Représentation par graphe de mots manuscrits dans les images pour la recherche par similarité

Peng Wang[1,2], Véronique Eglin[1], Christophe Garcia[1]
[1]Université de Lyon, LIRIS UMR5205, INSA-Lyon, France
{peng.wang, veronique.eglin, christophe.garcia}@liris.cnrs.fr

Christine Largeron[2]
[2]LAHC, CNRS UMR5516
Université Jean Monnet, Saint Etienne, France,
christine.largeron@univ-st-etienne.fr

Josep Lladós, Alicia Fornés
Computer Vision Center, Universitat Autònoma de Barcelona, Espagne
{josep, afornes}@cvc.uab.es

RÉSUMÉ. *Dans ce papier, nous proposons une nouvelle approche de la recherche de mots par similarité reposant sur une structure de graphes intégrant des informations sur la topologie, la morphologie locale des mots ainsi que des informations contextuelles dans le voisinage de chaque point d'intérêt. Chaque mot image est représenté par une séquence de graphes associés chacun à un objet connexe. Un graphe est construit sur la base d'un squelette décrit par un descripteur riche et compact en chaque point sommet: le contexte de formes. Afin d'être robuste aux distorsions de l'écriture et aux changements de scripteurs, l'appariement entre mots repose sur une distance dynamique et un usage adapté du coût d'édition approximé entre graphes. Les expérimentations sont réalisées sur la base de George Washington et la base de registres de mariages célébrés en la cathédrale de Barcelone. L'analyse de performances montre la pertinence de l'approche comparativement aux approches structurelles actuelles.*
MOTS-CLÉS : *recherche de mots par similarité, représentation par graphes, contexte de formes, distance d'édition, DTW, fusion d'information, interrogation par l'exemple*.

ABSTRACT. *Effective information retrieval on handwritten document images has always been a challenging task. In this paper, we propose a novel handwritten word spotting approach based on graph representation. The presented model comprises both topological and morphological signatures of handwriting. Skeleton-based graphs with the Shape Context labeled vertexes are established for connected components. Each word image is represented as a sequence of graphs. In order to be robust to the handwriting variations, an exhaustive merging process based on DTW alignment results introduced in the similarity measure between word images. With respect to the computation complexity, an approximate graph edit distance approach using bipartite matching is employed for graph matching. The experiments on the George Washington dataset and the marriage records from the Barcelona Cathedral dataset demonstrate that the proposed approach outperforms the state-of-the-art structural methods.*
KEYWORDS: word spotting, graph-based representation, shape context description, graph edit distance, DTW, block merging, query by example.

1. Introduction

Les applications fondées sur l'exploitation des manuscrits reposent pour une grande partie sur des versions numérisées. L'accès au contenu nécessite alors de pouvoir considérer les textes à travers une représentation qui doit être concise, discriminante et informante. Ce sont sur ces représentations que sont fondées les techniques de recherche par le contenu (recherche par similarités de formes, *word retrieval*, *word spotting*…). De nombreux défis scientifiques sont ainsi associés à la recherche par le contenu dans les images de traits. Autour de la recherche par similarité de formes écrites, un élément central concerne l'étude de la variabilité interne des écritures ainsi que celle qui permet de distinguer deux écritures de mains différentes. Il est désormais communément admis que les techniques d'OCR sont totalement inopérantes sur la plupart des textes écrits, en particulier sur les supports anciens et historiques souvent très dégradés et présentant des particularités graphiques associée à une grande diversité de styles d'écriture.

Généralement on associe le *word spotting* à deux types d'approches : une première famille repose sur la considération de mots prédéfinis en lien avec des mécanismes d'apprentissage spécifique dédiés à ces mots. On peut citer les travaux décrits par Fischer et Rodriguez dans [1] et [2] par exemple reposant sur des modèles de Markov cachés et des vocabulaires spécifiques. A côté de ces approches, on généralise les techniques de *word spotting* par le développement de techniques de représentation et de processus d'appariement flexible mettant les mots requêtes en correspondance avec les cibles issues de l'image [3], [4]. Dans les deux cas, l'accès au contenu dans l'image nécessite de disposer de représentations complètes assurant à la fois une bonne rigueur de description et une souplesse nécessaire pour absorber les variations présentes dans les pages d'écritures [1], [2], [3]. Finalement c'est sur la mesure de similarité permettant de déterminer les appariements acceptables qu'une attention particulière doit être portée. Cette mesure doit également offrir le maximum de robustesse aux déformations, aux changements d'échelles et irrégularités dans la formation des traits et enfin aux dégradations entamant généralement la description des contours et des extrémités de formes.

Les représentations structurelles des formes fondées sur les graphes sont très populaires pour leur reconnaissance car elles permettent de modéliser très finement les dimensions structurelles et topologiques des objets. Dans certains domaines de l'imagerie graphique, comme en reconnaissance de symboles et de formules chimiques, ces techniques ont trouvé un essor considérable ces dernières années [20]. Cependant, on constate que les représentations graphiques demeurent très insuffisamment exploitées dans le domaine de l'écrit ou bien souvent ce sont des mécanismes moins rigides qui sont employés car ils assurent une meilleure tolérance aux variations internes de l'écriture. On peut cependant citer quelques tentatives réalisées dans ce domaine et qui auront abouti à des premiers résultats prometteurs pour la reconnaissance du chinois et des textes à la dimension structurelle particulièrement prégnante [5], [6]. Bien que le modèle hiérarchique proposé dans [5] reflète la nature complexe des caractères écrits chinois, les mécanismes de représentation associés aux textes écrits sont généralement plus sophistiqués pour permettre de distinguer des formes similaires. On peut citer ici les

