Différenciation de documents textes Arabe et Latin par filtre de Gabor

Sofiene Haboubi, Samia Snoussi Maddouri et Noureddine Ellouze

Laboratoires des Systèmes et de Traitement de Signal
Ecole Nationale d'ingénieurs de Tunis. BP 37. le Bélvèdere 1002, Tunis, Tunisie
Sofiene.Haboubi@istmt.rnu.tn; Samia.Maddouri@enit.rnu.tn
Noureddine.Ellouze@enit.rnu.tn

Résumé Une des premières étapes dans le problème de la reconnaissance automatique de documents textes est l'identification de la langue. Dans cet article nous proposons une méthode d'identification de la langue qui traite le cas des écritures arabes et latines dans des documents imprimés ou manuscrits. Cette méthode est basée sur une analyse spatio-fréquentielle, en appliquant les filtres de Gabor, pour l'extraction des caractéristiques sous forme d'un vecteur de dimension 32. Ceci est effectuée après une normalisation du texte traité (correction d'inclinaison, normalisation d'interlignes et d'intermots,…). L'apprentissage est effectuée sur une base de 400 documents classés selon leur langue (arabe ou latine) et leur nature (imprimée ou manuscrite).

Mots clés différenciation multilingues, analyse spatio-fréquentielle, filtre de Gabor.

1 Introduction

Avec l'utilisation croissante des ordinateurs dans les affaires et d'autres secteurs, de plus en plus les organismes convertissent leurs documents papiers en documents électroniques qui peuvent être traités par des ordinateurs. Ceci mène également au développement de l'OCR. La reconnaissance de la langue du document est considérée comme une étape de prétraitement, cette étape est devenue délicate dans le cas de documents manuscrits. Les méthodes existantes pour la différenciation d'écritures sont regroupées par [6] en quatre classes principales selon les niveaux d'informations analysés : les méthodes basées sur une analyse de bloc de texte, les méthodes basées sur l'analyse de ligne de texte, les méthodes basées sur l'analyse d'objets connexes et les méthodes basées sur des analyses mixtes.

L'étude préliminaire effectuée, montre que la majorité des méthodes de différenciation ne traitent que les documents textes imprimés et qu'il y a peu qui s'intéressent à l'écriture arabe. Parmi ces dernières, la méthode proposée par [1] qui développe une stratégie de différenciation d'écritures exploitant les informations disponibles dans les trois principaux niveaux d'une entité textuelle à identifier : bloc, ligne ou mot, et entité connexe. La méthode proposée par [5] qui est basée sur l'extraction de caractéristiques morphologiques et géométriques en cherchant à exploiter les particularités de chacun des scripts arabe et latin. La méthode proposée par [7] qui est basée sur une analyse fractale du style de l'écriture.

Le cadre général de ce papier s'articule autour des systèmes de reconnaissance d'écritures multilingues. Notre objectif essentiel est de proposer une méthode de différenciation entre les écritures arabes et latines de ses deux natures imprimées ou manuscrites. Nous présentons les principales

caractéristiques de ces deux écritures dans la deuxième section. Dans la troisième section nous détaillons notre méthode de différenciation basée sur les filtres de Gabor. Les résultats expérimentaux sont présentés dans la section quatre.

2 Discrimination de l'écriture

2.1 Caractéristiques de l'écriture

Toute écriture présente une série de caractéristiques qui lui sont propres et qui tiennent, certes au groupe social, à la langue et à l'époque dont elle est l'expression, mais aussi au support sur laquelle l'écriture est tracée, au stylet et aux habitudes du scripteur.

Pour connaître une écriture, il faut tenir compte des notions suivantes :

- La forme qui représente l'aspect de la lettre.
- Le module qui indique les dimensions des formes, à savoir les largeurs, les hauteurs et les ordres des grandeurs relatives (rapport entre l'hauteur des lettres et l'hauteur du corps du mot).

L'alphabet latin contient 26 lettres, 26 lettres majuscules et des lettres accentuées. Sachant que les alphabets suivent quelques variations avec les langues et avec le formatage des textes (certains ne sont écrits qu'en majuscules, certains sont écrits sans accents,…).

L'écriture arabe est cursive et présente divers diacritiques. Un mot arabe est une séquence d'entités connexes entièrement disjointes nommées pseudo-mots, qui est à son tour formé de un ou plusieurs caractères.

