Nouvelle distance d'édition pour la reconnaissance d'images par chaînes de sacs de mots visuels

Hong-Thinh Nguyen, Cécile Barat, Christophe Ducottet

Nouvelle distance d'édition pour la reconnaissance d'images par chaînes de sacs de mots visuels

H.T. Nguyen[1] C. Barat[1] C. Ducottet[1]

[1] Université de Lyon, F-42023, Saint-Etienne, France,
CNRS, UMR5516, Laboratoire Hubert Curien, F-42000, Saint-Etienne, France,
Université de Saint-Etienne, Jean Monnet, F-42000, Saint-Etienne, France.

18 rue Benoit Lauras, 42000 Saint-Etienne, France
hong.thinh.nguyen@univ-st-etienne.fr

Résumé

La reconnaissance automatique d'images ou de classes d'images intervient dans de multiples problèmes, notamment lorsqu'il s'agit de classer ou de rechercher automatiquement des images par leur contenu. Depuis quelques années, la représentation d'images par sacs de mots visuels s'est imposée comme un modèle de référence pour ce type problème. L'un des inconvénients de ce modèle est de ne pas prendre en compte la répartition spatiale des mots dans les images. Dans cet article, nous proposons d'une part de représenter les images à partir de chaînes d'histogrammes de mots visuels issues d'un découpage régulier de l'image et d'autre part une nouvelle distance d'édition qui permet d'aligner deux chaînes en prenant en compte des opérations de fusions entre les histogrammes. Nous présentons l'algorithme de calcul de cette nouvelle distance et nous étudions ses performances pour la classification d'images.

Mots Clef

classification d'images, représentation d'image, sacs de mots visuels, distance d'édition, chaîne d'histogrammes.

Abstract

The two important problems of image classification based on visual content are the image content representation and the definition of a suitable metric to measure similarity between two images. Recently, the bag of visual words model has become the most popular representation image model for scene classification. However, an important drawbck of this model is that it does not take into account the spatial information of local features in the image. In this paper, we first propose a new image representation method based on strings of histograms of visual words. Second, we present a new edit distance operating on histogram strings and taking into account fusion of histograms. We study the performance of the method on an image classificaiton task.

Keywords

image classification, image representation, bag-of-visual-words, edit distance, string of histogram.

1 Introduction

La reconnaissance automatique d'images ou de classes d'images intervient dans de multiples problèmes, notamment lorsqu'il s'agit de classer ou de rechercher automatiquement des images par leur contenu. Depuis quelques années, la représentation d'images par sacs de mots visuels s'est imposée comme un modèle de référence pour ce type problème [12][5][17]. Elle consiste à représenter une image par un histogramme du nombre d'occurrences de mots visuels, obtenus par quantification d'un ensemble de descripteurs locaux.

L'un des inconvénients de ce modèle est de ne pas prendre en compte la répartition spatiale des mots dans les images. Pour pallier à cet inconvénient, l'approche la plus courante et la plus efficace est de découper l'image en différentes régions et de calculer un histogramme de mots visuels par région [5]. On obtient ainsi une représentation de type sacs de mots locaux. Dans ces représentations, après avoir ordonné les régions, on organise les différents histogrammes en un seul vecteur qui est utilisé comme représentation de l'image. La comparaison entre deux images se fait en comparant leurs vecteurs représentatifs.

Un inconvénient des approches sacs de mots locaux est qu'il faut fixer à l'avance le découpage de l'image. Ce découpage doit être appliqué de la même façon sur toutes les images sans tenir compte de leur contenu visuel. D'autre part, lors de la comparaison de deux images, une correspondance exacte entre les régions de ces deux images est recherchée là encore indépendamment de leur contenu. Par exemple, lorsque les images contiennent les mêmes objets mais à des positions différentes, la correspondance entre les régions ne sera pas assurée alors que le contenu visuel est similaire (cf. figure 1).

Dans cet article, nous proposons de définir des chaînes

de sacs de mots et de leur associer une nouvelle distance d'édition. Cette nouvelle distance permet d'aligner deux chaînes en prenant en compte des opérations de fusions entre les histogrammes composants ces chaînes. Ainsi, en fonction du contenu visuel, des régions peuvent être fusionnées, permettant ainsi d'une part de faire varier le nombre de régions d'une image à l'autre et/ou entre deux images et d'autre part de réaliser une meilleure mise en correspondance entre les sacs de mots.

Après avoir fait le lien avec les travaux antérieurs (section 2), nous expliquons dans la section 3 comment construire les chaînes de sacs de mots visuels. Dans la section 4 nous présentons alors la nouvelle distance d'édition associée à ces chaînes. Pour étudier les performances de cette distance, des expérimentations de classification d'images sont décrites et analysées dans la section 5. Enfin, la dernière section est dédiée à la conclusion et aux perspectives.

2 Travaux antérieurs

Un grand nombre de travaux récents s'intéressent à la prise en compte des informations spatiales dans les sacs de mots visuels. Essentiellement deux approches sont utilisées.