travaux de Fischer et al. dans [7] et [8] qui ont développé un modèle de représentation de graphique basé sur le squelette de l'image des mots à partir de l'encodage des sommets. En ajoutant une quantité suffisante de points d'intersection sélectionnés parmi les points d'intérêt de l'image, les informations structurelles demeurent préservées et peuvent prévenir des nombreuses insuffisances liées à l'usage exclusif d'une représentation fondée sur un squelette potentiellement bruité ou incomplet. Le risque alors est de disposer d'un trop grand nombre de sommets et de complexifier les calculs : les mécanismes fondés sur les graphes présentent par définition une complexité calculatoire exceptionnelle. En effet, le coût des correspondances entre formes graphiques, les transformations des représentations graphiques en leur équivalent vectoriel (vecteurs de caractéristique) rappellent ces considérations calculatoires importantes [9]. Dans [9], les auteurs présentent la construction d'un graphe direct acyclique pour chaque sous-mot présent dans le texte converti ensuite en un vecteur exprimant la signature topologique des mots. Lladós et al. dans [10] ont adapté le concept de sérialisation de graphes pour des applications d'appariement de graphes appliqués au word spotting. En extrayant et en classifiant les chemins acycliques du graphe en une représentation unidimensionnelle, les auteurs ont pu créer un mécanisme de description des mots fondé sur des « sacs de chemins » (Bag of Paths, *BoP*). Les performances restent encore faibles pour des applications de word spotting où les irrégularités, et les imprécisions demeurent omniprésentes.

Notre motivation à exploiter une représentation structurelle reposant sur une construction de graphe (plutôt que de partir d'un modèle d'apparence associé à une description en points d'intérêt par exemple ou à un ensemble d'indicateurs de formes, tels que les histogrammes de projection de points de contours, de courbures ou d'orientations) est lié à la stabilité structurelle des mots. Il existe en effet des règles d'exécution reproductibles que les points de jonction, de bifurcation et extrêmaux présents sur les traits d'écriture permettent d'encoder [19]. Les caractéristiques topologiques des traits et la nature bidimensionnelle de l'écriture nous semblent être des indications suffisantes pour écarter les descriptions unidimensionnelles reposant sur des valeurs scalaires uniquement [9].

Dans ce papier nous proposons une approche générique de *word spotting* ne nécessitant aucun paramétrage amont, n'ayant pas recours à l'apprentissage et reposant sur une représentation des mots par graphe. La description du graphe est fondée sur des primitives morphologiques obtenues par le descripteur contextuel de *Contexte de formes* (*SC* : Shape Context descriptor [11]). Cette description est intégrée au modèle de représentation pour indiquer localement en chaque sommet du graphe la nature de son voisinage et les relations de proximité que les contours ont entre eux. Ce descripteur est estimé sur la longueur totale d'un mot en tout point du graphe. Chaque mot est finalement représenté par une séquence de graphes formés à partir des connexités initialement repérés lors d'une étape de prétraitement. La comparaison entre les mots image (*requête* et *cibles*) est finalement obtenue par le calcul de la moyenne des distances d'édition entre graphes pair à pair. Au préalable, une mesure dynamique (Dynamic Time Warping) est exploitée entre les graphes *Requête* et *Cibles* comme processus de fusion de connexités garantissant les meilleures correspondances et les meilleurs appariements de graphes. Une distance

d'édition approximée initialement définie dans [12] a été choisie dans nos travaux afin de gagner en temps de calcul et donc en rapidité d'exécution. Les performances de la description comparées à d'autres approches sont étudiées pour des applications d'interrogations par l'exemple.

La suite du papier est organisée de la façon suivante: après une brève introduction portant sur les étapes de prétraitement nécessaire à l'obtention d'un squelette, des contours et des points structurels exploitables, nous présentons formellement le concept de graphe pour la représentation des formes écrites (section III). Nous développons ensuite les métriques retenues pour assurer une comparaison de graphes robuste aux variations internes d'écriture : la comparaison des graphes requêtes et cibles par l'application de la DTW permet de produire les meilleures configurations internes (fusion de graphes) qui sont ensuite évaluées par la distance d'édition approximée entre graphes (section IV). Les résultats expérimentaux et l'analyse des performances reposent sur la comparaison entre quatre approches de *word spotting* du domaine avec notre proposition (section V). Nous conclusions le papier sur de possibles améliorations pour des applications de *word spotting* sans segmentation.

2. Prétraitements

Afin de produire un système comparable à l'existant, nous sommes partis de l'a priori de l'existence d'une segmentation en mots. Nous disposons donc pour ce travail de mots *Requêtes* bien identifiés dans le benchmark de l'étude (et qui sera détaillé en section V) et de mots *Cibles* que nous avons extraits pour les besoins de l'étude. En réalité, cette étape est purement artificielle car le mécanisme complet de recherche de mots ne peut se passer d'une approche bas niveau de localisation de régions d'intérêt. Cette contribution n'est pas présentée dans ce papier. Pour obtenir les informations topologiques de l'écriture et un squelette unitaire constituant l'entrée du mécanisme de sélection de points structurels, nous avons appliqué l'algorithme de squelettisation de Zhang et Suen défini dans [13]. Cet algorithme possède une version parallélisable applicable sur des images binaires, voir Figure 1.c. Cette représentation permet d'extraire aisément les points de structure classifiés en trois familles : des points de forte courbure, des points de croisement, des points extrémaux. Ces points sont très génériques et totalement indépendants des types de langue et d'écriture considérés. La méthode d'extraction et de classification des points est présentée dans [14]. La Figure 1.d illustre cette décomposition en produisant une version segmentée de l'écriture en segments (ligne et courbes concaves ou convexes) séparés par des points d'accroche.

Afin de caractériser les informations morphologiques des mots, nous avons choisi d'utiliser le contour extérieur de l'écriture. Après une étape de comblement des trous et des zones de discontinuités dans le tracé, le détecteur de contour de *Canny-Deriche* est utilisé pour extraire des bords [21]. Une procédure de suivi de contour est ensuite appliquée pour assurer une extraction continue sans rupture de tracé (voir la Figure 1 (b)).

