Contribution à la Reconnaissance Automatique des Documents d'Entreprises

Djamel GACEB *Frank LEBOURGEOIS* *Véronique EGLIN* Hubert EMPTOZ
LIRIS UMR 5205CNRS, INSA de Lyon 69621 Villeurbanne Cedex

djamel.gaceb1@insa-lyon.fr flebourg@rfv.insa-lyon.fr veronique.eglin@insa-lyon.fr hubert.emptoz@liris.cnrs.fr

Résumé

Le traitement automatique de documents et courrier d'entreprises est un domaine exigeant en terme de performances et de vitesse. Les systèmes actuels utilisent des architectures modulaires dans lesquelles chaque étape du processus de reconnaissance est indépendante. Pour augmenter les performances, il est nécessaire de réintroduire une coopération entre les différents modules. Dans ce cadre, nous proposons une approche hybride de localisation des zones de textes et de binarisation des images. Ce couplage a permis à la fois de gagner en temps de calcul en évitant de traiter l'arrière plan de l'image et d'obtenir une meilleure segmentation en caractères pour l'OCR. Nous présenterons les résultats obtenus à partir de l'implémentation de notre nouvelle approche sur une ligne industrielle qui traite quotidiennement plusieurs tonnes de courrier et documents internes de grandes entreprises.

Mots clefs

Localisation de textes, segmentation des images, courrier d'entreprises.

1 Introduction

Le domaine du traitement automatique du courrier d'entreprises possède en générale plusieurs contraintes :

- Très grande variété de documents (texte manuscrit ou imprimé, qualité, couleur et texture de papier différentes)
- Contraintes de temps réel (temps de traitement limité)
- Adaptation au mode de capture par système de caméra linéaire (on devra développer les outils d'analyse d'image à la particularité de cette prise d'image pour optimiser les temps de calcul)
- Une obligation de résultats (Le système doit être le plus performant possible pour éviter les coûteuses interventions manuelles).

On retrouve aussi d'autres contraintes particulièrement liées à l'application industrielle qui nous concerne:

- Les images à traiter sont réparties en catégories correspondantes aux familles de courriers des clients d'entreprise : Courrier interne manuscrit (CIM), Courrier interne dactylographique (CID), formulaire (FRM), planus (PL), carte bleue (CB), listing A3(LA3), listing A4(LA4), NPAI, chèque circulant (CHC). Ces images sont très différentes du point de vue de leur taille, de leur orientation, des couleurs du fond et du texte, de la position de texte dans l'image, de la taille des caractères et des types d'écritures (imprimés, imprimés matriciels, manuscrits…). Les documents sont traités par lots ou bien arrivent en vrac.
- Temps de traitement limité, pour l'acquisition de l'image, sa binarisation, la localisation des zones de textes.
- La résolution actuelle de la caméra CCD utilisée est d'environ 200dpi (10 pouces/2048 pixels) et ne peut prendre qu'une seule image par document.
- Les documents non reconnus sont immédiatement traités manuellement. L'échec de reconnaissance s'explique généralement par un dysfonctionnement des étages de prétraitements et en particulier des étages de segmentation et de localisation [1][2].

2 Comparaison des méthodes existantes

2.1 Architectures logicielles linéaires et approches coopératives

Les limites atteintes par les systèmes de vision actuels sont dues à l'organisation linéaire du traitement de l'information. Le taux de rejet et le taux d'erreur des systèmes industriels sont élevés à cause de l'indépendance des processus engagés dans la reconnaissance.