La première, déjà évoquée précédemment, consiste à partitionner l'image en différentes régions et à construire des histogrammes locaux dans chaque région. Les régions peuvent être issues d'une segmentation de l'image [3] ou d'un découpage régulier. La méthode qui constitue l'état de l'art dans ce domaine est le modèle de pyramide spatiale proposé par Lazebnik et al [5] qui repose sur un découpage régulier multi-échelles. Parmi les extensions de cette approche, citons les travaux de Zhou et al [20] qui utilisent des mélanges de gaussiennes (GMM) associés à un découpage hiérarchique pour éviter le problème de quantification des mots visuels.

La deuxième approche consiste à rechercher des relations spatiales entre mots visuels de types différents ou directement entre descripteurs, soit pour constituer des corrélogrammes ou histogrammes spatiaux [11, 6, 9], soit pour définir des chaînes de mots visuels [19, 18, 13, 7]. L'image est alors le plus souvent représentée par un histogramme d'occurrences de ces combinaisons de descripteurs en utilisant éventuellement un nouveau vocabulaire (par exemple les corrélatons dans [11]).

Dans les approches précédentes, l'étape de classification est effectuée à partir d'une fonction mesurant la distance ou la similarité entre les représentations issues des différentes images. La représentation d'une image étant la plupart du temps non structurée (vecteur ou histogramme), la fonction utilisée ne permet généralement pas d'échanges entre valeurs issues de composantes différentes. Pour pallier à cet inconvénient, quelques travaux mentionnent l'utilisation de distances de type EMD (Earth Movers Distance) [10] ou de type distances d'éditions [1, 7].

L'originalité de notre approche est de combiner une représentation par sac de mots locaux à une distance d'édition ce qui n'a, à notre connaissance, jamais été réalisé. Par rapport à la distance EMD, la distance d'édition présente l'avantage de prendre en compte la relation d'ordre entre les éléments de la chaîne tout en bénéficiant d'algorithmes de calcul relativement efficaces.

3 Représentation d'image par chaînes d'histogrammes

Notre modèle de représentation d'une image vise à prendre en compte les relations spatiales des descripteurs locaux des images et s'inspire de modèles de sacs de mots ordonnés [2]. A partir d'un découpage en régions d'une image, il s'agit de construire une ou plusieurs chaînes d'histogrammes, chaque histogramme correspondant au sac de mots visuel classique d'une région. Nous expliquons dans cette partie comment construire les chaînes, avant de décrire la représentation sac de mots utilisée.

3.1 Construction des chaînes

Dans les images, il existe souvent un axe privilégié selon lequel la projection des mots visuels de l'image permet de capturer l'information géométrique. Intuitivement, et comme mentionné dans [2], dans les scènes naturelles, les relations spatiales entre descripteurs locaux sont souvent verticales ou horizontales. Par exemple, dans le cas d'un coucher de soleil vu d'une plage, les descripteurs ciel se trouvent au dessus de ceux de la mer, qui eux mêmes se situent au dessus de ceux de la plage. Dans [13], pour les images d'objet, une méthode est proposée pour trouver un axe correspondant à l'orientation principale de l'objet, selon lequel les points d'intérêts sont projetés dans le même ordre indépendamment de la rotation ou de la translation de l'objet dans l'image.

Pour représenter une image comme un ensemble de chaînes, notre méthode consiste tout d'abord à diviser cette image en x bandes de largeur égale selon un axe priviligié défini. Chaque bande est alors subdivisée en n régions égales, orthogonalement à l'axe initial. Pour chaque bande, on forme ainsi une chaîne de n régions en la parcourant dans une direction donnée. Dans ce papier, nous nous limiterons à des divisions selon des axes verticaux et horizontaux. Toutefois, il est à noter que la méthode est généralisable à n'importe quelles orientations.

Dans le cas d'un nombre de bandes fixé à 1, on ne considère alors qu'une seule chaîne de régions ordonnées selon la direction de la bande (cf. figure 1). Dans les autres cas, nous aurons autant de chaînes que de bandes. Il nous reste à voir quel modèle sac de mots est associé à chacune des régions d'une chaîne.

3.2 Les sacs de mots visuels

Pour obtenir les sacs de mots visuels de chaque région, un vocabulaire visuel doit tout d'abord être défini. Le processus nécessite de choisir un ensemble de points d'intérêt, ainsi qu'un descripteur de ces points, puis d'appliquer un algorithme de clustering pour définir des mots visuels.

Dans ce travail, les points d'intérêt sont obtenus par dé-

coupage dense de l'image en imagettes carrées de côté c, centrées sur une grille régulière de taille t, autorisant le recouvrement entre points lorsque $t < c$. Ils sont décrits par le descripteur SIFT [8], de dimension d. L'algorithme de clustering K-means a été retenu pour obtenir k cluters de descripteurs, chaque centre de cluster représentant un mot visuel [4].
Chaque descripteur local d'une image est alors associé au mot du vocabulaire le plus proche au sens de la distance euclidienne. Le sac de mots visuels d'une région est l'histogramme du nombre d'occurrences des mots visuels qu'elle contient.
Pour comparer des images décrites au moyen de chaînes d'histogrammes, il est nécessaire de définir une métrique adaptée.

