Différenciation d'écritures arabe et latine de natures imprimée et manuscrite par approche globale

Karim Baati, Slim Kanoun, Mohamed Benjlaiel

Différenciation d'écritures arabe et latine de natures imprimée et manuscrite par approche globale

Karim Baati*— Slim Kanoun* —Mohamed Benjlaiel*

** REGIM : REsearch Group on Intelligent Machines, Université de Sfax, Ecole Nationale d'Ingénieurs de Sfax (ENIS), BP 1173, Sfax, 3038, Tunisie.*

karim.baati@ieee.org ; slim.kanoun@yahoo.fr ; benjlaiel@yahoo.fr

RÉSUMÉ. *Cet article s'inscrit dans le cadre du problème de la différenciation d'écritures. Notre objectif est d'utiliser l'approche globale pour traiter le problème de la différenciation de l'arabe et du latin de natures imprimée et manuscrite. Dans le cadre de notre méthode, les caractéristiques sont extraites à l'aide de trois outils à savoir les filtres de Gabor, les matrices de cooccurrence de niveaux de gris et les ondelettes. Les résultats de différenciation obtenus sur une base de 800 documents latins et arabes, imprimés et manuscrits sont présentés. Le meilleur taux global de bonne identification est obtenu avec les matrices de cooccurrence de niveau de gris (84.75%). Il s'agit d'un taux assez satisfaisant en tenant compte de la simplicité de l'approche et de la complexité du problème traité.*

ABSTRACT. *In this paper, we treat the problem of script differentiation. Our aim is to use the global approach to resolve the problem of differentiation between the Arabic and the Latin scripts in printed and handwritten natures. In our method, features are extracted using three tools: Gabor filters, gray-level co-occurrence matrices and wavelets. The results of differentiation obtained on a base of 800 Arab and Latin documents, printed and handwritten are presented. The highest rate of correct identification is obtained with gray-level co-occurrence matrices (84.75%). We can consider this rate as satisfactory taking into account the simplicity of the approach and the complexity of the treated problem.*

MOTS-CLÉS : *différenciation d'écritures, approche globale, filtres de Gabor, matrices de cooccurrence de niveaux de gris, ondelettes.*

KEYWORDS: *script differentiation, global approach, Gabor filters, gray-level co-occurrence matrices, wavelets.*

1. Introduction

Depuis des années, les documents produits chaque jour dans le monde entier ont commencé à contenir des langues différentes et surtout dans les environnements internationaux. Par conséquent, l'automatisation de la différenciation d'écritures est devenue une étape de pré-reconnaissance nécessaire dans tout système de traitement automatique de documents multilingues.

Les recherches actuelles dans le domaine de l'écrit et du document visent à concevoir et à mettre en œuvre des systèmes automatiques capables de différencier un certain nombre d'écritures afin de sélectionner le système de reconnaissance (OCR) approprié à un document donné.

Trois approches peuvent être utilisées pour concevoir ces systèmes de différenciation d'écritures. Une première approche dite approche globale consiste à traiter les blocs de texte dans leur globalité. Une telle approche considère un bloc de texte comme étant une seule entité et donc ne fait pas recours à d'autres analyses au niveau ligne de texte et entité connexe (Wood et al. 1995). Elle suppose que le bloc de texte à identifier est normalisé (hauteur et largeur égaux), uniforme (espaces inter-lignes et inter-mots constants) et homogène (contient une seule écriture) (Tan, 1998 ; Tao et Tang, 2001 ; Busch et al., 2005). Une deuxième approche dite approche locale consiste à traiter les détails du bloc de texte en se basant ou bien sur une analyse des lignes de textes et des mots (Elgammal et Ismail, 2001 ; Fan et al., 1998) ou bien sur l'analyse d'entités connexes qui peuvent être soit naturellement segmentées dans le cas d'écritures non connexes comme le Latin, l'Asiatique, etc. soit issus d'une segmentation explicite dans le cas d'écriture de nature connexe comme l'Arabe, le Bangla, le Devanagari, etc. (Spitz, 1997 ; Hochberg et al., 1997). Une troisième approche de différenciation appelée approche mixte combine les deux premières approches en exploitant les informations disponibles dans les trois principaux niveaux d'une entité textuelle à identifier : bloc, ligne ou mot, et entité connexe (Chaudhury et Sheth, 1999 ; Bennasri *et al*., 2000 ; Pal et Chaudhuri, 2001).