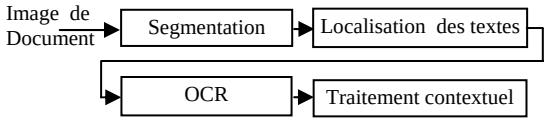

Figure 1 – *Architecture linéaire de systèmes de vision*

Cette séparation des processus est adaptée à la répartition des tâches sur plusieurs ordinateurs connectés, mais l'échec d'une seule étape du processus conduit irrémédiablement le système à rejeter ou bien à commettre une erreur d'interprétation. Certains travaux font déjà référence à des architectures plus avancées. [3] propose un système multi-agents pour l'échange des données et la collaboration entre les différents modules d'acquisition et de reconnaissance. [4] décrit une architecture

collaborative des différents modules pour reconnaître les adresses et les codes postaux. [5] décrit une approche probabiliste pour combiner la localisation, la segmentation et la reconnaissance. Enfin, [6] décrit une segmentation en mots dirigée par une étape de reconnaissance.
C'est de dans ce contexte que s'inscrivent nos travaux pour réduire les taux de rejet et les erreurs du système de vision existant en introduisant une meilleure coopération entre les différentes étapes de la reconnaissance tout en restant dans les limites d'un processus de temps réel. Nous allons donc étudier une architecture non linéaire du processus de reconnaissance en introduisant des bouclages d'informations possibles entre les différents étages (classification des documents, localisation des zones d'intérêts, localisation des zones de texte, segmentation, OCR, reconnaissance de la structure du document et classification du type de document…). Parmi les couplages possibles, nous proposons dans cet article de commencer par une coopération entre la segmentation et la localisation des zones textuelles. Cette coopération devrait nous permettre à la fois d'économiser le temps de traitements et d'améliorer la qualité de la segmentation.

2.2 Comparaison des méthodes de binarisation des documents

La numérisation des documents et courriers avec une caméra CCD, donne des images en niveaux de gris. La réduction de la quantité d'informations à analyser pour l'OCR, nécessite souvent une étape préliminaire de binarisation. La binarisation est le passage irréversible d'une image en niveaux de gris qui permet une classification entre le fond (image du support papier) en blanc et la forme (traits, graphique, caractères) en noir. Le mauvais choix du seuil, peut détruire une grande part d'information contenue dans l'image de l'enveloppe. En effet, une bonne binarisation doit être capable de conserver à la fois tous les caractères et les gravures sans récupérer trop de bruit.

On peut trouver dans la littérature de très nombreux travaux concernant la binarisation de documents. Les plus simples et les plus rapides utilisent l'histogramme de l'image comme les célèbres méthode d'Otsu [7], de Fisher [8] ou d'entropie [9] pour déterminer un seuil qui leur sera appliqué. Ces méthodes globales ont l'avantage d'être extrêmement rapide mais la variation d'éclairage sur le document fait chuter la qualité de la binarisation (Figure2).

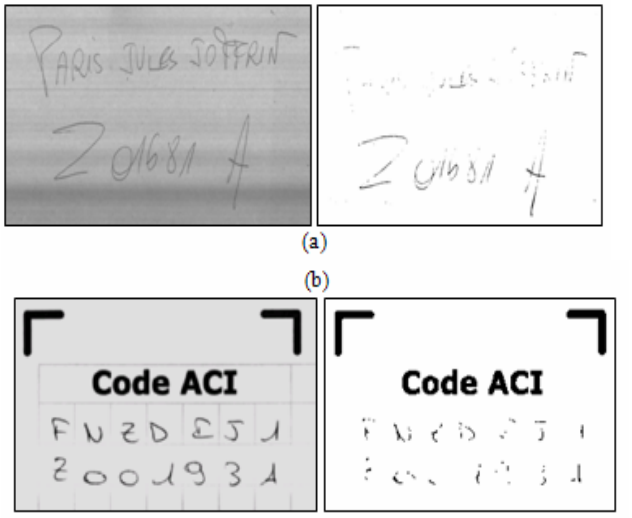

Figure 2 - *Binarisation, (a) par la méthode de Fisher, (b) par la méthode d'Otsu.*

D'autres, telles que les méthodes introduites par Niblack [10], Sauvola[11], Wolf [12] utilisent une approche locale aux pixels pour déterminer une valeur de seuil, pour chaque pixel de l'image, en analysant son voisinage. Leur adaptation locale aux changements de contraste explique l'efficacité de ces méthodes sur les images de manuscrits ou encore sur les documents qui utilisent des couleurs d'encre différentes. Cette approche permet d'obtenir un résultat faiblement dépendant des variations de luminosité sur la page (Figure 3).