4 Nouvelle distance d'édition sur des chaînes d'histogrammes

L'idée sous-jacente à cette partie est d'utiliser une distance d'édition pour permettre un meilleur alignement de nos chaînes d'histogrammes. Nous rappelons d'abord la distance d'édition classique avant de proposer son adaptation à notre cas.

4.1 Distance d'édition classique

La distance d'édition classique ou distance de Levenshtein [14] permet de comparer des chaînes constituées de symboles issus d'un alphabet noté Σ. Elle définit trois opérations de base entre les symboles d'une chaîne : l'insertion d'un nouveau symbole, la suppression d'un symbole existant et la substitution d'un symbole en un autre. Il s'agit alors de calculer le nombre minimal d'opérations de base qui permettent de passer d'une chaîne d'entrée $\mathbf{x}(\mathbf{T})$ en une chaîne de sortie $\mathbf{y}(\mathbf{V})$. On peut aussi introduire une fonction de coût d'édition $c(x_r, y_k)$ qui pénalise une opération en tenant compte de la nature des symboles x_r et y_k la concernant. Si l'on note λ le symbole vide, cette fonction permet aussi de fixer les coûts de suppression et d'insertion d'un symbole x notés respectivement $c(x, \lambda)$ et $c(\lambda, x)$. Le calcul de la distance revient à la recherche de la séquence d'opérations de coût total minimal, ce qui est possible avec une complexité $\mathcal{O}(T \times V)$ à partir d'un algorithme de programmation dynamique qui se ramène au remplissage d'une matrice de taille $(T+1) \times (V+1)$ (cf. Algorithme 1). Notons que l'algorithme précédent n'est valable que si la fonction de coût vérifie l'inégalité triangulaire. De plus, si la fonction de coût est symétrique et définie positive, alors la distance est une métrique.

4.2 Nouvelle distance d'édition

La distance d'édition précédente peut s'appliquer à des chaînes d'histogrammes si l'on considère les histogrammes comme des symboles. Ainsi, l'ensemble des symboles Σ est l'ensemble des histogrammes possibles c'est à dire $\mathcal{R}^N$ si N désigne la taille des histogrammes (on considère que les valeurs de l'histogramme peuvent être réelles).

Entrées : Deux chaînes $\mathbf{x}(T)$ et $\mathbf{y}(V)$
Sorties : Distance d'édition $D(T, V)$ entre $\mathbf{x}(T)$ et $\mathbf{y}(V)$

$D(0,0) \leftarrow 0$;
pour *r=1 à T* **faire**
  $D(r,0) \leftarrow D(r-1,0) + c(x_r, \lambda)$;
fin
pour *k=1 à V* **faire**
  $D(0,k) \leftarrow D(0,k-1) + c(\lambda, y_k)$;
fin
pour *r=1 à T* **faire**
  pour *k=1 à V* **faire**
    $d_1 \leftarrow D(r-1,k) + c(x_r, \lambda)$; // Suppr.
    $d_2 \leftarrow D(r,k-1) + c(\lambda, y_k)$; // Inser.
    $d_3 \leftarrow D(r-1,k-1) + c(x_r, y_k)$; // Subst.
    $D(r,k) \leftarrow \min(d_1, d_2, d_3)$;
  fin
fin
retourner $D(T, V)$;

Algorithme 1: Distance d'édition classique retournant le coût du script minimal pour transformer la chaîne $\mathbf{x}(T)$ en $\mathbf{y}(V)$.

Si l'on ne considère pour l'instant que les opérations de substitutions (c'est à dire que l'on suppose que les coûts sont infinis pour les insertions et suppressions), une fonction de coût assez naturelle est la distance entre deux histogrammes soit $c(x, y) = d(x, y)$ si x et y désignent les deux histogrammes. On peut alors vérifier facilement que la distance d'édition est équivalente à la somme des distances entre les histogrammes issus des différentes régions. C'est ce type de distance qui est utilisé dans les méthodes de sacs de mots locaux.
La question des insertions et suppressions est plus délicate. Il convient de se poser la question du sens que l'on veut donner à de telles opérations, compte tenu de l'image d'origine dont les chaînes sont issues. Plutôt que de considérer que ces opérations conduisent à l'ajout ou à la suppression de mots visuels dans l'image d'origine, nous avons choisi d'introduire des opérations de fusion entre symboles c'est à dire entre les régions d'où sont issus les histogrammes. Une opération de fusion entre deux symboles se définit comme la somme des deux histogrammes correspondants, ce qui produit l'histogramme qui aurait été obtenu si les deux régions initiales avaient été fusionnées.
Ainsi, pendant le calcul de la distance d'édition entre deux chaînes $\mathbf{x}(\mathbf{T})$ et $\mathbf{y}(\mathbf{V})$, l'opération de suppression d'un symbole x_r est interprétée comme la fusion de ce symbole avec le symbole suivant dans la chaîne, soit x_{r+1}. Pour l'insertion, nous remarquons d'abord que l'insertion dans la chaîne d'entrée est équivalente à une suppression dans la chaîne de sortie. Ainsi, nous pouvons interpréter l'insertion d'un nouveau symbole y_k dans la chaîne d'entrée par la fusion du symbole courant y_k avec son suivant y_{k+1} dans la chaîne de sortie $\mathbf{y}(\mathbf{V})$. Nous proposons alors de fixer le coût d'une opération de fusion comme le coût de substi-