Contrairement à l'écriture latine, les caractères arabes s'écrivent de droite à gauche, et ne comportent pas de lettres majuscules. La forme des caractères varie suivant leur position dans le mot : initiale, médiane, finale et isolée.

Dans le cas du manuscrit, les caractères, arabe ou latin, peuvent varier dans leurs propriétés statiques et dynamiques. Les variations statiques concernent la taille et la forme, tandis que les variations dynamiques concernent le nombre de segments et de diacritiques ainsi que leur ordre. La présence des points diacritiques et leur position ainsi que leur nombre jouent un rôle primordial pour séparer les caractères en famille. En effet, en arabe, on compte 15 lettres, parmi les 28 de l'alphabet qui possèdent des points. La notion de voyelle n'existe pas sous sa forme classique. Les voyelles arabes se placent au-dessus ou en dessous du caractère. Dans la figure 1, nous donnons un exemple pour chaque nature d'écritures présentées.

(a) écriture arabe manuscrite

(b) écriture latine manuscrite

هذه كتابة عربيّة مطبعيّة

(c) écriture arabe imprimée

C'est une écriture latine imprimée

(d) écriture latine imprimée

Figure 1 : Exemples d'écritures

2.2 Analyse du document texturé

Tout document texte peut être vu comme une image texturée. L'analyse de la texture est utilisée dans des domaines de plus en plus variés : la caractérisation, la segmentation d'image ou la reconnaissance des formes [2]. Il existe différentes approches d'analyse de texture : les approches structurelles, statistiques, spatio-fréquetielles et fractales. Les approches structurelles conviennent plus pour les textures périodiques, alors que les autres approches conviennent plus pour les textures aléatoires. L'approche spatio-fréquentielle comprend la méthode par filtrage de Gabor et la transformée par ondelettes. Dans cette contribution, nous nous sommes intéressés à la méthode par filtre de Gabor.

3 Différenciation de l'écriture Arabe-Latine par filtre de Gabor

La méthode que nous proposons pour la différenciation des écritures arabe et latine est inspirée de la méthode appliquée par Tan T.N [8,9], pour l'identification du scripteur d'un texte écrit en anglais, et [4] pour l'identification du scripteur d'un texte écrit en arabe. Notre méthode de différenciation comprend trois phases : prétraitement, extraction de caractéristiques par filtre de Gabor et classification.

3.1 Prétraitement et normalisation de l'écriture

L'image d'un document peut contenir différents espaces interlignes et inter-mots, différentes longueurs de lignes. Aussi, les lignes peuvent êtres inclinées. Pour minimiser l'influence de ces facteurs sur la caractérisation de la texture, le document à analyser doit être normalisé pour créer un bloc de texte uniforme. La phase de normalisation consiste à :

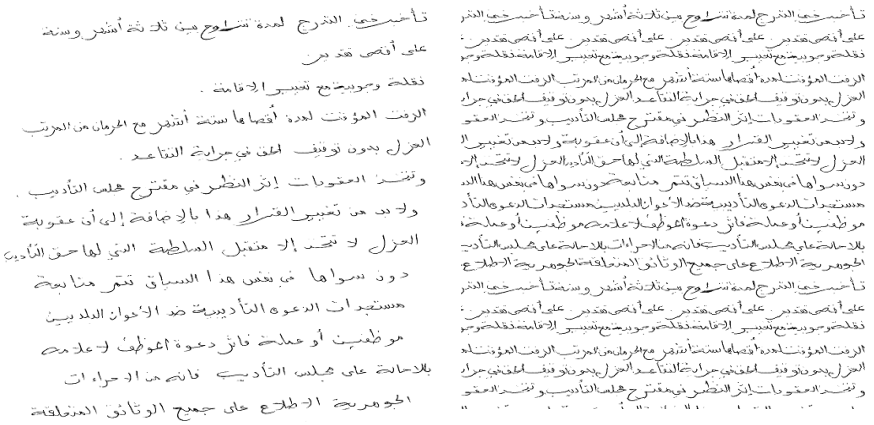
- Corriger l'inclinaison : cette étape, permet de détecter et de corriger l'angle d'inclinaison des lignes de texte en se basant sur le calcul de la projection horizontale et l'entropie E de l'image texte donnée par la formule (1) [3].

$$E = -\sum_i P_i \log(P_i) \quad et \quad P_i = N_i / N \tag{1}$$

Avec N_i : le nombre de pixels noirs dans la i^{ieme} ligne, et N : le nombre total de pixels noirs.