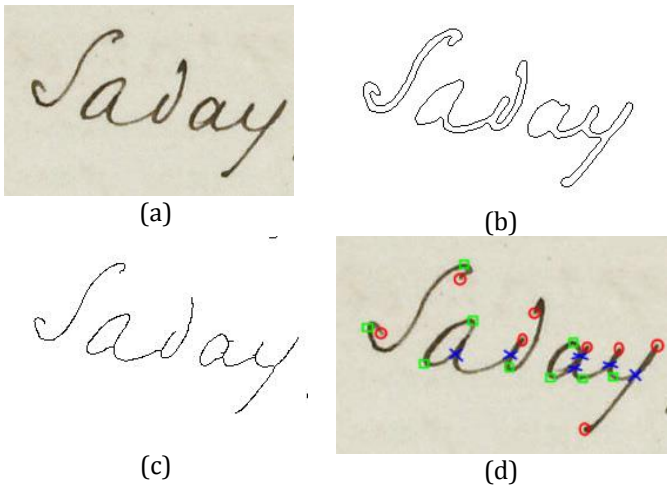


Figure 1. (a) mot manuscrit (b) son contour (c) son squelette (d) points structurels (cercle : points extrémaux, carré : points de haute courbure, croix : point de croisement)

3. Modèle de graphe pour la représentation structurelle des mots

Les représentations reposant sur les modèles de graphes possèdent l'avantage de préserver les propriétés topologiques et dispositionnelles des segments internes présents dans l'écriture. Nous avons choisi de représenter les propriétés structurelles de l'écriture en nous intéressant à son squelette qui contient à lui seul toutes les indications morphologiques que nous souhaitons retenir. Les sommets du graphe correspondent aux points structurels extraits par l'analyse des configurations locales du squelette et les arêtes (voir Figure 3). L'information morphologique contextuelle de chaque sommet en relation avec son proche voisinage est décrite par le *contexte de formes* (*SC* : Shape Context descriptor [11]), qui capture la distribution de contours du voisinage d'un point en un vecteur riche et informant. Ce descripteur a été exploité dans des contextes similaires pour la reconnaissance de formes graphiques [18], voir Figure 2. Il est associé ici à chaque sommet du graphe. Les arêtes du graphe sont également décrites par la longueur des segments estimée à partir du décompte de points du squelette entre deux sommets adjacents.

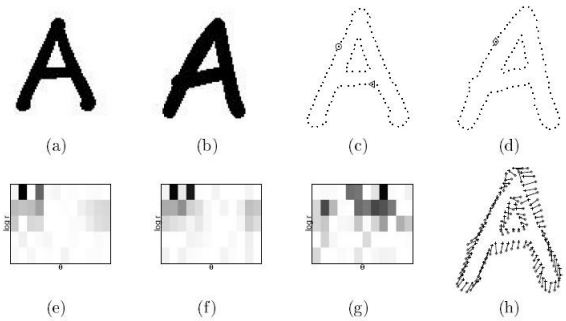


Figure 2. Illustration du descripteur de Contexte de Formes (SC) [11].

Chaque composante connexe génère ainsi un graphe propre. Un mot est donc constitué d'un ensemble de graphes définis individuellement et à ce stade ne possédant aucune connexion entre eux. La Figure 3 illustre une représentation en séquence de graphes du mot "*Savay*" : les trois graphes sont réellement constitués des portions de mots "*S*", "*av*" et "*ay*".

La définition du graphe peut ainsi se présenter de la façon suivante:

Définition 1. (Graph) Un graphe est un 4-tuple $g = (V, E, \mu, v)$, où

- V représente l'ensemble des sommets correspondant aux points structurels
- $E \subseteq V \times V$ est l'ensemble des arêtes, correspondant aux segments entre deux sommets adjacents
- $\mu: V \rightarrow L$ est la fonction qui associe à chaque sommet son vecteur de contexte de formes (SC : 5 niveaux concentriques et 12 secteurs angulaires, soit 60 dimensions)
- $v: E \rightarrow L'$ est la fonction qui associe à chaque arête du graphe sa longueur estimée entre deux sommets.

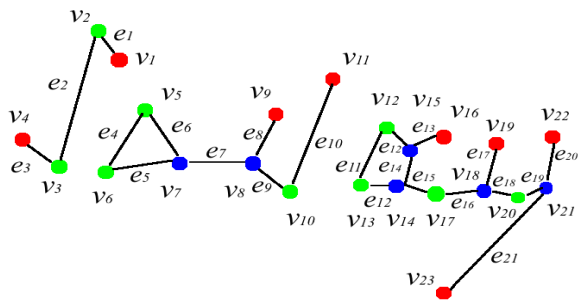


Figure 3. Représentation en une séquence de trois graphes du mot "*Savay*".

4. Appariements entre mots et mesures de similarité

Au-delà des métriques reposant sur une comparaison exhaustive des composants du graphe et interdisant toute distorsion ou souplesse dans l'alignement des arêtes et des nœuds à comparer, nous avons choisi d'exploiter des mesures de similarités non linéaires permettant :

- De comparer chaque paire de composantes connexes d'un mot à l'autre à l'aide d'une distance d'édition approximée (définie en section 4.1)
- D'estimer les similarités internes interprétées comme des critères de fusion, simplifiant ainsi le modèle de représentation des mots cibles en diminuant leur nombre de connexités (section 4.2)
- D'estimer la distance finale entre deux mots par le cumul des distances internes

L'approche structurelle sur laquelle nous fondons la comparaison de mots souligne tout l'intérêt de la prise en compte de la structure des entités dans la métrique de comparaison, complexifiant inévitablement la vision qui ramène le problème à un espace mathématique où les objets sont comparables linéairement selon un calcul de distance entre vecteurs de caractéristiques. L'idée de procéder à une approche relevant de l'étude de l'alignement structurel entre deux graphes revient donc à prendre en compte non seulement les correspondances entre sommets mais aussi les ressemblances dans leurs connexions (arêtes).