Figure 3 - *Binarisation adaptative par la méthode de Sauvola, (a) Image CHC, fenêtre 7×7, (b) Image CIM, fenêtre 9×9, (c) Image 'Insert' fenêtre 15×15.*

Malgré leur efficacité, les méthodes locales possèdent les inconvénients suivants :
- temps de calculs prohibitifs en fonction la taille de la fenêtre d'analyse

- sur-segmentation des défauts et de la texture de papier sur l'arrière plan de l'image
- Traitement difficile des documents dont la taille des caractères est très variable, la fenêtre d'analyse étant fixe durant tout le traitement.

Le tableau suivant contient les temps moyens de traitement, calculés sur une base de 9341 images.

Méthodes de binarisation Type de documents	Niblack	Sauvola	Wolf	Fisher
CHC	4,44	4,47	4,38	0,32
NPAI	3,75	2,28	2,30	0,29
CB	4,34	3,46	4,30	1,74
LA3	4,38	4,46	4,32	0,31
LA4	2,42	2,43	2,42	0,23
FRM	3,59	3,50	4,33	0,25
PL	4,53	4,53	4,45	1,69
CIM	4,50	4,61	4,39	0,36
CID	1,30	1,28	1,29	0,15

Tableau 1 - Vitesses d'exécution des méthodes classiques de binarisation (en secondes).

Après cette analyse, on peut conclure qu'aucune des méthodes classiques ne remplit toutes les conditions imposées (efficacité et rapidité).

2.3 La localisation des zones de texte

On regroupe les travaux sur la localisation des blocs adresses en plusieurs classes :
- Les méthodes basées sur la multi-résolution
- Les méthodes agrégatives par filtrage
- Les méthodes ascendantes basées sur l'adjacence des composantes connexes
- Les méthodes basées sur la segmentation d'images
- Les méthodes basées sur l'apprentissage

Les contraintes de temps réel et la grande variabilité des tailles des caractères et d'espaces entre les mots ont amené plusieurs chercheurs à utiliser la multi-résolution. La localisation du bloc adresse par multi-résolution ne nécessite pas une binarisation préalable de l'image et s'appuie sur une construction pyramidale permettant de faire apparaître à un niveau de résolution approprié la structure d'un bloc de lignes [13]. L'approche pyramidale permet aussi une analyse de type descendante pour construire un arbre d'inclusion de composantes connexes segmentées aux différents niveaux de résolution [14]. D'autres travaux utilisent les méthodes agrégatives classiques de type RLSA [15] qui sont rapides car elles ne nécessitent pas un calcul coûteux de capture des connexités. Cependant ces méthodes sont sensibles à l'inclinaison des documents et nécessitent une bonne orientation et un parfait alignement des lignes de texte. Ces approches agrégatives, ne sont pas nouvelles au regard des premiers travaux sur la localisation des textes dans les images par G. Nagy pour extraire la structure physique du courrier et des formulaires au service des grandes entreprises [Nagy 68]. Les ordinateurs de l'époque n'ayant pas la puissance de calcul nécessaire pour des algorithmes évolués, il eut l'idée d'utiliser la défocalisation progressive de l'optique de la caméra pour rendre l'image progressivement floue dans laquelle les caractères deviennent des «taches» qui s'agglomèrent progressivement entre elles pour désigner les mots, les lignes et les blocs de texte.

Les travaux sur la localisation des blocs adresses basés sur le regroupement des composantes connexes sont nombreux et ne sont pas adaptés aux contraintes de temps réel. En effet la capture de toutes les connexités et la binarisation « aveugle » de toute l'image est trop coûteuse en temps de calculs [18]. De plus ces méthodes nécessitent une classification complexe des connexités en fonction de leur alignement et un rejet des connexités qui ne correspondent pas à des éléments textuels [16][17]. Enfin ces méthodes nécessitent une binarisation préalable de l'image.