Entrées : Deux chaînes $\mathbf{x}(T)$ et $\mathbf{y}(V)$
Sorties : Distance d'édition $D(T,V)$ entre $\mathbf{x}(T)$ et $\mathbf{y}(V)$

1 $D(0,0) \leftarrow 0$;
2 $S_x(0,0) \leftarrow x_1$;
3 $S_y(0,0) \leftarrow y_1$;
4 **pour** *r=1 à T* **faire**
5 $\quad D(r,0) \leftarrow D(r-1,0) + c(S_x(r-1,0), x_{r+1})$;
6 $\quad S_x(r,0) \leftarrow \texttt{Fusion}(S_x(r-1,0), x_{r+1})$;
7 $\quad S_y(r,0) \leftarrow y_1$;
8 **fin**
9 **pour** *k=1 à V* **faire**
10 $\quad D(0,k) \leftarrow D(0,k-1) + c(S_y(0,k-1), y_{k+1})$;
11 $\quad S_x(0,k) \leftarrow x_1$;
12 $\quad S_y(0,k) \leftarrow \texttt{Fusion}(S_y(0,k-1), y_{k+1})$;
13 **fin**
14 **pour** *r=1 à T* **faire**
15 $\quad$ **pour** *k=1 à V* **faire**
16 $\quad\quad d_1 \leftarrow D(r-1,k) + c(S_x(r-1,k), x_{r+1})$;
17 $\quad\quad d_2 \leftarrow D(r,k-1) + c(S_y(r,k-1), y_{k+1})$;
18 $\quad\quad d_3 \leftarrow$
$\quad\quad D(r-1,k-1) + c(S_x(r-1,k-1), S_y(r-1,k-1))$;
19 $\quad\quad D(r,k) \leftarrow \min(d_1, d_2, d_3)$;
20 $\quad\quad$ **si** $d_1 \leq d_2$ *et* $d_1 < d_3$ **alors** // Suppr.
21 $\quad\quad\quad S_x(r,k) \leftarrow \texttt{Fusion}(S_x(r-1,k), x_{r+1})$;
22 $\quad\quad\quad S_y(r,k) \leftarrow S_y(r-1,k)$;
23 $\quad\quad$ **fin**
24 $\quad\quad$ **sinon si** $d_2 < d_3$ **alors** // Inser.
25 $\quad\quad\quad S_x(r,k) \leftarrow S_x(r,k-1)$;
26 $\quad\quad\quad S_y(r,k) \leftarrow \texttt{Fusion}(S_y(r,k-1), y_{k+1})$;
27 $\quad\quad$ **fin**
28 $\quad\quad$ **sinon** // Subst.
29 $\quad\quad\quad S_x(r,k) \leftarrow x_{r+1}$;
30 $\quad\quad\quad S_y(r,k) \leftarrow y_{r+1}$;
31 $\quad\quad$ **fin**
32 $\quad$ **fin**
33 **fin**
34 **retourner** $D(T,V)$;

Algorithme 2: Nouvelle distance d'édition prenant en compte des opérations de fusion entre les symboles

tution du symbole source vers le symbole avec lequel ce symbole est fusionné, c'est à dire qu'aux lignes 10 et 11 de l'algorithme 1, nous aurons :

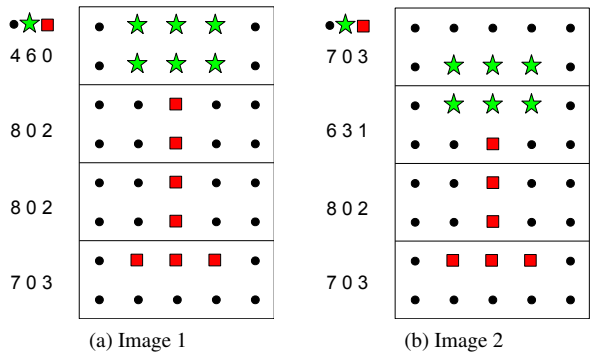
$$c(x_r, \lambda) = c(x_r, x_{r+1})$$
$$c(\lambda, y_k) = c(y_k, y_{k+1})$$

L'interprétation en terme de fusion implique également d'autres modifications dans l'algorithme initial. En effet, lors d'une opération de fusion, le symbole situé après le symbole fusionné est modifié. Cette modification a lieu soit dans la chaîne source (ligne 10), soit dans la chaîne destination (ligne 11) et uniquement si le coût résultant obtenu minimise la distance (ligne 13). Dans chacun de ces cas, il faut stocker le nouveau symbole pour pouvoir l'utiliser lors des opérations suivantes qui le concerneront. Notons qu'il ne faut pas modifier les chaînes initiales car d'autres scripts ne réalisant pas cette fusion seront envisagés lors de l'algorithme de programmation dynamique.