Dans la majorité des méthodes utilisant l'approche globale, on considère souvent que celle-ci est plus rapide et plus simple qu'une approche locale qui reste toujours dépendante des caractéristiques spécifiques de l'écriture étudiée. Pourtant, l'approche globale n'était jamais exploitée afin de résoudre un problème qui combine la différenciation de l'écriture arabe et de la nature manuscrite.

Notre objectif est d'utiliser l'approche globale pour pouvoir différencier les écritures arabe et latine de nature imprimée et manuscrite. Ainsi notre système sera confronté à un problème de différenciation des quatre classes suivantes : l'arabe imprimé, le latin imprimé, l'arabe manuscrit et le latin manuscrit. La difficulté majeure qui en découle provient principalement des similarités qui existent entre ces

trois dernières classes à cause de la nature cursive de l'écriture arabe et du latin manuscrit.

Dans une première partie, nous rappelons les différents travaux portant sur la différenciation d'écritures par approche globale. Dans une seconde partie, nous présentons notre méthode de différenciation par approche globale. Enfin, nous exposons les résultats d'identification obtenus avec les différents outils utilisés.

2. Synthèse des travaux sur la différenciation d'écritures par approche globale

L'état de l'art des travaux réalisés sur la différenciation d'écritures par approche globale montre les constatations suivantes :

– la majorité des travaux n'ont pas traité la différenciation de l'écriture arabe, à l'exception de (Tan, 1998) qui propose une méthode qui identifie cette écriture parmi six types d'écritures imprimées,

– la plupart des travaux portent uniquement sur l'identification d'écritures de nature imprimée, à l'exception de (Singhal *et al*. 2003) qui propose une méthode pour l'identification d'écritures manuscrites,

– aucun travail n'a combiné l'identification de l'écriture arabe et de l'écriture manuscrite,

– bien que plusieurs outils puissent être utilisés pour l'extraction des caractéristiques dans le cadre de l'approche globale, telles que les matrices de cooccurrence de niveaux de gris (Peake et Tan, 1997 ; Busch *et al*., 2005), les ondelettes (Busch *et al*., 2005) et les dimensions fractales (Tao et Tang, 2001), les filtres de Gabor sont les outils les plus utilisés dans la majorité des travaux (Peake et Tan, 1997 ; Tan, 1998 ; Singhal *et al*. 2003 ; Busch *et al*., 2005),

– dans la majorité de ces travaux, un taux global d'identification très satisfaisant a été obtenu. Les taux obtenus pour les différentes références sont récapitulés dans le tableau 1.

Référence	Ecritures et natures	Taux
(Tan, 1998)	– Chinois, Coréen, Anglais, Grecque, Russe, Perse, et Malayalam. – Imprimée	96.7%
(Tao et Tang, 2001)	– Orientales (Chinois, Japonais, Coréen et Indien) et Euro-américaines (Anglais, Russe, Français, Allemand et Italien). – Imprimée	93.31%
(Manthalkar et Biswas, 2002)	– Sept écritures indiennes et l'Anglais. – Imprimée.	100 %
(Singhal *et al*., 2003)	– Romain (Anglais), Devanagari (Hindi), Bangla et Telugu. – Manuscrite	91.6 %
(Busch *et al*., 2005)	– Latin, Chinois, Japonais, Grec, Cyrillique, Hébreu, Sanskrit et Farsi. – Imprimée.	99 %
(Pan *et al*., 2005)	– Chinois, Coréen, Japonais et Anglais. – Imprimée.	98.76%
(Joshi *et al*., 2006)	– 10 écritures indiennes. – Imprimée.	97.11%