Les méthodes basées sur la segmentation de l'image avec des méthodes classiques comme le Split&merge [19] permettent de localiser rapidement les régions non uniformes de l'image susceptibles de contenir du texte. D'autres méthodes de segmentation utilisent aussi les informations sur la texture avec des filtres de Gabor [21][20] ou les ondelettes [23]. Ces méthodes localisent à la fois les zones pertinentes de l'image sans capturer les connexités, mais elles différencient les zones de texte des éléments non textuels à partir de leurs textures. Cependant ces approches intéressantes sont néanmoins très coûteuses en temps de calculs.

Les systèmes de localisation par apprentissage [22] [24][25] nous paraissent difficiles à mettre en œuvre devant la grande variété des documents que nous avons a traiter. De plus certains chercheurs admettent que les systèmes à apprentissage sont moins performants que les systèmes dont les règles ont été ajustés manuellement au problème posé [26].

3 Notre proposition

3.1 Couplage binarisation/localisation

La séparation entre l'étape de la binarisation et celle de localisation des textes, augmente à la fois le temps de calcul et conduit à une sur-segmentations du bruit et de la texture de papier sur des zones vides de l'image. Nous avons pu optimiser notre méthode de binarisation en appliquant les calculs de seuils adaptés uniquement à

proximité des zones de texte. Pour cela nous détectons très rapidement des zones de textes afin d'y appliquer une méthode de segmentation locale de type Sauvola. Nous évitons ainsi de binariser les zones vides qui représentent la plus grande partie de l'image. Cette approche nous permettra aussi de corriger le défaut de sur-segmentation des méthodes adaptatives sur les zones non textuelles de l'image.

3.2 Application à la localisation des zones textuelles

La localisation doit s'effectuer directement sur l'image en niveaux de gris issue de la caméra. La méthode développée doit aussi réduire le plus possible le nombre de fausses détections et ajuster les zones au voisinage du texte. Nous avons utilisé une méthode robuste qui permet de localiser rapidement toutes les zones textuelles dans une scène naturelle sans éclairage particulier ni contrainte lors de la prise d'image. Ce procédé consiste à agglomérer certaines périodicités caractéristiques des lignes de texte qui proviennent des variations lumineuses sur des contours des caractères ou générées par les alternances entre les traits ou entre les caractères. Ces périodicités sont calculées à partir de séquences de pixels à gradients élevés. Pour éviter de filtrer ces points et d'introduire de nouveaux seuils, on effectue localement, dans un voisinage V en chaque point (x_0,y_0), une simple sommation des normes des gradients normalisée par le nombre N de pixels du voisinage $V(x_0,y_0)(1)$.

$$G(x_0, y_0) = \frac{1}{N} \sum_{(x,y) \in V(x_0, y_0)} \frac{\partial f(x,y)}{\partial \vec{v}} \qquad (1)$$

Ce filtre de « gradients cumulés », initialement développé pour la localisation de textes dans les images vidéos [27], a été utilisé pour la localisation des titres dans les vidéos non contraintes comme les archives télévisuelles [12] et la segmentation de l'imprimé composite couleur [28]. Cette méthode des « gradients cumulés » possède plusieurs inconvénients pour notre application. Nous proposons de l'adapter à notre environnement :

Le filtre suppose que la direction de l'enveloppe est a priori connue. En effet, les dérivées sont calculées dans la direction supposée du texte et sommés dans cette même direction. Pour rendre le filtrage insensible à la rotation de l'image du document, nous allons calculer les dérivées horizontales et verticales et les sommer dans les deux directions (2). Nous utiliserons une approximation grossière mais rapide pour le calcul des dérivées (3). Le coût du calcul de la sommation en chaque point de l'image dans un voisinage V est trop élevé pour notre application. Nous allons réduire ce coût de calcul en effectuant la sommation par blocs en multi-résolution. Nous divisons l'image en blocs rectangulaires de taille dx×dy puis nous calculons dans chaque bloc la somme des gradients verticaux et horizontaux (Figure 4).