Ainsi, le symbole fusionné alors que l'on considère un couple de symboles (x_r, y_k) de l'algorithme ne sera utilisé que par les opérations futures impliquant la case (r,k) de la matrice D. Il est donc nécessaire de stocker deux nouvelles matrices S_x et S_y contenant respectivement, pour chaque position (r,k), le symbole suivant dans la chaîne $\mathbf{x}(\mathbf{T})$ et le symbole suivant dans la chaîne $\mathbf{y}(\mathbf{V})$. Ces deux matrices devront être mises à jour à chaque itération soit avec les symboles x_{r+1} et y_{k+1} (s'il y a une substitution), soit avec des symboles issus de fusions avec ces même symboles (s'il y a une fusion). L'algorithme correspondant est donné par l'algorithme 2.

Notons que les matrices S_x et S_y doivent être utilisées chaque fois que l'on considère les symboles issus d'une cellule précédemment calculée dans la matrice D. Notons également que nous avons fait le choix de privilégier les opérations de substitution en cas d'égalité avec une opération de fusion et de privilégier une fusion dans la chaîne source en cas d'égalité avec une fusion dans la chaîne destination (cf. conditions des lignes 20 et 24).

En guise d'exemple, la figure 2 présente les matrices résultat de cet algorithme pour la comparaison des deux chaines issues des images de la figure 1. Notons, que lors des calculs de distance, afin de pouvoir comparer correctement des régions de tailles différentes, il est nécessaire de prendre soin de renormaliser les histogrammes en fonction du nombre de régions fusionnées. Prenons l'exemple de l'insertion du symbole 7|3|0 dans la chaîne $\mathbf{x}(\mathbf{T})$. Cette insertion est traitée comme une fusion avec le symbole suivant dans la chaîne $\mathbf{y}(\mathbf{V})$, ici 6|3|1. La somme des deux histogrammes est alors divisée par 2 pour conduire à l'histogramme fusionné 6.5|3|0.5. A partir de ce nouveau symbole, une nouvelle insertion correspondra à une fusion avec le symbole suivant 8|0|2, cela consiste à reprendre la somme des 2 histogrammes précédents non normalisée,



(a) Image 1 (b) Image 2

FIGURE 1 – Exemple de découpage vertical avec deux images similaires

D S_x S_y	●★■ 0|0|0	●★■ 7|3|0	●★■ 6|3|1	●★■ 8|0|2	●★■ 7|0|3
●★■ 0|0|0	0 4|6|0 7|3|0	2 4|6|0 6.5|3|0.5	8 4|6|0 7|2|1	12 4|6|0 7|1.5|1.5	22 4|6|0 7|1.5|1.5
●★■ 4|6|0	12 6|3|1 7|3|0	6 8|0|2 6|3|1	8 8|0|2 8|0|2	10 8|0|2 7.5|0|2.5	20 8|0|2 7.5|0|2.5
●★■ 8|0|2	18 6.7|2|1.3 7|3|0	6 8|0|2 6|3|1	8 8|0|2 8|0|2	8 8|0|2 7|0|3	11 8|0|2 0|0|0
●★■ 8|0|2	22 6.8|1.5|1.8 7|3|0	8 7.7|0|2.3 6|3|1	10 7.7|0|2.3 8|0|2	8 7|0|3 7|0|3	10 7|0|3 0|0|0
●★■ 7|0|3	32 6.8|1.5|1.8 7|3|0	18 7.7|0|2.3 6|3|1	14 0|0|0 8|0|2	10.7 0|0|0 7|0|3	8 0|0|0 0|0|0

FIGURE 2 – Matrice résultat de l'algorithme de distance d'édition entre chaîne d'histogrammes appliqué aux chaînes de la figure 1 pour la distance L_1. Chaque case de la matrice contient la valeur de D et les deux histogrammes des matrices S_x et S_y. Le script d'édition minimal est représenté par des flèches. La distance obtenue est 8 alors que la la distance entre histogrammes locaux donnerait 12.

soit 13|6|1, et à ajouter l'histogramme 8|0|2, donnant ainsi l'histogramme 21|6|3. Ce dernier est alors normalisé par une division par 3, nombre de régions en jeu dans cette deuxième fusion, ce qui aboutit au symbole 7|2|1 de la matrice. En revanche, à partir du symbole 6.5|3|0.5, si on procède à une substitution, la distance entre les histogrammes 6.5|3|0.5 et 4|6|0 est calculée, donnant 6, valeur à laquelle il convient d'ajouter la distance précédemment calculée pour aboutir au symbole 6.5|3|0.5, soit 2.

5 Expérimentations

Nous évaluons notre approche sur une tâche de classification d'images avec la collection Simplicity [1][15]. Après avoir présenté le protocole expérimental et les paramètres utilisés, nous discuterons les résultats.