Tableau 1. *Tableau récapitulatif des taux obtenus dans le cadre de la différenciation d'écritures par approche globale.*

3. Notre méthode

3.1. *Démarche globale*

Comme dans la majorité des travaux qui ont utilisé l'approche globale (Tan, 1998 ; Busch *et al*., 2005 ; Singhal *et al*. 2003) , notre méthode est basée sur trois principales étapes :

– prétraitement,

– extraction de caractéristiques,

– classification.

Dans les paragraphes qui suivent, ces étapes seront détaillées.

3.2. *Prétraitement*

L'étape de prétraitement est une étape indispensable pour appliquer l'approche globale (Tan, 1998 ; Busch *et al*., 2005). Pour notre système, cette étape consiste à extraire à partir des documents de la base, des blocs de texte normalisés et uniformes. Nous avons choisi, dans notre cas, une taille normalisée des blocs de 512 x 512 pixels. Il s'agit d'une taille qui permet d'avoir une quantité d'information assez suffisante pour pouvoir caractériser la texture de chaque document. Pour construire le contenu de chaque bloc normalisé, nous commençons par l'extraction de la première ligne du bloc de texte original. Par la suite, les mots de cette ligne sont placés dans la première ligne du document normalisé en respectant un espace inter-mots constant et en veillant à ne pas dépasser la taille prédéfinie pour la largeur du bloc (les entités connexes situés au delà de 512 pixels sont supprimées). Le bloc normalisé est enfin obtenu en dupliquant cette première ligne sur tout le bloc, en respectant un espace inter-lignes constant et en veillant à ne pas dépasser la hauteur prédéfinie du bloc (512 pixels). Un exemple de prétraitement d'un bloc de texte arabe imprimé est présenté par la figure 1.

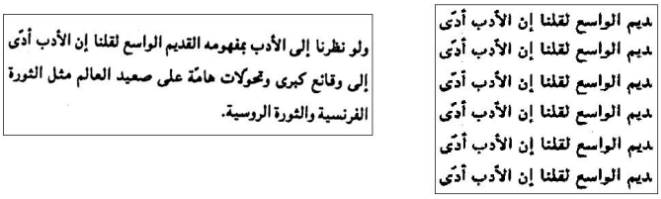

Figure 1. *Exemple de prétraitement d'un bloc de texte arabe imprimé.*

3.3. *Extraction de caractéristiques*

Après avoir effectué les prétraitements précités sur les différents documents, une étape d'extraction de caractéristiques et nécessaire. Notre méthode se sert de trois outils : les filtres de Gabor, les matrices de cooccurrence de niveaux de gris et les ondelettes. Dans la suite, nous allons présenter les outils utilisés par notre méthode ainsi que les valeurs numériques des différents paramètres.

3.3.1. *Les filtres de Gabor*

L'énergie de l'image de sortie après application d'un banc de filtres de Gabor sur un document original est toujours considérée comme une caractéristique pouvant permettre d'avoir de bons résultats d'identification même en présence d'une petite base de documents.

En 2D, les filtres de Gabor sont représentés par une paire de filtres réels définis par :

$$h_e(x, y; f, \Theta) = g(x, y)\cos\big(2\pi f(x\cos\Theta + y\sin\Theta)\big)$$
$$h_o(x, y; f, \Theta) = g(x, y)\sin\big(2\pi f(x\cos\Theta + y\sin\Theta)\big)$$

Les termes f et Θ étant respectivement la fréquence radiale et l'orientation du filtre.