4.1 *Correspondance entre graphes et distance d'édition approximée*

Puisque les composantes connexes sont représentées par des graphes, leur comparaison se ramène à un problème de correspondance de graphes. Afin d'éviter les débordements calculatoires vite atteints dans de telles situations, nous avons opté pour l'exploitation d'une comparaison approximée entre graphes proposée initialement par K. Riesen and H. Bunke dans [12] reposant sur la recherche du coût d'édition minimal entre deux graphes.

Définition 2. (Distance d'édition entre graphes) Soit $g_1 = (V_1, E_1, \mu_1, v_1)$ le graphe associé à la représentation de l'image *Requête* et $g_2 = (V_2, E_2, \mu_2, v_2)$ le graphe associé à l'image *Cible*. La distance d'édition entre les deux graphes g_1 et $g_2$est définie par :

$$d(g_1, g_2) = \min_{(e_1,\ldots,e_k)\in\gamma(g_1,g_2)} \sum_{i=1}^{k} c(e_i) \qquad (1)$$

où $\gamma(g_1, g_2)$ représente l'ensemble des chemins possibles permettant la transformation du graphe g_1 en graphe g_2. La fonction $c(e_i)$ représente le coût cumulant les coûts intermédiaires nécessaires pour transformer le graphe g_1 en g_2. Les opérations indépendantes de coût e_i peuvent être : des opérations d'insertion, de suppression ou de substitution de sommets et d'arêtes.

L'algorithme de comparaison sous-optimal est basé sur la considération des graphes internes à chaque connexité (une connexité étant, rappelons-le constituée de

n sommets et *m* arêtes). Les graphes internes sont donc ici définis comme un ensemble centré en un sommet et sa structure adjacente locale. Par conséquent, la distance d'édition entre deux graphes peut être reformulée comme la recherche d'un optimal entre sommets et leurs structures locales respectives. Au lieu d'utiliser les ressources coûteuses de la programmation dynamique, l'appariement bipartite entre deux graphes est adapté et rapide. L'appariement de graphes est donc traité comme un problème d'affectation. Pour cela l'algorithme *hongrois* de Munkres (*Hungarian matrix*) a été choisi pour résoudre ce problème de recherche d'optimal entre graphes minimisant le coût lié aux transformations de sommets et d'arêtes des graphes considérés.

Afin d'estimer les coûts de substitution d'un sommet en un autre, deux composantes ont été mesurées : le premier indice permet de comparer la description locale reposant sur le contexte de formes entre deux sommets et le second repose sur la comparaison de la structure locale entre deux graphes impliquant les longueurs de segments séparant deux sommets (voir Equation 2).

Nous avons ainsi testé les performances de la distance d'édition entre graphes avec différentes valeurs de pondérations de ces coûts intermédiaires : w_1= 0.2 et w_2= 0.8, w_1= 0.5 et w_2= 0.5, w_1= 0.8 et w_2= 0.2. Constatant sur un grand jeu de tests que la description reposant sur la description des sommets par le contexte de formes demeure plus discriminante que l'information contenue dans la structure voisine de chaque sommet, nous avons choisi de fixer les valeurs de poids à 0.8 pour w_1 et 0.2 pour w_2 . Les deux constantes a et d ont été choisies expérimentalement.

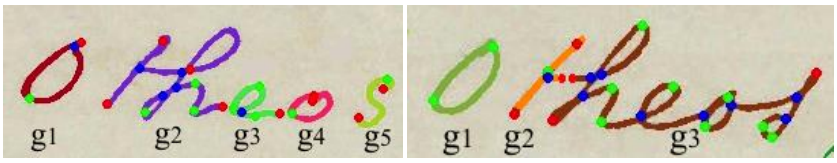
$$C_{substitution} = w_1 C_{SC\_cost} + w_2 C_{local\_structure} \quad (2)$$
$$C_{suppression} = d \quad (3)$$
$$C_{insertion} = a \quad (4)$$

Compte tenu les intervalles de valeurs choisis pour exprimer les coûts de substitution, les deux constantes restantes liées au coût d'insertion et de suppression sont fixées à 0.5. C_{SC} représente le coût associé au descripteur de contexte de formes des sommets. Il s'exprime par la distance du χ^2 square entre deux descripteurs de contexte de formes et $C_{local\_structure}$ le coût de substitution lié aux valeurs de longueurs des arêtes. Pour un sommet considéré, on estime le coût de substitution de la façon suivante : $C_{local\_structure} = 1 - e_{short}/e_{long}$, avec e_{short} la longueur du contour le plus court rattachant le sommet considéré à un voisin direct et e_{long} la longueur la plus importante.

Typiquement, la distance d'édition entre graphes est calculée avec une approche de type recherche arborescente présentant une complexité calculatoire exponentielle. L'utilisation d'une approximation de la distance d'édition par l'algorithme de Munkres résout ce problème en temps polynomial. Ces différences sont importantes à considérer pour le traitement de masses de données conséquentes, ce qui est notamment le cas pour les applications d'exploration de collections manuscrites anciennes.

4.2 Appariement complet de mots

Puisque un mot est représenté par une séquence de graphes (représentant les objets connexités), la distance dynamique DTW est la mesure la plus adaptée pour permettre un appariement souple entre graphes en intégrant la distance d'édition présentée en section 4.1. Un des éléments fondamentaux de cette distance est de permettre d'intégrer une variation parfois importante de connexités entre deux mots à comparer. En effet, on peut rapidement observer que les mots présents dans un texte manuscrit sont marqués par la présence de ruptures en des points non réguliers du texte, ceci se traduit par la présence d'objets connexités en nombre variable entre l'image *Requête* et l'image *Cible.* Ces variations sont remarquables entre scripteurs mais également pour un même scripteur, voir Figure 4. Cette figure présente un exemple de deux instances d'un même mot écrit par la même main possédant respectivement cinq et trois connexités. Cette particularité graphique va plus généralement pouvoir se résoudre par la comparaison de *n* graphes (issus du mot *Requête*) avec *m* graphes (issus du mot *Cible*).