5.1 Protocole expérimental

Données. La collection Simplicity contient 1000 images extraites de la base d'images COREL, divisées en 10 catégories de 100 images chacune. Chaque image (384×256 pixels) ne peut appartenir qu'à l'une des 10 catégories : peuples d'Afrique, plages, buildings, bus, dinosaures, éléphants, fleurs, nourriture, chevaux et montagne (figure 3).

Classification et évaluation. La classification est effectuée selon la règle du 1-plus proche voisin. Le critère d'évaluation utilisé est le taux de bien classés. Ce critère est estimé par une procédure de validation croisée stratifiée à 5 niveaux. Le classifieur est entraîné avec 80 images de chaque classe, et testé avec les 20 images restantes de chaque classe. Ainsi, 800 images sont utilisées pour l'apprentissage et 200 pour le test.

Représentation sac de mots et chaînes. Pour l'échantillonnage dense, la constitution du vocabulaire visuel et la représentation d'image sous forme de chaînes d'histogrammes présentés Section 3, nous avons choisi les paramètres suivants :

1. http://wang.ist.psu.edu/~jwang/test1.zip

c =	16	dimension du point d'intérêt
t =	12,14	paramètre de recouvrement
d =	128	dimension du descripteur SIFT
k =	200, 400	taille du vocabulaire
x =	1	nombre de bande par image
n =	1 à 12	nombre d'histogrammes par chaîne

De nombreux paramètres interviennent en général dans les représentations sacs-de-mots. Dans le cadre de ce papier, nous ne pouvons étudier l'influence de chacun de ces paramètres sur le résultat de classification et avons donc restreint notre étude à certains. Notamment, nous avons choisi de travailler avec deux tailles de vocabulaires, identiques à celles choisies dans [5]. Des résultats obtenus $t = 12$ et $t = 14$ seront comparés afin d'étudier l'influence du nombre de points d'intérêts retenus par régions. Une seule chaîne d'histogrammes par image sera considérée, orientée verticalement. L'impact de la longueur de cette chaîne sera analysé en variant le paramètre n entre 1 et 12.

Fonction de coût. Deux distances d'histogrammes seront considérées pour la fonction de coût : la distance L_1 et la distance du χ_2. Notre approche sera notée $DECH_{L_1}$ ou $DECH_{\chi_2}$ selon la métrique utilisée (Distance d'Edition entre Chaînes d'Histogrammes).

Comparaison avec d'autres approches. Afin de montrer l'intérêt de notre méthode, nous proposons de comparer les résultats avec une approche sacs de mots locaux, notée DHL (Distance entre Histogrammes Locaux). Dans cette approche, une image est représentée au moyen d'un seul histogramme obtenu par concaténation des histogrammes de chaque région. La distance entre deux images est alors calculée au moyen de celle utilisée comme fonction de coût dans la méthode $DECH$. Ainsi, on distinguera les méthodes DHL_{L_1} et $DECH_{\chi_2}$, à comparer respectivement à $DECH_{L_1}$ et $DECH_{\chi_2}$. Le test statistique non paramétrique de la somme des rangs signés de Wilcoxon sera utilisé pour indiquer si une méthode est significativement meilleure qu'une autre [16].

FIGURE 3 – Exemple d'images de la collection SIMPLIcity.

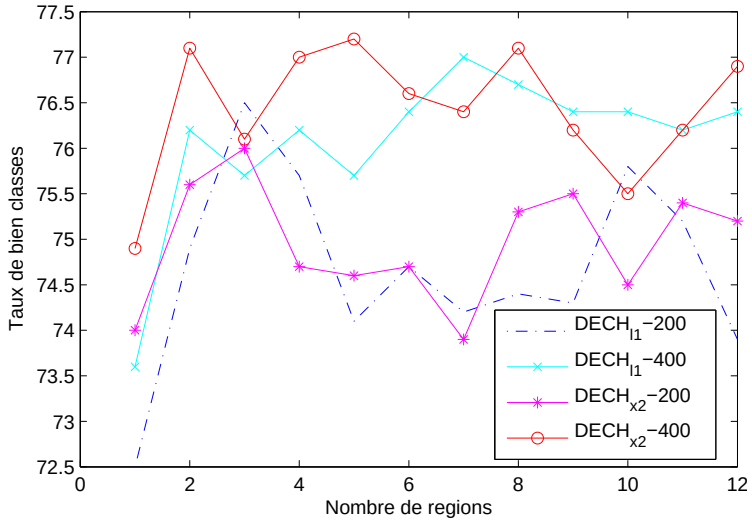


FIGURE 4 – Taux de bien classés en fonction du nombre de régions obtenus par $DECH$ calculée avec $t = 14$, deux tailles de vocabulaires 200 et 400, deux fonctions de coût.

