

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique
DER de Génie Electrique et Informatique



THESE

Présentée par

Mr Hamza AKROUM

Ingénieur d'état en électronique de l'E.N.P.

pour l'obtention du

Grade de MAGISTER

EN ELECTRONIQUE

Option : Télécommunications

Sujet :

PRETRAITEMENT ET SEGMENTATION DES IMAGES DE DOCUMENTS

Soutenue publiquement le : 28/06/1999 devant le jury composé de :

M^{lle} M. GUERTI
Mr D. BERKANI
Mme L. HAMAMI
Mr A. BELOUCHIRANI
Mr M. MEHenni
Mr L. SAADAoui

Maître de conférence à l'E.N.P
Professeur à l'E.N.P
Chargée de cours à l'E.N.P
Docteur à l'E.N.P
Maître de conférence à l'E.N.P
Chargé de cours à l'E.N.P

Président
Rapporteur
Rapporteur
Examineur
Examineur
Examineur

ملخص

إن أغلبية الوثائق عموماً توجد على هيئة مطبوعات ورقية. تحويلها إلى معلومات إلكترونية يتم عن طريق أنظمة التحليل والتعرف على الوثائق. المرحلة الأولى في هذه الأنظمة تتمثل بصفة عامة في تحويل الصورة للرقمية للورقة من 256 مستوى رمادي إلى صورة ذات مستويين، وتخفيض نسبة الضوضاء، وتصحيح زلوية ميل الورقة. أما الخطوة الثانية فتتمثل في تجزئ الصورة إلى قطع متجانسة. بعد ذلك تأتي مراحل أخرى يتم من خلالها معالجة القطع المستخرجة على حسب نوعيتها (كتلية، رسم، صورة، الخ...). تفصيل الإجراءات المبينة خلال المرحلتين الأولىين هو المحور الأساسي لهذه الأطروحة.

كلمات المفتاح

المعالجة الوثائقية، تحليل صور الوثائق، تصحيح ميل الورقة، تجزئ الورقة، التحليل التصاعدي، التحليل التبايزي.

Abstract

Most Documents still exist only in printed form. Conversion from paper to electronic media is the role of "Document Image Analysis and recognition (DIAR) Systems". The first step, in DIAR Systems, is pixel-level processing (also called preprocessing). Such processing includes: thresholding to reduce a gray-scale image to a binary image, reduction of noise to reduce extraneous data, and skew elimination. The second step is page segmentation, thus creating blocks or regions of homogenous data. Each zone is processed appropriately by further steps. The first and the second step are the subject of this Master thesis.

Keywords

Document processing, document image analysis, skew estimation, page segmentation, top-down analysis, bottom-up analysis.

Résumé

La majorité des documents n'existent que sur support papier. Les systèmes d'Analyse et de reconnaissance de Document (ARD) ont pour rôle de convertir le papier en un média électronique. La première étape, dans les systèmes ARD, comprend la binarisation, la réduction du bruit et la correction automatique de l'inclinaison. la deuxième étape est la segmentation de la page en zones homogènes. D'autre étapes effectuent un traitement spécifique aux différentes zones identifiées. Nous développons dans notre travail les deux premières étapes.

Mots clés

Traitement de document, analyse de document, estimation de l'inclinaison, segmentation, méthode ascendantes, méthodes descendantes.

TABLE DES MATIERES

INTRODUCTION	1
I. NORMES ET STRUCTURATION DU DOCUMENT	4
I.1. INTRODUCTION	4
I.2. STRUCTURATION DES DOCUMENTS	4
I.2.1 Structure logique	4
I.2.2 Structure physique	4
I.3. NORMES DE REPRESENTATION DE DOCUMENTS	5
I.3.1. La norme ODA	6
I.3.2. La norme SGML	8
I.4. CONCLUSION	9
II. PRETRAITEMENT	10
II.1. INTRODUCTION	10
II.2. BINARISATION	11
II.2.1. Seuillage par sélection itérative	11
II.2.2. Seuillage par corrélation	12
II.2.3. Discussion	15
II.3. REDUCTION DU BRUIT	18
II.3.1. filtrage morphologique	18
II.3.2. filtres spécifiques aux images de document	21
II.4. CORRECTION AUTOMATIQUE DE L'INCLINAISON	23
II.4.1. Méthode des moindres carrés	23
II.4.2. Méthode fondée sur la transformée de Hough	25
II.4.3. Méthode fondée sur la Projection de Profil	32
II.4.4. Discussion	32
II.4.5. Redressement	34
II.5. CONCLUSION	35

III. SEGMENTATION EN BLOCS	37
III.1. INTRODUCTION	37
III.2. STRATEGIE ASCENDANTE	37
III.3. STRATEGIE DESCENDANTE	40
III.3.1. Technique de la Transformée de Hough	40
III.3.2. Technique de suivi de contour	41
III.3.3. Technique de lissage directionnel (RLSA)	42
A. Etape d'agglomération	42
B. Etape de séparation des composantes 8-connexes	46
C. La première partition de l'image	46
D. Segmentation intra-bloc et récursivité	46
E. Applications et résultats obtenus	47
III.2.4. Technique de projection	60
III.3. STRATEGIE MIXTE	63
III.4. DISCUSSION	63
IV. RECONNAISSANCE DES BLOCS	71
IV.1. INTRODUCTION	71
IV.2. METHODE DE WONG, CASEY ET WAHL	71
IV.3. METHODE DE FISHER, HINDS ET D'AMATO	74
IV.4. METHODE DE BELAÏD ET AKINDEL	74
IV.5. METHODE DE P. CHAUVET	76
IV.6. METHODE ADOPTÉE	78
IV.6.1. Remarques Préliminaires	78
IV.6.2. Partition Adoptée	80
IV.6.3. Algorithme	82
IV.7. RESULTATS ET DISCUSSION	82
CONCLUSION	94
BIBLIOGRAPHIE	96