La fonction $g(x, y)$ est la fonction gaussienne en 2D définie par :

$$g(x, y) = \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_{x^2}} + \frac{y^2}{\sigma_{y^2}}\right)\right]$$

Les termes σ_x et σ_y étant les variances de la gaussienne respectivement suivant les axe x et y.

Afin d'appliquer les filtres de Gabor pour l'extraction des caractéristiques des différents documents, nous avons adopté la même démarche présentée dans (Tan, 1998). Cette démarche est basée sur les deux étapes suivantes :

– à partir de la matrice $I = \{i(x, y)\}$ d'une image originale, on applique les filtres de Gabor pour obtenir à la sortie une matrice $Q = \{q(x, y)\}$. Les éléments $q(x, y)$ de la matrice Q sont calculés de la manière suivante :

$$q(x, y; f, \Theta) = \sqrt{q_e^2(x, y; f, \Theta) + q_o^2(x, y; f, \Theta)}$$
$$q_e(x, y; f, \Theta) = \mathrm{conv}[h_e(x, y; f, \Theta), i(x, y)]$$
$$q_o(x, y; f, \Theta) = \mathrm{conv}[h_o(x, y; f, \Theta), i(x, y)]$$

La fonction conv étant la fonction convolution en 2D,

– à partir de la matrice Q, certaines caractéristiques peuvent être extraites afin de caractériser la texture du document initial. Parmi ces caractéristiques, on peut calculer la moyenne et l'écart-type des valeurs de la matrice Q (Tan, 1998).

Ainsi nous remarquons que pour chaque valeur f et pour chaque valeur Θ des filtres de Gabor utilisés, deux caractéristiques peuvent être extraites à chaque fois. Il est clair alors qu'en augmentant le nombre de fréquences et d'orientations choisies, nous augmentons la taille du vecteur caractéristique caractérisant les différents blocs.

Concernant notre système, nous avons fixé les valeurs $\sigma_x = \sigma_y = 1.5$ et nous avons choisi expérimentalement 3 valeurs de fréquences ($f = 8$, 16 et 32) et 8 valeurs pour les orientations comprises entre 0 et 2π et équidistantes de $\pi/4$.

Deux caractéristiques étant extraites pour chaque valeur de f et de Θ, le vecteur caractéristique est alors de taille 48.

3.3.2. *Les matrices de cooccurrence de niveaux de gris*

La matrice de cooccurrence de niveaux de gris définit les relations entres les différents niveaux de gris d'une image. Pour une image à L niveaux de gris, la matrice de cooccurrence de niveaux de gris est une matrice de taille L x L dont les éléments sont définis par :

$$N_{d,\theta}(i,j) = \frac{C_{d,\theta}(i,j)}{\sum_i \sum_j C_{d,\theta}(i,j)}$$

Avec $C_{d,\theta}(i,j)$ le nombre de pixels de niveau de gris j situés d'une distance d et d'un angle θ par rapport aux pixels de niveau de gris i.

Le terme $N_{d,\theta}(i,j)$ est appelé version normalisée de $C_{d,\theta}(i,j)$. Parfois on peut utiliser la version symétrique de $C_{d,\theta}(i,j)$ et dans ce cas la matrice de cooccurrence de niveaux de gris est une matrice symétrique. La version symétrique est définie par :

$$S_{d,\theta}(i,j) = C_{d,\theta}(i,j) + C_{d,\theta+\pi}(i,j)$$

A partir des termes $N_{d,\theta}(i,j)$ ou $S_{d,\theta}(i,j)$ de la matrice de cooccurrence de niveaux de gris, on peut extraire certaines caractéristiques telles que l'entropie, l'énergie, le contraste, l'homogénéité et la corrélation.