5.2 Résultats

Influence de la taille du vocabulaire et de la fonction de coût. La figure 4 présente 4 graphes de résultats des méthodes $DECH_{L_1}$ et $DECH_{\chi_2}$ pour les deux tailles de vocabulaires 200 et 400, le paramètre t étant fixé à 14. Globalement, les résultats obtenus avec $k = 400$ sont supérieurs à ceux obtenus avec $k = 200$. Ces écarts sont statistiquement significatifs selon le test de Wilcoxon (p<0,01 dans les 2 cas). Par contre, pour une même taille de vocabulaire, l'utilisation de l'une ou l'autre des fonctions de coût n'influe pas. Les écarts ne sont pas considérés comme statistiquement significatifs. Dans la suite des résultats, nous ne montrerons que les résultats obtenus avec L_1.
Le premier point de chacune de ces courbes correspond au cas où aucun découpage n'est appliqué à l'image, soit à la représentation sac de mots visuels standard, prise comme référence. Quelle que soit la taille du vocabulaire et la fonction de coût, on constate que les taux de classification augmentent à partir du moment où l'on travaille sur des chaînes d'histogrammes composées d'au moins deux éléments. Pour un vocabulaire de 200, les courbes atteignent un maximum vers la valeur 3, puis décroissent pour atteindre un pallier dont la valeur moyenne reste au dessus de la valeur de référence. Pour un vocabulaire de 400, le comportement est proche. Les courbes atteignent un maximum, puis décroissent. La décroissance est observée plus tardivement, pour un découpage en nombre de régions supérieur à 10. Ces résultats montrent qu'il est clairement intéressant d'utiliser $DECH$ avec une représentation des images sous forme d'une chaîne d'histogrammes composées de 3 éléments.

Influence du nombre de points (paramètre t). Au vu des conclusions précédentes, nous présentons des résultats calculés avec L_1, un vocabulaire de 400 et les deux valeurs de t, 12 et 14. A l'exception des points d'abscisse 2 et 3, les résultats obtenus avec $t = 14$ sont supérieurs à ceux obtenus pour $t = 12$. Les écarts sont considérés comme significatifs. On en déduit que plus le nombre de descripteurs locaux est important, plus la performance de classification s'améliore.

Comparaison entre méthodes $DECH$ et DHL. La table 1 et la figure 6 montrent les résultats des approches $DECH_{L_1}$ et DHL_{L_1} lorsque l'on augmente le nombre de régions dans la chaîne, pour un vocabulaire de taille

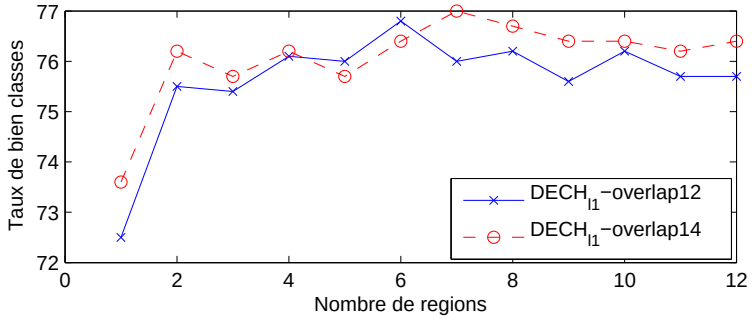


FIGURE 5 – Taux de bien classés en fonction du nombre de régions obtenus par $DECH_{L_1}$ pour $t = 12$ ou $t = 14$, $k = 400$

$k = 400$ et $t = 14$. Pour un découpage inférieur ou égal à 3 régions, les deux courbes se superposent. Au delà de 3, la courbe $DECH_{L_1}$ atteint un maximum pour 7 régions, puis tend à se stabiliser autour de la valeur 76.5, alors que les résultats obtenus par DHL_{L_1} décroissent. Le test de Wilcoxon donne une p-valeur inférieure à 0,01 indiquant la supériorité de la méthode $DECH_{L_1}$ sur DHL_{L_1}. Notons que cette stabilisation peut s'expliquer par les opérations de fusion qui sont permises par la méthode $DECH$. Au delà de 3 ou 4, le nombre de régions optimal est déterminé automatiquement par la distance.

TABLE 1 – Taux de bien classés en fonction du nombre de régions obtenus par $DECH_{L_1}$ et DHL_{L_1} pour $t = 14$, $k = 400$. Moyenne (μ) et écart-type (σ) calculés sur 5 niveaux de validation croisée stratifiée.

Nombres de régions	$DECH_{L_1}$		DHL_{L_1}	
	μ	σ	μ	σ
1	73,6	3,9	73,6	3,9
2	76,2	3,1	76,2	3,1
3	75,7	3,3	75,7	3,3
4	76,2	3,2	76,1	3,2
5	75,7	2,8	75,3	3,2
6	76,4	3,2	75,9	3,7
7	**77**	3,3	**76,2**	3,6
8	76,7	3,4	74,5	3,4
9	76,4	2,9	74,7	3,3
10	76,4	2,8	75,4	4,5
11	76,2	2,9	75,2	5,2
12	76,4	3,2	74,8	3,6

6 Conclusion

Dans cet article, nous avons présenté une nouvelle distance d'édition permettant de comparer des images représentées sous la forme chaînes de sacs de mots. Etant donnée une direction de référence les chaînes sont construites en découpant les images en un nombre fixé de régions perpendiculairement à cette direction. Chaque symbole de la chaine correspond à l'histogramme des mots visuels de la région d'origine. Par rapport aux approches sac de mots locales, la nouvelle distance d'édition permet de réaliser des opérations de fusion entre les régions pour faciliter la mise en correspondance entre deux chaînes. Nous avons évalué les performances de cette approche sur une tâche de classification d'image. Après avoir étudié l'influence de la taille du vocabulaire et du nombre de mots visuels par image, nous avons montré que la nouvelle distance produit un taux d'images de bien classées toujours supérieur à la distance classique entre les histogrammes locaux.