(a) (b)

Figure 4. Exemple de deux instances du mot "*Otheos*" composées d'un nombre variable de connexités (*Correspondances clandestines 18ème siècle*, projet CITERE : ANR Blanc SHS 2009-2011, Circulations, Territoires et Réseaux de l'âge classique aux Lumières.*)*

Pour rendre la mesure de similarité robuste aux variations internes de connexités, nous avons proposé de les reconsidérer en réduisant leur nombre par un processus de fusion. Les graphes affectés à une même connexité à l'issue de l'appariement par DTW sont proposés à la fusion et reconsidérés comme un unique graphe. Dans ce cas, la séquence de graphes de la *Requête* et la séquence de l'image *Cible* peuvent être transformées. A l'issue de ces opérations de fusion, la distance d'édition entre graphes est à nouveau calculée créant ainsi de nouvelles correspondances. La distance moyenne obtenue à l'issue de ces nouvelles évaluations internes entre graphes est la mesure finale retenue entre les deux mots. Considérons pour cela un exemple reposant sur un mot de taille moyenne : "*Otheos*" représenté à la Figure 4. On considère $a = g_a^1, \cdots, g_a^5$ la séquence de graphes représentant le mot de la Figure 4(a) et b= $g_b^1, \cdots, g_b^3$ représente le mot de la Figure 4(b). A l'issue de l'étape d'alignement obtenue par la distance dynamique DTW, les graphes g_a^1 et g_b^1 sont appariés, et les graphes g_a^2 et g_a^3 associés au graphe g_b^2 et les trois graphes g_a^3, g_a^4, g_a^5 au graphe g_b^3. Par conséquent, en respectant la procédure de fusion, les graphes g_a^2 , g_a^3 , g_a^4 et g_a^5 sont rassemblés en un seul et même grapherenommé pour l'exemple $(g_a^2)'$, ayant en commun le graphe g_a^3 réagissant

positivement à deux graphes communs du mot *b*. Les graphes g_b^2 et g_b^3 sont également fusionnés en un unique graphe $(g_b^2)'$, car tous deux sont appariables à g_a^3. Ainsi les deux représentations des images (a) et (b) peuvent être reformulées comme deux nouvelles séquences réduites de deux graphes. La distance d'édition entre les deux nouvelles séquences de graphes est à nouveau calculée : c'est-à-dire entre les graphes g_a^1et g_b^1, puis entre les graphes $(g_a^2)'$ et $(g_b^2)'$, distances que nous notons respectivement $ged(g_a^1, g_b^1)$ et $ged((g_a^2)', (g_b^2)')$. La distance finale entre les deux mots (a) et (b) s'exprime alors par la moyenne des deux distances.

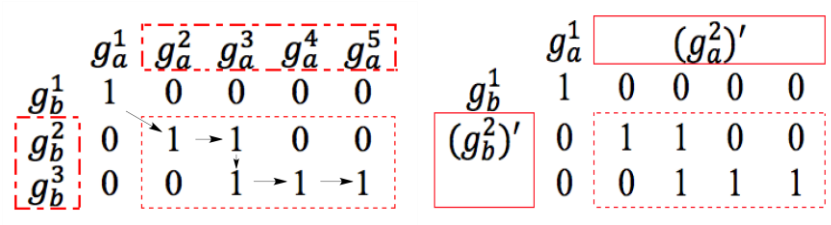


Figure 5. Illustration du processus de fusion de blocs reposant sur la distance DTW (les flèches indiquent le chemin optimal suivi pour l'appariement).

5. Expérimentations et protocoles

Les expérimentations sur lesquelles portent cette étude reposent sur deux ensembles de données : la base de données désormais usuelle pour les applications de word spotting et composée des lettres manuscrites de Georges Washington (1780) [3] et le registre de mariages célébrés en la Cathédrale de Barcelone dont les manuscrits sont datées des périodes allant de 1451 à 1905 et qui nommé la base *5CofM* (« *The Five Centuries of Marriages Database* ») [16]. Les résultats obtenus sur ces deux bases de tests ont été comparés à quatre autres approches développées dans le domaine du word spotting. La première approche comparative a été développée dans [3] par Manmatha et al. et traite de l'alignement de séquences utilisant la distance DTW exploitant des attributs structurels de contours. La seconde approche proposée par J. Lladós concerne la construction de sacs de mots visuels à plusieurs échelles générant un modèle statistique pour la comparaison de mots sur la base d'un clustering de formes [10]. La troisième approche vise le développement d'un modèle pseudo-structurel utilisant une représentation reposant sur les descripteurs Loci [10]. Ces descripteurs structurels encodent la fréquence d'intersections entre le mot centré en un point caractéristique choisi dans l'écriture et huit orientations autour de ce point. La quatrième approche servant de référence est structurelle et fondée sur une modélisation par graphe utilisant les descripteurs nommés *bag-of-paths* par analogie avec la notion plus communément admise de *bag-of-word* [10]. L'étude des résultats est réalisée à partir de l'exploitation d'indicateurs statistiques de rappel, de précision et de précision moyenne illustrés dans la section suivante.