Concernant notre système, la méthode proposée est celle présentée dans (Busch *et al.*, 2005). Il s'agit de binariser chaque document afin d'obtenir une image à deux niveaux de gris. La matrice de cooccurrence de niveaux de gris est, dans ce cas, de taille 2 x 2. En choisissant de créer une matrice de cooccurrence de niveaux de gris symétrique pour une distance d et une orientation θ donnés, 3 termes significatifs peuvent être extraits. En variant les valeurs de d et de θ, on augmente la taille du vecteur caractéristique.

Dans notre cas, 5 distances ($d = 1, 2, 3, 4$ et 5) et 4 orientations ($\theta = 0, 45, 90$ et 135) sont choisies ce qui donne un vecteur caractéristique de taille 60.

3.3.3. *Les ondelettes*

En 2D, une transformée discrète en ondelettes peut être définie par:

$$A_j = [H_x * [H_y * A_{j-1}]_{\downarrow 2,1}]_{\downarrow 1,2}$$
$$D_{j1} = [G_x * [H_y * A_{j-1}]_{\downarrow 2,1}]_{\downarrow 1,2}$$
$$D_{j2} = [H_x * [G_y * A_{j-1}]_{\downarrow 2,1}]_{\downarrow 1,2}$$
$$D_{j3} = [G_x * [G_y * A_{j-1}]_{\downarrow 2,1}]_{\downarrow 1,2}$$

Les termes A_j et D_{jk} étant respectivement les coefficients d'approximation et de détails directionnels (respectivement suivant la verticale, l'horizontale et la diagonale), H et G : des filtres passe-bas et passe-haut et $\downarrow x, y$ l'opérateur de sous-échantillonnage suivant les axes x et y.

Pour notre système et comme dans (Busch *et al.*, 2005), les énergies des trois images de détails sont utilisées comme caractéristiques pour l'identification d'écritures. Ces énergies sont calculées de la manière suivante :

$$E_{jk} = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} D_{jk}(m,n)}{MN}$$

Les termes M et N étant respectivement la hauteur et la largeur de l'image de détails.

Ainsi pour chaque niveau de décomposition J, trois caractéristiques sont extraites (E_{j1}, E_{j2} et E_{j3}). Concernant notre système, nous avons choisi d'utiliser des ondelettes de type biorthogonales et d'effectuer 12 décompositions (J = 12) ce qui donne un vecteur caractéristique de taille 36.

3.4. Classification

Le classifieur des K plus proches voisins a été toujours considéré comme un classifieur simple et efficace. C'est pour cette raison que nous l'avons choisi pour notre système. Le choix de la valeur de K dépend généralement de l'application à traiter. Pour notre cas, nous avons choisi K=5.

Concernant le calcul de la distance, plusieurs types de distances peuvent être également choisis. Dans notre cas, nous avons opté pour la distance de Canberra.

4. Expérimentations et résultats

4.1. *Présentation de la base utilisée*

Afin d'évaluer notre méthode proposée, nous avons choisi d'utiliser une base de 800 blocs de texte.

Afin d'utiliser cette base, les 800 blocs de la base ont été divisés en deux parties égales : 400 pour l'apprentissage et 400 pour le test (100 pour l'apprentissage et 100 pour le test pour chacune des quatre classes à identifier).

4.2. *Résultats d'identification*

Nous présentons les différents résultats d'identification obtenus pour les différents outils utilisés (tableau 2, 3 et 4).

Script	% Ident.	% Conf.	Matrice de confusions			
			arabe imprimé	latin imprimé	arabe manuscrit	latin manuscrit
arabe imprimé	69 %	31 %	69	1	16	14
latin imprimé	60 %	40 %	18	60	0	22
arabe manuscrit	91 %	9 %	0	0	91	9
latin manuscrit	81 %	19 %	1	0	18	81
Moyenne	75.25%	24.75 %	-	-	-	-