En perspective, nous envisageons d'étudier le comportement de cette distance sur d'autre bases d'images. Il serait aussi intéressant d'évaluer d'autres méthodes de construction des chaînes en faisant varier la direction de référence et le nombre de bandes comme proposé dans la section 3. Notons également que cette distance pourrait être utilisée dans un noyau d'édition pour pouvoir bénéficier de méthodes de classifications telles que les SVM. Enfin, pour mieux exploiter les propriétés spatiales des images, il serait intéressant de généraliser cette distance sur d'autre structures de données comme les arbres pour lesquels il existe déjà des distances d'édition.

Références

[1] C. Barat, C. Ducottet, E. Fromont, AC. Legrand, and M. Sebban. Weighted Symbols-based Edit Distance for String-Structured Image Classification. In *Proceedings of ECML PKDD 2010*, volume 6321 of *Lecture Notes in Computer Science*, pages 72–86. Springer, 2010.

[2] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial-bag-of-features. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3352–3359, 2010.

[3] X. Chen, X. Hu, and X. Shen. Spatial weighting for bag-of-visual-words and its application in content-based image retrieval. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '09, pages 867–874, Berlin, Heidelberg, 2009. Springer-Verlag.

[4] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu. An efficient k-means clustering algorithm : Analysis and implementation.

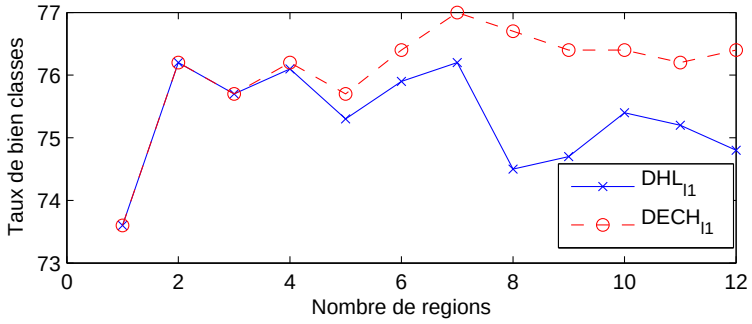


FIGURE 6 – Taux de bien classés en fonction du nombre de régions obtenus par $DECH_{L_1}$ et DHL_{L_1} pour $t = 14$, $k = 400$

IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7) :881–892, 2002.

[5] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories. In *CVPR 2006, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178. IEEE, 2006.

[6] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR 2008, IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[7] Y. Liu and V. Caselles. Spatial string matching for image classification. In *ICPR 2010, 20th International Conference on Pattern Recognition*, pages 2937–2940. IEEE, 2010.

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

[9] N. Morioka and S. Satoh. Building compact local pairwise codebook with joint feature space clustering. In *ECCV 2010, European Conference on Computer Vision*, pages 692–705. Springer, 2010.

[10] Y. Rubner, C. Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2) :99–121, 2000.

[11] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *CVPR 2006, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2033–2040. IEEE, 2006.

[12] J. Sivic and A. Zisserman. Video Google : A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003.

[13] P. Tirilly, V. Claveau, and P. Gros. Language modeling for bag-of-visual words image categorization. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, CIVR '08, pages 249–258, New York, NY, USA, 2008. ACM.

[14] Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1) :168–173, 1974.

[15] J.Z. Wang, J. Li, and G. Wiederhold. Simplicity : Semantics-sensitive integrated matching for picture libraries. In *VISUAL '00 : Proceedings of the 4th International Conference on Advances in Visual Information Systems*, pages 360–371, London, UK, 2000. Springer-Verlag.

[16] F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6) :80–83, 1945.

[17] J. Yang, YG. Jiang, A. G. Hauptmann, and CW. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *MIR '07 : Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206, New York, NY, USA, 2007. ACM.

[18] J. Yuan, Y. Wu, and M. Yang. Discovery of Collocation Patterns : from Visual Words to Visual Phrases. In *CVPR '07. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[19] QF. Zheng, WQ. Wang, and W. Gao. Effective and efficient object-based image retrieval using visual phrases. In *Proceedings of the 14th annual ACM international conference on Multimedia*, MULTIMEDIA '06, pages 77–80, New York, NY, USA, 2006. ACM.

[20] X. Zhou, N. Cui, Z. Li, F. Liang, and T.S. Huang. Hierarchical gaussianization for image classification. In *ICCV 2009, IEEE 12th International Conference on Computer Vision*, pages 1971–1977. IEEE, 2009.