5.1 Données expérimentales et évaluation de performances

La première base sur laquelle repose les tests est la base composée de vingt pages de la collection de lettres manuscrites de George Washington (*GW*) [3]. La seconde évaluation repose sur l'exploitation des données issues de vingt sept pages du registre de mariages de la Cathédrale de Barcelone (*5CofM*) [16]. Les deux corpus sont conçus comme des corpus étalons disposant d'une segmentation en mots transcrits et permettant donc de réaliser une étude de performance pertinente. Pour la collection *GW*, on dispose de 4860 mots segmentés avec 1124 transcriptions, et pour la collection *5CofM* on dispose de 6544 mots associés à 1751 transcriptions. Tous les mots possédant au minimum 3 lettres et apparaissant au minimum 10 fois dans la collection sont sélectionnés comme mot requête. Par conséquent, les expérimentations se fondent sur précisément 1847 requêtes correspondant à 68 mots différents pour la base *GW* et 514 requêtes de 32 mots pour la base *5CofM*. La Figure 6 illustre quelques extraits de mots de la base des registres de mariages (*5CofM*).

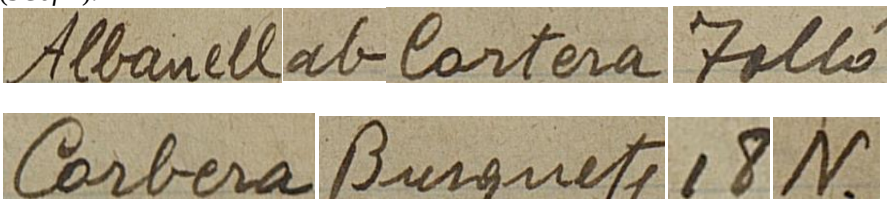


Figure 6. Exemples de mots du registre de mariages de la cathédrale de Barcelone, *5CofM* (1825).

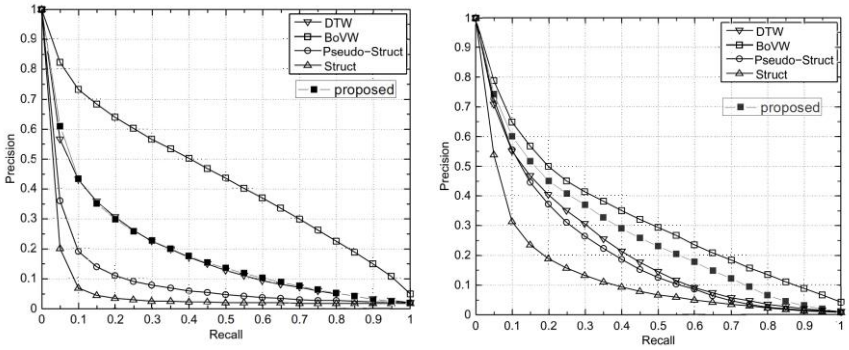
Afin d'évaluer les performances de l'approche que nous avons proposée, nous avons choisi trois indices relevant des valeurs de rappel et de précision. Considérant une requête, on notera *Rel* l'ensemble de réponses pertinentes relativement à cette requête et *Ret* l'ensemble des éléments effectivement retrouvés dans la base de tests.

- *P@n* — est la précision obtenue au rang *n*, considérant uniquement les *n* premiers retours du système. Dans nos tests, nous ne considérons que les retours aux rangs *n*= 10 et 20.
- *R-Précision* — la *R-Précision* est la précision obtenue après que *R* images aient été retrouvées, avec *R* le nombre de documents pertinents pour la requête considérée.
- *mAP* (*mean* Average Precision) — *mAP* ou *moyenne des précisions* obtenues chaque fois qu'un mot image pertinent est retrouvé. Il correspond à l'indice calculé à partir de chaque valeur de précision pour chaque rang. Pour une requête donnée, en notant *r(n)* la fonction indiquant le nombre de retours positifs du système au rang *n*, cette précision moyenne s'exprime comme le rapport suivant :

$$mAP = \frac{\sum_{n=1}^{|ret|}(P@n \times r(n))}{|rel|} \qquad (5)$$

5.2 *Résultats et discussion*

La Figure 7 présente la courbe de précision-rappel obtenues pour l'ensemble des tests pour les deux bases de données traitées (*GW* et *5CofM)*. Les Tables 1 et 2 illustrent les résultats complets d'évaluation de performances pour les deux bases.



(a) (b)

Figure 7. Courbes de précision-rappel pour (a) la base George Washington et (b) le registre de mariages de la cathédrale de Barcelone.

On peut tout d'abord constater sur les deux bases que notre proposition surpasse les performances des méthodes structurelles et pseudo-structurelles (dénommées respectivement « Pseudo-struct » et « Struct » et faisant références aux travaux décrits dans [10]). La méthode dénommée *BoVW* reposant sur les sacs de mots visuels réalise les meilleures performances sur les deux jeux de données, cela est sans doute en lien avec le processus de description des mots basé directement sur l'image à niveaux de gris sans segmentation binaire des traits d'écriture. Les autres méthodes comparées dans cette étude basent toutes leur description sur une version binaire des images, augmentant ainsi la sensibilité des mécanismes mis en jeux aux données bruitées. Dans notre approche, c'est davantage la qualité du squelette qui est mise en cause en provoquant des discontinuités dans la reconstruction des traits et qui génère un nombre de connexités souvent supérieur à celui qui est effectivement observé, voir Figure 8. Et bien que notre approche de l'appariement soit tolérante aux distorsions de l'écriture, la distance finale entre deux mots similaires peut être fortement impactée par les cassures de squelette.