Tableau 2. *Résultats d'identification avec les filtres de Gabor*

Script	% Ident.	% Conf.	Matrice de confusions			
			arabe imprimé	latin imprimé	arabe manuscrit	latin manuscrit
arabe imprimé	86 %	14 %	86	4	8	2
latin imprimé	95 %	5 %	3	95	1	1
arabe manuscrit	83 %	17 %	0	7	83	10
latin manuscrit	75 %	25 %	0	10	15	75
Moyenne	84.75%	15.25 %	-	-	-	-

Tableau 3. *Résultats d'identification avec les matrices de cooccurrence*

Script	% Ident.	% Conf.	Matrice de confusions			
			arabe imprimé	latin imprimé	arabe manuscrit	latin manuscrit
arabe imprimé	71 %	29 %	71	1	17	11
latin imprimé	80 %	20 %	14	80	4	2
arabe manuscrit	88 %	12 %	3	1	88	8
latin manuscrit	90 %	10 %	0	0	10	90
Moyenne	82.25%	17.75 %	-	-	-	-

Tableau 4. *Résultats d'identification avec les ondelettes*

L'analyse des résultats présentés dans les tableaux ci-dessus montre que les taux d'identification dépendent de l'outil utilisé. Cette dépendance par rapport à l'outil concerne le taux d'identification global ainsi que les taux d'identification de chacune des quatre classes. Dans notre cas, le meilleur taux global de bonne identification est obtenu à l'aide des matrices de cooccurrence de niveau de gris avec un taux de 84.75%. Ce taux peut être considéré comme un taux satisfaisant en tenant compte de la simplicité de la méthode qui ne nécessite pas de segmentation préliminaire comme dans le cas d'autres approches de différenciation, et de la complexité du problème traité à cause de l'aspect cursif de l'arabe (imprimé ou manuscrit) et du latin manuscrit.

5. Conclusion et perspectives

Le travail présenté s'inscrit dans le cadre du problème de la différenciation d'écritures. Notre objectif consistait à créer un système permettant la différenciation de l'arabe et du latin de natures imprimée et manuscrite en se basant sur l'approche globale.

Le meilleurs taux d'identification obtenu à l'aide de notre méthode est de 84.75%. Ce taux a été obtenu en utilisant les matrices de cooccurrence de niveaux de gris comme outils d'extraction de caractéristiques.

La simplicité de la méthode et la complexité du problème à résoudre permettent de considérer le taux obtenu comme un taux satisfaisant.

Cependant, la comparaison de ce taux avec les taux obtenus dans le cadre d'autres méthodes qui ont adopté l'approche globale pour différencier d'autres classes d'écritures (100% pour (Manthalkar et Biswas, 2002) et 99% pour (Busch *et al.*, 2005)), met en évidence la nécessité d'ajouter certaines améliorations à notre système afin d'obtenir de meilleurs résultats.

Ces améliorations peuvent affecter le système en conservant l'approche globale. Parmi les améliorations proposées dans ce cas, nous citons :

– la possibilité d'utiliser une base de blocs de texte de taille plus grande. En effet, la taille de la base d'apprentissage a un effet considérable sur les résultats d'identification,

– la possibilité d'utiliser d'autres d'outils pour l'extraction de caractéristiques tels que, par exemple, les filtres de Log-Gabor (Joshi et al., 2006).

D'autre part, les améliorations peuvent faire intervenir d'autres approches et d'autres modèles de systèmes. Parmi ces améliorations, nous citons :

– la possibilité d'adopter une approche mixte (Chaudhury et Sheth, 1999 ; Bennasri *et al*., 2000 ; Pal et Chaudhuri, 2001 ; Kanoun *et al.,* 2002). Dans ce cas, quelques caractéristiques locales doivent être ajoutées au système proposé,

– la possibilité de mettre en œuvre un système de différenciation d'écritures à plusieurs niveaux de décision (Ben Jlaiel *et al.,* 2007). L'approche globale peut être exploitée, dans ce cas, dans certains niveaux d'un tel système.