- Normalisation d'interlignes : l'objectif de cette étape est de supprimer les espaces vides pour avoir une image texturée uniforme. Pour cela, il faut réduire au maximum l'espace interlignes. Ceci nécessite d'abord une localisation des lignes par projection horizontale, et puis une recherche des bords inférieur et supérieur de cette ligne. Enfin, une élimination de l'espace entre le bord inférieur de la ligne au-dessus et le bord supérieur de la ligne en dessous.
- Normalisation d'espacement : pour chaque ligne, nous passons une projection verticale qui sera examiné de droite à gauche ou inversement, on élimine les pixels blancs par un simple décalage.
- Normalisation des marges : l'objectif de cette étape est de compléter les lignes incomplètes. Pour cela, on doit parcourir l'image texte ligne par ligne, l'espace entre le dernier pixel de la ligne et l'extrémité gauche de l'image doit être remplit par un bloc de même taille extrait de l'entête de la ligne considérée. Après la normalisation d'espacement, l'image subit un rétrécissement et le bloc devient partiellement vide. Pour avoir une image complète, il suffit de copier un bloc de même taille à partir de l'entête de l'image.

Le résultat de prétraitement est illustré dans la figure 2.



(a) Texte avant le prétraitement (b) Texte après le prétraitement

Figure 2 : Textes avant et après prétraitement

3.2 Extraction des caractéristiques

Après avoir normalisé l'image de texte manuscrit, il faut s'intéresser à l'étape de sa caractérisation, qui a pour objectif de déterminer le vecteur représentatif de l'image en utilisant les filtres de Gabor. Un filtre de Gabor est une fonction sinusoïdale modulée par une enveloppe gaussienne. La fonction sinusoïdale est caractérisée par sa fréquence et par son orientation. Ainsi un filtre de Gabor peut être vu comme un détecteur d'arêtes d'orientation particulières, puisqu'il réagit aux arêtes perpendiculaires à la direction de propagation du sinus [10]. Le filtrage par Gabor conserve les aspects temporels et fréquentiels du signal. Dans le domaine spatial, l'application des filtres de Gabor est effectuée en calculant la convolution de l'image avec une fonction réglée à une des textures. Le processus est montré dans l'équation (2).

$$q(x, y) = p(x, y) \times h(x, y) \tag{2}$$

$q(x, y)$ est le pixel filtré de l'image résultat.

$h(x, y)$ est le filtre de Gabor

$p(x, y)$ est le pixel de l'image original.

Théoriquement, nous pouvons faire tous les calculs dans le domaine fréquentiel ; le produit de convolution se réduit à une simple multiplication des transformées de Fourier. Le processus est montré dans l'équation (3).

$$q(x, y) = TF^{-1}(P(u, v)\, H(u, v)) \quad tel\ que\ P(u, v) = TF(p(x, y))\ et\ H(u, v) = TF(h(x, y)) \tag{3}$$

Le filtre de Gabor multi-canal 2D est représenté par l'équation (4).

$$h(x, y) = g(x', y') \exp(2\pi i f x') \tag{4}$$

Avec $g(x', y')$ est la fonction gaussienne 2D donnée par la formule (5).

$$g(x', y') = \exp\left[-\frac{1}{2}(x'^2 + y'^2) \Big/ \sigma^2\right] \tag{5}$$

avec $(x', y') = (x\cos\theta + y\sin\theta, -x\cos\theta + y\sin\theta)$ les coordonnées (x, y) tournées d'un angle θ .

Les paramètres f et θ représentent la fréquence et l'orientation du signal sinusoïdale et constitue le paramètre de l'espace du filtre de Gabor. Dans le domaine spatial, la fonction de Gabor est représentée par l'équation (6).

$$h(x,y) = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{1}{2}\left(\frac{x^2+y^2}{\sigma^2}\right)\right]\cos(2\pi fx) \tag{6}$$

La transformée de Fourier du filtre peut aussi être représentée par l'équation (7)

$$H(u,v) = \exp\left[-2\pi^2\sigma^2((u-f)+v^2)\right] + \exp\left[-2\pi^2\sigma^2((u+f)+v^2)\right] \tag{7}$$

