des experts.

Comme indiqué auparavant, nous avons décomposé l'histogramme en 17 coupes de couleur ($1 \leq k \leq 17$), 17 coupes pour la luminance ($1 \leq m \leq 17$) et 12 coupes de chroma ($1 \leq n \leq 12$). Il est clair que le nombre de partitions pour chaque k, m et n a une forte influence sur le comportement des mesures de précision et de rappel ainsi que sur le temps de calcul.

4.1 Environnement d'expérimentation

Le système de recherche d'images par le contenu (Missaoui et al 2003) est implanté en Java sur un Pentium III de 900 MHz avec une mémoire RAM de 256 MB. L'expérimentation vise la détermination de la qualité de chacune des mesures de similarité sur une base de données de 2069 images composée d'une dizaine de catégories distinctes d'images extraites de sites Web. Dix requêtes (images) sont soumises à la base de données par quatre intervenants (étudiants et professeurs). Pour chacune des quatre mesures de similarité, le système produit un ordonnancement d'images cibles.

4.2 Analyse quantitative

La comparaison des quatre stratégies d'analyse de similarité est faite en fixant le rappel à un seuil donné variant entre 10% et 100% et en déterminant la précision moyenne (sur l'ensemble de requêtes et d' intervenants) pour chacun des niveaux de rappel.

Rappel (%)	Précision (%)			
	DE	DS*	Gibbs	Dirichlet
40	98.5	100	100	100
50	93	95	95.84	100
60	76.10	83.83	87.45	98
70	67.72	77.21	84.09	90.03
80	52.36	69.63	77.94	87.2
90	44.03	59.79	67.86	78.5
100	34.17	42.31	50.19	72.5

TAB. 2.

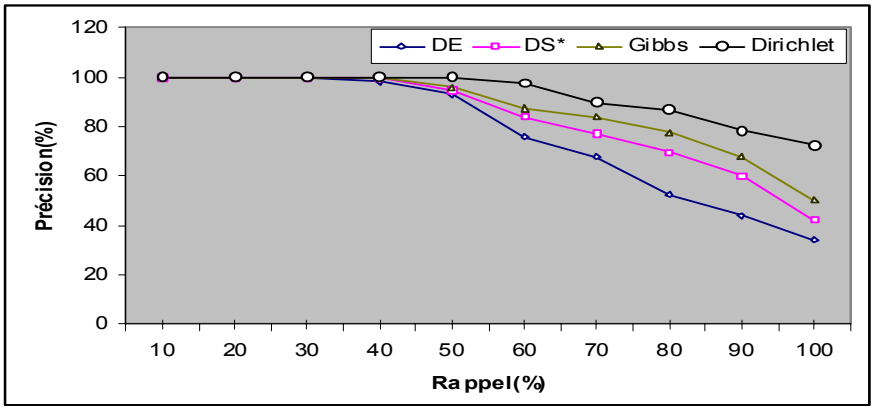

FIG. 5 - *Rappel et précision*

Les résultats de l'analyse montrent que pour un niveau de rappel déterminé dépassant 0.35, la courbe de précision est toujours la plus élevée dans le cas de la variante Dirichlet et la plus

faible dans le cas de la distance Euclidienne. Ainsi donc, la variante Dirichlet est meilleure que celle de Gibbs, laquelle est meilleure que la mesure de dissimilitude DS*, laquelle se comporte mieux que la distance Euclidienne. En outre, l'écart entre les précisions se creuse au fur et mesure que le rappel augmente. De même, la précision pour la variante Dirichlet reste au-dessus de 80% pour un rappel ≥ 80%. Cependant, le temps de calcul avec les variantes les plus performantes représente environ 2.5 fois celui de la distance Euclidienne.

5 Conclusion

Dans cet article, nous avons défini deux nouvelles mesures : la distance de dissimilitude DS*, et la distance de similarité E. La seconde est intégrée à la formule de la distribution de Gibbs (en utilisant seulement le pouvoir discriminant qu'apporte l'exponentiation des erreurs) et de celle de la mixture de Dirichlet généralisée, alors que la première est comparée à trois autres variantes d'estimation de la similarité : la distance Euclidienne et les distributions de Gibbs et Dirichlet intégrant la distance de similarité E. Les tests des quatre mesures de similarité porte sur les histogrammes de couleurs d'une collection d'images et montre que l'efficacité de la recherche mesurée par le rappel et la précision est la plus importante pour la distribution de Dirichlet modifiée et la plus faible pour la distance Euclidienne.

Remerciements

Nous tenons à remercier les évaluateurs anonymes pour leurs judicieuses remarques. La réalisation de cette recherche a été rendue possible grâce à la participation financière de VRQ, Canarie et Patrimoine Canada et des partenaires industriels du consortium CoRIMedia.

Références

Bouguila B., Ziou D. et Vaillancourt J. (2002), Maximum Likelihood Estimation of the Generalized Dirichlet Mixture, Rapport Technique 2002, Département de math & info, Université de Sherbrooke.

Faloutsos C. et al., (1994), Efficient and Effective Querying by Image Content, Journal of Intelligence Information System, Vol. 3, pp 231-262, 1994.

Missaoui R., Sarifuddin M., Hamouda Y., Vaillancourt J., et Laggoune H. (2003), A Framework for Image Mining and Retrieval, Visual Communications and Image Processing (VCIP2003), juillet 2002, Lugano, Suisse.

Rémillard B. et Beaudoin C., (1999), Statistical Comparison of Images Using Gibbs Random fields, Vision Interface'99, pp 612-617, 1999.

Raghavan, V.V., Jung G.S. et Bollmann B., (1989). A Critical Investigation of Recall and Precision as measures of Retrieval System Performance, ACM TOIS 7(3): pp 205-229.

Rui Y., Huang T.S. et Chang S-F. (1999), Image Retrieval: Current techniques, promising directions and open issues, Journal of Visual Communication and Image Representation, Vol. 10, pp. 39-62, March 1999.

Santini S. et Jain R. (1999), Similarity Measures, IEEE Trans. on PAMI, Vol.12, No. 9, pp. 871 – 883, September 1999.

Wand M.P. (1996), Data-Based Choice of Histogram Bin Width, Australian Graduate School of Management Working Paper Series, May 1996, University of New South Wales.