6. Bibliographie

Ben Jlaiel M., Kanoun S., Alimi A.M., Mullot R., « Three decision levels strategy for Arabic and Latin texts differentiation in printed and handwritten natures », *In Proceedings of 2007 International Conference on Document Analysis and Recognition ICDAR'07*, vol. 2, pp. 1103-1107.

Bennasri A., Zahour A., Taconet B., « Arabic Script Preprocessing and Application to Postal Addresses », *Proc. of ACIDCA'2000*, March 22-24, Monastir, Tunisia, pp. 74-79, 2000.

Busch A., Boles W.W., Sridharan S., « Texture for Script Identification », *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI'05*, vol. 27, pp. 1720-1732.

Chaudhury S., Sheth R., « Trainable Script Identification Strategies for Indian Languages », *In Proceedings of 1999 International Conference on Document Analysis and Recognition ICDAR'99*, pp. 657-660, 1999.

Elgammal A. M., Ismail M.A., « Techniques for Language Identification for Hybrid Arabic-English Document Images », *In Proceedings of 2001 International Conference on Document Analysis and Recognition ICDAR'01*, pp. 1100-1104, 2001.

Fan K., Wang L., Tu Y., « Classification of machine-printed and handwritten texts using character block layout variance », *In International Journal of Pattern Recognition IJPR*, Volume 31, Number 9, pp. 1275-1284, 1998.

Hochberg J., Cannon M., Kelly P., White J., « Page Segmentation Using Script Identification Vectors: A First Look », *In Proceedings of the 1997 Symposium on Document Image Understanding Technology SDIUT'97*, pp. 258-264, 1997.

Joshi G.D., Garg S., Sivaswamy J., « Script Identification from Indian Documents », *In Proceedings of IAPR Workshop on Document Analysis System DAS'06*, vol.3872, pp. 255-267.

Kanoun S., Ennaji A., Alimi A.M., Lecourtier Y., « Script and Nature Differentiation For Arabic and Latin Text Images », *Eight IAPR International Workshop on Frontiers in Handwriting Recognition IWFHR'2002*, 6-8 August, 2002, pp. 309- 313, Niagara-on-the-Lake, Ontario, Canada.

Manthalkar R., Biswas P.K., « An Automatic script identification scheme for Indian Languages », *NCC'02*, pp. 31-34

Pal U., Chaudhuri B.B., « Automatic Identification of English, Chinese, Arabic, Devnagari and Bangla Script Line », *In Proceedings of 2001 International Conference on Document Analysis and Recognition ICDAR'01*, pp. 790-794, 2001.

Pan W.M., Suen C.Y., Bui T.D., « Script Identification Using Steerable Gabor Filters », *In Proceedings of 2005 International Conference on Document Analysis and Recognition ICDAR'05*, pp. 883-887

Peake G.S., Tan T.N, « Script and Language Identification from Document Images », *In Proceedings of Eighth British Machine Vision Conference BMVC'97*, Volume 2, pp. 230-233, septembre 1997.

Singhal V., Navin N., Ghosh D., « Script-based Classification of Hand-written Text Documents in a Multilingual Environment », *RIDE- MLIM'03*, pp. 47-54.

Spitz A. L., « Determination of the the Script and Language Content of Document Images », *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, Volume 19, Number 3, pp. 235-245, 1997.

Tan T.N., « Rotation Invariant Texture Features and Their Use in Automatic Script Identification », *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, Volume 20, Number 7, pp. 751-756, 1998.

Tao Y., Tang Y.Y, « Discrimination of Oriental and Euramerican Scripts Using Fractal Feature », *In Proceedings of 2001 International Conference on Document Analysis and Recognition ICDAR'01*, pp. 1115-1119, 2001.

Wood S. L., Yao X., Krishnamurthi K., Dang L., « Language identification for Printed Text Independent of Segmentation », *In Proceedings of 1995 IEEE International Conference on Image Processing ICIP'95*, pp. 428-431, 1995.