Pour évaluer notre méthode, nous avons commencé par générer la fonction de Gabor dans un espace discret de taille 61*61. Après nous lui appliquons une normalisation pour que les valeurs soient comprises dans l'intervalle [0..1]. Ensuite, nous filtrons l'image texte par les différents filtres obtenus. Les paramètres les plus importants du filtre de Gabor sont la fréquence radiale et l'orientation, ils définissent la localisation du canal dans le plan fréquentiel. D'après Tan [8], chaque image de taille N*N, les fréquences les plus significatives sont dans l'intervalle [0..N/4] et à la puissance 2. Étant donné que l'extraction des caractéristiques doit être effectuée sur des blocs de texte de même taille, nous avons fait le choix d'une taille de blocs égale à 128*128. Vu cette taille, nous devons utiliser dans notre système les fréquences : 4, 8, 16 et 32. La largeur de la bande passante est choisie de façon à être inversement proportionnelle à la fréquence centrale : 8, 4, 2 et 1. Pour chaque fréquence centrale, nous considérons les quatre orientations suivantes : 0°, 45°, 90° et 135°, donnant en total 16 images filtrées (4 fréquences et 4 orientations).

Pour chaque image, on collecte la moyenne et l'écart type pour former un vecteur représentatif composé de 32 caractéristiques. Dans le tableau 1, nous donnons 4 exemples de vecteurs représentatifs pour des textes arabe et latin de nature imprimée et manuscrite.

Tableau 1 : Exemple de vecteurs représentatifs

(a) texte arabe imprimé															
2.81	1.19	3.27	1.19	14.8	12.0	40.0	14.1	1.67	0.78	1.77	0.78	23.6	18.1	27.3	16.1
1.14	0.58	1.13	0.58	29.2	17.9	28.8	19.1	1.34	0.45	1.28	0.45	17.6	15.8	21.5	16.3
(b) texte arabe manuscrit															
3.32	1.38	3.83	1.39	7.39	22.6	8.60	2.02	2.60	0.92	2.21	0.92	14.6	14.2	19.7	10.6
1.34	0.69	1.44	0.69	15.0	17.6	20.2	13.1	1.57	0.54	1.63	0.54	13.6	17.0	18.6	16.2
(c) texte latin imprimé															
2.92	1.15	3.22	1.15	12.0	4.98	36.7	5.26	1.73	0.81	2.22	0.81	13.5	8.83	23.0	9.81
1.19	0.61	1.33	0.61	31.6	18.5	29.4	18.1	1.37	0.48	1.48	0.48	35.7	32.9	30.5	33.4
(d) texte latin manuscrit															
2.95	1.10	2.55	1.09	13.2	7.31	37.2	9.77	1.73	0.76	1.78	0.76	14.9	9.19	22.2	13.7
1.14	0.68	1.30	0.59	33.9	17.1	26.1	21.8	1.33	0.47	1.52	0.47	25.1	25.0	22.1	28.4

3.3 Classification

L'identification de la langue du document texte en se basant sur les vecteurs caractéristiques est un problème typique en reconnaissance. Nous utilisons la distance euclidienne pondérée donnée par la formule (8)

$$d(k) = \sum_{i=1}^{N} \frac{(f_i - f_i^k)^2}{(v_i^k)^2} \tag{8}$$

f_i est la i^{ieme} caractéristique du document inconnu.

f_i^k est la i^{ieme} caractéristique d'un document du k^{ieme} classe et v_i^k son écart type.

Dans notre approche de différenciation, nous disposons de quatre classes selon la langue de l'écriture et la nature du document traité: arabe imprimé, latin imprimé, arabe manuscrit et latin manuscrit. Chaque classe est définie par un nombre de vecteurs caractéristiques. Un document inconnu appartient à la même classe que la k^{ieme} vecteur qui possède la valeur minimale de *d(k)*.

4 Résultats

Dans nos premières expérimentations, nous nous sommes limités sur des documents textes écrits en arabe ou en latin, imprimé ou manuscrit. La base d'apprentissage utilisée contient 400 documents : 100 textes arabes imprimés en 10 fontes différentes, 100 textes latins imprimés en 10 fontes différentes, 100 textes arabes manuscrits de 10 scripteurs différents et 100 textes latins manuscrits de 10 scripteurs différents. Dans le tableau 2, nous donnons des résultats statistiques sur les distances entre classes, et dans le tableau 3 les distances maximales trouvées entre les vecteurs de la même classe.