	P@10	P@20	*R*-précision	mAP
Proposed	0.372	0.312	0.206	0.175
Manmatha	0.346	0.286	0.191	0.169
BoVW	0.606	0.523	0.412	0.422
Pseudo-Struct	0.183	0.149	0.096	0.072
Structural	0.059	0.049	0.036	0.028

Table 1. Performances pour l'ensemble des méthodes testées sur la base *GW*

	P@10	P@20	*R*-précision	mAP
Proposed	0.342	0.241	0.270	0.246
Manmatha	0.288	0.201	0.214	0.192
BoVW	0.378	0.289	0.303	0.3
Pseudo-Struct	0.273	0.189	0.199	0.178
Structural	0.155	0.12	0.118	0.097

Table 2. Performances pour l'ensemble des méthodes testées sur la base *5CofM*

L'approche statistique *BoVW* décrite à partir d'une sélection de points d'intérêt SIFT permet une description bas niveau robuste aux variations, elle est également compacte et complète dans sa prise en compte de l'information de luminance en chaque point d'intérêt. Elle est globalement plus insensible aux distorsions internes de l'écriture. Notons cependant que la classification des mots visuels issus de la méthode *BoVW* nécessite une comparaison reposant sur une représentation pyramidale des données (Algorithme *SPM* : Spatial Pyramidal Matching [10]) qui encode les relations spatiales non intégrées dans les codebooks visuels. Elle offre ainsi un meilleur pouvoir discriminant mais augmente de façon significative les temps de calcul, ce qui réduit son champ d'application (à de plus larges collections de documents).

Il est intéressant dans notre cas de constater que notre proposition réalise ses meilleures performances sur la base *5CofM*. La raison vient du fait déjà évoqué précédemment que les images de la collection *GW* présentent de plus nombreuses dégradations que celles contenues dans la base *5CofM*. Celles-ci impactent fortement la construction du squelette établi sur un algorithme simple d'amincissement morphologique. L'exploitation d'approches variationnelles comme les tenseurs d'inertie ou la diffusion anisotrope pourrait permettre localement de rehausser les traits dégradés et d'offrir une représentation plus continue respectant la morphologie de l'axe médian des écritures. Des travaux en ce sens ont été entrepris par Lebourgeois dans [17].

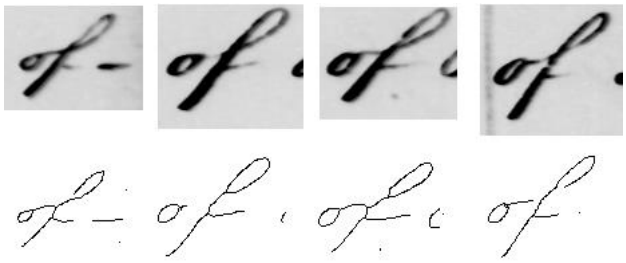


Figure8. Exemples de squelettes extrait par la méthode de Zhang [13] pour différentes instances d'un même mot extrait de la collection George Washington.

A l'exception de l'approche *BoVW*, notre proposition s'est montrée plus performante que les approches structurelles et pseudo-structurelles reposant sur des

graphes et procédant sans apprentissage. Cela illustre en particulier l'efficacité du choix de points structurels définis sur le squelette des mots pour diriger la structure du graphe qui le décrit, au lieu de points maximisant une variation locale d'intensité lumineuse, comme cela est généralement le cas pour les points d'intérêt plus standards. Par ailleurs, les informations contextuelles apportées par une description par *contexte de formes* (*SC*) constituent une description très riche de ces points.

Le coût global permettant de conduire la recherché de mot par similarité est difficile à établir du fait du grand nombre d'étapes allant de la description bas niveau à l'appariement proprement dit (pour l'approche BoVW, il est nécessaire de considérer la construction du codebook, la classification des points SIFT et enfin l'appariement complet). Nous pouvons cependant proposer de comparer qualitativement les durées de traitement, les performances relatives (déduites des taux figurant dans les Tables 1 et 2) et différents critères usuellement exploités pour caractériser les capacités d'adaptation des systèmes (facilité du passage à l'échelle, aisance à l'indexation), voir Table 3. Les indications '+' et '−' traduisent respectivement une bonne (et faible) capacité pour chaque critères (en colonne).

	Taille	Temps	Indexabilité	Prétraitement	Performance
Proposed	+	−	+ +	− −	+ +
Manmatha	+	− −	−	−	+
BoVW	− −	+	+	+	+ +
Pseudo-Struct	+ +	+ +	+ +	−	+
Structural	+ +	+ +	+ +	− −	−

Table 3. Comparaison qualitative de cinq approches de word spotting (reprise des tableaux de comparaisons 1 et 2)

6. Conclusion et perspectives

Ce papier est la présentation de deux contributions dans le domaine de la recherche de mots par similarité. Tout d'abord, nous avons proposé une nouvelle approche de la description des mots images fondée sur une structure de graphes construits sur une sélection de points structurels et morphologiques de l'écriture. Dans un second temps, nous avons proposé un nouveau mécanisme de capture de similarités souples basé sur une distance dynamique permettant de comparer les séquences de graphes formant les mots à l'aide d'une distance d'édition approximée et permettant de reconstruire une séquence optimale facilitant les appariements. L'information bas niveau de l'image est exploitée dans ses expressions de contours (par le descripteur de contexte de formes) et de squelette (permettant une sélection rapide de points d'intérêt structurels). La souplesse offerte par l'usage de la distance dynamique (DTW) pour l'appariement et son rôle dans le processus de fusion des objets connexes permet de compenser les distorsions subies par l'écriture lors de la formation des mots et les irrégularités relevées par l'étape de squelettisation. Celle-ci peut parfois créer artificiellement des connexités non-visibles ou ne faisant pas sens. Notre proposition repose ainsi sur un mécanisme de comparaison de graphes et

un usage intensif d'une distance d'édition approximée sans aucun apprentissage préalable. Cette approche n'avait jamais été exploitée dans un contexte de recherche de mots manuscrits. L'application à de nouveaux corpus plus volumineux constitue le prochain enjeu de cette étude, en lien avec les données du projet ANR CITERE aux volumes très conséquents. C'est donc sur la complexité calculatoire qu'un effort devra être réalisé, sachant que la comparaison de graphes même partielle est très consommatrice de puissance de calculs (DTW, coût d'édition en deux passes et itérations en un très grand nombre de fenêtres d'analyse sur une page de texte exploitant un descripteur de formes de dimension 60). Notre volonté de conserver les deux dimensions de la représentation par graphe au lieu de ramener la description à une séquence 1D de caractéristiques est un défi que nous tenterons de maintenir malgré l'augmentation de taille des corpus à analyser.