$$J(x_0, y_0) = \frac{1}{dx\ dy} \sum_{i=1}^{dy} \sum_{j=1}^{dx} \left| \frac{\partial I(x_0 + j, y_0 + i)}{\partial x} \right| + \left| \frac{\partial I(x_0 + j, y_0 + i)}{\partial y} \right| \qquad (2)$$

$$\frac{\partial I}{\partial x}(u,v) = I(u-2,v) - I(u+2,v)$$
$$\frac{\partial I}{\partial y}(u,v) = I(u,v-2) - I(u,v+2) \qquad (3)$$

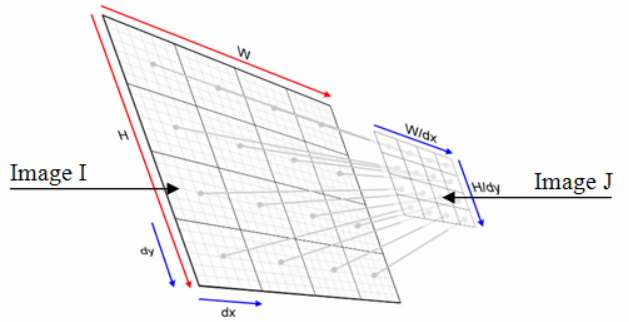

Figure 4 - *Réduction de la taille d'image avec traitement sur le voisinage*

Figure 5 - *Etapes de la binarisation hybride*

Nous obtenons une image J de taille réduite (Figure4, Figure5) dans laquelle les zones claires représentent les zones textuelles dans l'image originale. Cette sommation par bloc ne donne pas les mêmes résultats qu'une sommation en chaque pixel. Nous devons effectuer un prétraitement morphologique sur l'image J pour obtenir un filtrage équivalent à celui de l'algorithme original. Sur cette image J, nous appliquons consécutivement d1 fois l'opérateur dilatation, e1 fois érosions, d2 fois dilatations et e2 fois érosions (4).

$$K = Ee2(Dd2(Ee1(Dd1(J)))) \qquad (4)$$

Le but d'appliquer ces transformations morphologiques est d'une part, de re-densifier le texte et donc de l'agglomérer en blocs, et d'une autre part de prendre une marge suffisante au tour de trait afin d'inclure

l'information pertinente de l'arrière plan (la texture et la couleur) pour un meilleure seuillage. Les paramètres d1, e1, d2, e2 du masque ainsi que la taille de la fenêtre dx×dy ont été fixés pour le moment arbitrairement. On peut constater qu'une augmentation de dx et dy mène à une détection grossière du texte et plus rapide alors que l'augmentation de d1 et e1 détecte mieux les zones de textes agglomérées entre elles. Donc une étude de la stabilité du résultat sur les différents types d'images peut aboutir à un compromis satisfaisant. (e1=d1=2, pour détecter les zones de textes et e1=d1=1, pour détecter les mots). Le surcoût de calcul des opérations morphologiques est négligeable puisqu'il est effectué sur l'image réduite.

3.3 Méthode utilisée pour la binarisation

Nous avons choisi d'utiliser la méthode de Sauvola d'une part pour sa rapidité (table 1) et d'autre part pour ses performances (la méthode Wolf est spécifique aux images vidéo et ne convient pas pour notre application). Le temps économisé nous a permis d'utiliser une grande taille de fenêtre 21x21 pour l'application de l'algorithme de Sauvola ce qui permet d'obtenir de très bons résultas sur les documents imprimés ou manuscrits avec des tailles de caractères très variables.

4 Résultats

Les temps de traitement écoulés sont très proches de ceux écoulés par une binarisation globale et beaucoup moins importants qu'avec les techniques de binarisations adaptatives classiques. Nous avons pu améliorer également les résultats de reconnaissance (TAB2), ce qui a permis à la société d'avoir des résultats de lecture nettement supérieurs à ceux qu'elle avait avant l'utilisation de notre méthode. Tout en sachant que la société utilisait la méthode de binarisation fournie avec le module de l'OCR commercial, les résultats de la reconnaissance sont une moyenne de six jours sur six mois successifs sachant que la société traite moyennement 29225 courriers par jour.