Tableau 2 : Les distances inter-classes

Distance	Arabe Imprimé	Arabe Manuscrit	Latin Imprimé	Latin Manuscrit
Arabe Imprimé	0	2.96	7.31	6.13
Arabe Manuscrit	2.96	0	11.36	6.19
Latin Imprimé	7.31	11.36	0	6.79
Latin Manuscrit	6.13	6.19	6.79	0

Tableau 3 : Les distances intra-classes

Distance intra-classes	Arabe Imprimé	Arabe Manuscrit	Latin Imprimé	Latin Manuscrit
	29.94	25.93	41.73	30.60

La base de test contient 200 documents textes écrits en arabe et latin sous les deux natures imprimée et manuscrite. Les résultats de différenciation sont présentés dans le tableau 4.

Tableau 4 : Les taux de reconnaissance

	Arabe	Latin	A/L
Imprimé	44%	96%	70%
Manuscrit	62%	86%	74%
I/M	82%	92%	

Les résultats présentés dans le tableau 4, montrent que l'utilisation des filtres de Gabor dans la différenciation des écritures arabe et latine est possible. Le taux de détermination de l'écriture arabe est estimé à 82%, et celui de Latine à 92%. Les résultats du tableau 2, présentent les distances euclidiennes qui séparent les différentes classes (LatinImprimé, LatinManuscrit, ArabeImprimé et ArabeManuscrit). La distance minimale trouvée est celle qui sépare les deux natures de l'écriture arabe, qui justifie la confusion présentée dans le tableau 4 (44% et 62%). La deuxième confusion probable est entre l'écriture latine manuscrite et l'écriture arabe, qui justifie le taux de différenciation trouvé 86%.

5 Conclusion

Nous avons présenté dans cet article, une méthode d'identification des écritures arabe et latine. Cette méthode est basée sur une analyse spatio-fréquentielle en appliquant les filtres de Gabor. Les premiers résultats obtenus sont très encourageants et indique que l'on peut encore les améliorer. Parmi les améliorations possibles, nous pouvons faire une étude sur les coefficients du filtre de Gabor pour maximiser les distances entre les 4 classes et minimiser les distances inter-classes. Le but de cette étude est de minimiser la probabilité de confusion entre les classes.

Nous pensons encore à généraliser notre méthode aux documents mixtes (arabe et latin), tel qu'on peut discriminer les écritures, arabes et latines sous ses deux natures imprimées et manuscrite, d'un même document texte.

Références

[1] BENNASRI A., ZAHOUR A, TACONET B., Arabic script preprocessing and application to postal adresses, Proc. of ACIDCA'2000, Tunisia, 2000, pp. 74-79.

[2] BLOCH I., MATIGNON D., PESQUET B., SCHMITT H., SIGELLE F., Le traitement des images, Cours ANIM Version 5.0, Département TSI, Telecom Paris, France, 2004.

[3] COTE M., CHERIET M., SUEN C. Y., LECOLINET E., Détection des Lignes de Base de Mots Cursifs à l'aide de l'Entropie, Colloque sur l'Intelligence Artificielle dans les Technologies de l'Information, Université McGill, Montré (Canada), 13-17 Mai 96.

[4] FADDAOUI N., HAMROUNI K., Personal identification based on texture analysis of arabic handwritten text. ICTTA 06, Syria, April 2006.

[5] KANOUN S., ENNAJI A., LECOURTIER Y., ALIMI A., Une approche de discrimination Arabe / Latin, Imprimé / Manuscrit, CIFED'00, 2000, pp. 121-129.

[6] KANOUN S., Approche affixale pour la reconnaissance de textes arabes dans des documents multilingues", Thèse, Université de Rouen, France, 2001

[7] SEROPIAN A., GRIMALDI M., VINCENT N., Différenciation entre alphabets dans des textes manuscrits, CIFED'04, 2004.

[8] TAN T. N., Texture Features Extraction via visual cortical channel modelling, Proceeding of the 11th IAPR International Conference Pattern Recognition vol. III, 1992.

[9] TAN T. N., SAID H. E. S., BAKER K. D., Personal identification based on handwriting, National Laboratory of Pattern Recognition , Institue of Automation, Chinese Academy of Sciences, Décembre 2000.

[10] VINCENT L., Texture Segmentation Using Gabor Filters, Center For Intelligent Machines, McGill University, December 2000.