Actuellement, nos travaux portent sur l'élaboration d'une stratégie complète pré-sélectionnant des régions d'intérêt dans les images au fort potentiel en rapport avec les propriétés morphologiques du mot-requête soumis au système. Cette étape vise ainsi à rejeter massivement des propositions non pertinentes et de ne traiter que les fenêtres d'analyse candidates. Une reconsidération locale de la description par contexte de formes autour des sommets des graphes décrivant les objets connexes pourrait ainsi soutenir cette étape de recherche de zones candidates.

7. Remerciements

Nous tenons à remercier le professeur Antony McKenna responsable du projet CITERE et d'un fond de manuscrits numérisés uniques : « *Les correspondances clandestines de l'Europe des Lumières* » ainsi que « *Les correspondances de Pierre Bayle* ». Cette recherche est soutenue par un projet régional d'ARC de la région Rhônes-Alpes en lien avec le projet ANR CITERE et soutenue également par le projet espagnol TIN2012-37475-C02-02 ainsi que le projet européen ERC-2010-AdG- 20100407-269796 au sein du CVC de Barcelone.

8. Références

[1] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "HMM-based word spotting in handwritten documents using subword models," in Proc. of the ICPR, pp. 3416–3419, 2010.

[2] J. Rodriguez-Serrano, and F. Perronnin, "Handwritten word spotting using hidden markov models and universal vocabularies," PR, vol. 42, no. 9, pp. 2106–2116, 2009.

[3] T. Rath, and R. Manmatha, "Word spotting for historical documents," in Proc. of the ICDAR, vol. 9, no. 2-4, pp. 139–152, 2007.

[4] Y. Leydier, A. Ouji, F. LeBourgeois, and H. Emptoz, "Towards an omnilingual word retrieval system for ancient manuscripts," PR, vol. 42, no. 9, pp. 2089–2105, 2009.

[5] S. Lu, Y. Ren, and C. Suen, "Hierarchical attributed graph representation and recognition of handwritten Chinese characters," PR, vol. 24, no. 7, pp. 617–632,

1991.
[6] M. Zaslavskiy, F. Bach, and J. Vert, “A path following algorithm for the graph matching problem,” PAMI, vol. 31, no. 12, pp. 2227–2241, 2009.
[7] A. Fischer, C. Y. Suen, V. Frinken, K. Riesen, and H. Bunke, “A fast matching algorithm for graph-based handwriting recognition,” in Lecture Notes in Computer Science, vol. 7887, pp. 194–203, 2013.
[8] A. Fischer, K. Riesen, and H. Bunke, “Graph similarity features for HMM-based handwriting recognition in historical documents,” in Proc. of the ICFHR, pp. 253–258, 2010.
[9] Y. Chherawala, R. Wisnovsky, and M. Cheriet, “Tsv-lr: Topological signature vector-based lexicon reduction for fast recognition of premodern Arabic subwords,” in Proc. of the workshop HIP, pp. 6–13, 2011.
[10] J. Lladós, M. Rusinol, A. Fornés, D. Fernandez, and A. Dutta, “On the influence of word representations for handwritten word spotting in historical documents,” Pattern Recognition and Artificial Intelligence, vol. 26, no. 5, pp. 1 263 002.1–1 263 002.25, 2012.
[11] J. Puzicha, S. Belongie, and J. Malik, “Shape matching and object recognition using shape contexts,” PAMI, vol. 24, pp. 509–522, 2002.
[12] K. Riesen, and H. Bunke, “Approximate graph edit distance computation by means of bipartite graph matching,” Image and Vision Computing, vol. 27, no. 7, pp. 950–959, 2009.
[13] T. Y. Zhang, and C. Y. Suen, “A fast parallel algorithm for thinning digital patterns,” in Communication of the ACM, vol. 27, pp. 236–239, 1984.
[14] P. Wang, V. Eglin, C. Largeron, A. McKenna, and C. Garcia, “A comprehensive representation model for handwriting dedicated to word spotting,” in Proc. of the ICDAR, pp. 450-454, 2013.
[15] J. Munkres, “Algorithms for the assignment and transportation problems,” the Society for Industrial and Applied Mathematics, vol. 5, pp. 32–38, 1957.
[16] D. Fernandez, J. Lladós, and A. Fornés, “Handwritten word spotting in old manuscript images using a pseudo-structural descriptor organized in a hash structure,” Pattern Recognition and Image Analysis, vol. 6669, pp. 628–635, 2011.
[17] F. Lebourgeois, and H. Emptoz, “Skeletonization by Gradient Regularization and Diffusion,” In Proc. of the ICDAR, pp.1118-1122, 2007
[18] T-H. Do, S. Tabbone, and O.R. Terrades, “New Approach for Symbol Recognition Combining Shape Context of Interest Points with Sparse Representation,” In Proc. of the ICDAR, pp. 265-269, 2013.
[19] D. Hani, V. Eglin, S. Bres, and N. Vincent, “A new approach for centerline extraction in handwritten strokes: an application to the constitution of a codebook,” International Workshop on Document Analysis Systems, pp. 425-425, 2010.
[20] M.M. Luqman, J-Y. Ramel, J. Lladós, and T. Brouard, “Subgraph Spotting through Explicit Graph Embedding: An Application to Content Spotting in Graphic Document Images,” International Conference on Document Analysis and Recognition, pp. 870-874, 2011.
[21] R. Deriche, “Using Canny's criteria to derive a recursively implemented optimal edge detector,” Int. J. Computer Vision, vol. 1, pp. 167–187, 1987.