Documents model	Méthode hybride (localisation/ Segmentation)	Amelioration de l' OCR
CCH	0,56	+2%
NPAI	1,19	+26%
BC	1,42	+21
LA3	0,51	+11%
LA4	0,27	+11%
FMR	0,68	+30
PLN	1,12	+20%
HIM	1,64	+76%
TIM	0,23	+16%

Tableau 2 - Vitesses d'exécution de notre algorithme (en secondes) et amélioration de la reconnaissance.

Parallèlement, nous avons pu obtenir de meilleurs résultats sur les enveloppes manuscrites, car c'est sur ce type de document qu'on trouve le plus de variations locales :

- Variation de la taille des caractères : suivant le style d'écriture des gens.
- Variation de l'épaisseur du trait : suivant le stylo, crayon ou fluorescent utilisé.
- Variation de la couleur du trait : suivant la couleur du stylo utilisé.
- Variation du fond due aux différents papiers utilisés pour les enveloppes internes (papier craft, pochette plastique).

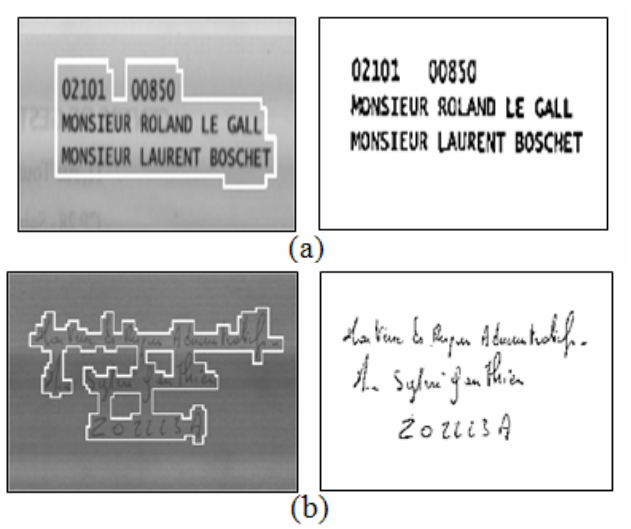

Figure 6 – *Efficacité de notre méthode sur : (a) le texte imprimé, (b) le texte manuscrit (même de faible contraste).*

5 Conclusion et perspectives

Notre méthode nous a permis à la fois de réduire le temps de calcul et d'augmenter la qualité de la binarisation (par une meilleure séparation fond/écriture). Les temps de traitement écoulés sont très proches de ceux écoulés par une binarisation globale et beaucoup moins importants qu'avec des techniques de binarisation adaptatives classiques. Nous avons pu améliorer également les résultats de reconnaissance par l'OCR ce qui a permis à la société d'avoir des résultats de lecture nettement supérieurs à ceux qu'elle avait avant l'utilisation de notre méthode. On peut également, étendre notre combinaison des différentes étapes de reconnaissance pour assurer une

meilleure coopération et interaction entre tous les modules de système de tri.
Ce travail est adopté par la société CESA (www.cesa.fr).

Références

[1] N. Gorski and al, A new A2iA bankcheck recognition system, Handwriting Analysis and Recognition, IEEE Third European Workshop, 1998, pp.1-6.

[2] N. Gorski and al, A2IA check reader, ICDAR'99, pp. 523-526.

[3] U. Miletzki, Documents on the Move, DA&IR-Driven Mail Piece Processing Today and Tomorrow, DAS' 96, pp. 547-563.

[4] S.Srihari, E. Kuebert, Integration of hand-written address interpretation Technology into the United States Postal Service Remote Computer Reader System, ICDAR 97, V.2, pp. 892-896.

[5] Y. Lu and al., An implementation of postal numerals segmentation and recognition system for Chinese business letters, ICDAR99, pp. 725-728.

[6] J. Zhou and al, A feedback-based approach for segmenting handwritten legal amounts on bank cheques, in proc. of ICDAR'01, pp. 887-891.

[7] N. Otsu, A threshold selection method from grey-level histogram, IEEE trans system, man and cybernetics, vol 9, 1979, pp. 62-66.

[8] J. Fisher, S. Hinds, K. D'Amato, A Rule-Based System for Document Image Segmentation, in proc. of the 10th Int'l Conf. Pattern Recognition, 1990, pp. 567-572.

[9] A. Abutaleb, Automatic thresholding of grey-level pictures using two-dimensional entropy, computer vision graphics Image processing, 1985, pp. 22-32.

[10] [Nib86] W. Niblack, An Introduction to Digital Image Processing, Englewood Cliffs, N.J.:Prentice Hall, pp. 115–116, 1986.

[11] J. Sauvola, and al. Adaptive Document Binarization, ICDAR'97, vol 1, pp. 147–152, 1997.

[12] C. Wolf C, J.M. Jolion, F. Chassaing, Text Localization, Enhancement and Binarization in Multimedia Documents, ICPR, 2002, pp. 1037-1040.

[13] O. Deforges, C.Viard-Gaudin, D.Barba, Gray-level Document Image Analysis, 2nd French-Korean Workshop, Man-Machine Handwritten Communication, CNRS Ile de France, 1996, pp. 139-149.

[14] C.Viard-gaudin, D. Barba, Localisation du bloc adresse par une approche multi-résolution, ICDAR91, pp.954-962.

[15] Wahl F, Wong K., Casey G., Block segmentation and text extraction in mixed text/image documents, Computer graphics and image processing, 1982, pp.375-390.

[16] J.C. Oriot, d. Barba, J. Salome, Adress Block Locating Method Based On Transition Analysis Approach: Design And evaluation on flats objects, ICDAR 91, pp.665-673.

[17] J.C. Oriot, D. Barba, M. Gilloux, Localisation du bloc adresse sur les objets postaux par une méthode de segmentation ascendante : évaluation et optimisation, Traitement du Signal, 1995.

[18] B. Yu, A. K. Jain and M. Mohiuddin, Address Block Location on Complex Mail Pieces, in proc. of ICDAR'97, V.2, pp. 897-901.

[19] M. Wolf, H. Niemann, W. Schmidt, Fast Address Block Location on Handwritten and Machine Printed Mail-piece Images, ICDAR 97,V.2, pp. 753-757.

[20] O. Deforges, D. Barba, A fast multiresolution text-line and non text line structures extraction and discrimination scheme for document image analysis, in proc. ICPR 94, pp. 134-137.

[21] A. K. Jain, Y. Chen. Address block location using color and texture analysis, Computer Vision, Graphics and image processing : image understanding, 1994, pp.179-190.

[22] C. Jarousse, C. Viard-Gaudin, Localisation du code postal par réseau de neurones sur bloc adresse manuscrit non contraint, CIFED'98, pp. 72-81.

[23] D. Menoti and al., Segmentation of postal envelopes for address block location : an approach based on feature selection in wavelet space, ICDAR 03, pp. 699-703.

[24] H. Walischewski, Learning regions of interest in postal automation, ICDAR'99, pp. 317-320.

[25] U. Miletzki and al., Continuous learning systems postal address readers with built-in learning capability, ICDAR'99, pp. 329-332.

[26] K. Nitz, An Image-based mail facing and orientation system for enhanced postal automation, ICDAR '03, pp. 694-698.

[27] LeBourgeois F., Robust multifont OCR system from gray level images, fourth ICDAR, International Conference on Document Analysis and Recognition, Ulm, 1997, p. 1-5.

[28] F. LeBourgeoisF. , H. Emptoz H., Document Analysis in Gray Level and Typography Extraction Using Character Pattern Redundancies, Int. Conf. On Doc. Analysis and Recognition ICDAR'99, 1999, India, pp.177-180.