AVIS DE SOUTENANCE

THESE DE DOCTORAT

Présentée par

Mr : YOUSSEF ELFAKIR

Spécialité : Traitement d'image et Informatique

Sujet de la thèse : Indexation des documents manuscrits arabes anciens scannés.

Formation Doctorale : Sciences de l'ingénieur Sciences Physiques, Mathématiques et Informatique.

Thèse présentée et soutenue le mercredi 03 juillet 2019 à 10h à l'Amphi Al Khawarizmi devant le jury composé de :

Nom Prénom	Titre	Etablissement	
Driss CHENNOUNI	PES	Ecole Normale Supérieure de Fès	Président
Mohamed SABBANE	PES	Faculté des Sciences de Meknès	Rapporteur
Hassan QJIDAA	PES	Faculté des Sciences Dhar El Mehraz de Fès	Rapporteur
Ismail ESSAOUDI	PH	Faculté des Sciences de Meknès	Rapporteur
Saad BENNANI DOSSE	PES	Ecole Nationale des Sciences Appliquées de Fès	Examinateur
Khadija LAHRECH	PH	Ecole Nationale des Sciences Appliquées de Fès	Examinateur
Mostapha MRABTI Ghizlane KHAISSIDI	PES PH	Ecole Nationale des Sciences Appliquées de Fès Ecole Nationale des Sciences Appliquées de Fès	Directeurs de thèse

Laboratoire d'accueil : Informatique et Physique Interdisciplinaire.

Etablissement : Ecole Nationale des Sciences Appliquées de Fès

Titre de la thèse : Indexation des documents manuscrits arabes anciens scannés.
Nom du candidat : Youssef ELFAKIR
Spécialité : Traitement d'image et Informatique

Résumé de la thèse

La numérisation des archives et des documents anciens apparaît aujourd'hui comme une nécessité pour en préserver l'intégrité et palier la raréfaction de l'espace. Cependant, la numérisation n'est ainsi que la première étape d'un processus plus large qui consiste à qualifier, classer et indexer les images des documents d'archives pour pouvoir en exploiter toute la richesse informationnelle. Le travail présenté dans cette thèse est consacré à la conception d'un système d'indexation hors-ligne des documents manuscrits Arabes.

Au début, nous proposons une approche pour le prétraitement des documents manuscrits ainsi pour la segmentation de la ligne de texte, où une fenêtre glissante est utilisée pour localiser les régions de documents les plus similaires à la requête. Les histogrammes de gradient orienté (HOG) sont utilisés comme vecteurs caractéristiques pour représenter les images de documents, combinés avec la machine à vecteur de support (SVM) pour classifier les régions pertinentes à la requête. Enfin, l'application de la technique de reclassement est utilisée pour mieux reclasser les résultats obtenus.

Ensuite, nous présentons une approche sans utiliser l'étape de segmentation, car toute erreur affectée sur les représentations de mots engendre une perturbation dans l'étape de correspondance, ceci explique pourquoi les chercheurs dans le domaine de repérage et la recherche de mots se sont orientés vers des méthodes sans segmentation. Basée sur le détecteur de Harris dans l'étape d'extraction des points d'intérêts, la correspondance entre ces caractéristiques est établie avec la technique de corrélation ZNCC suivie d'une phase de relaxation pour éliminer les faux appariements.

Enfin, nous présentons l'impact de traitement d'image dans l'indexation, et son influence sur l'extraction des caractéristiques informatives et discriminatives dans le cas des documents manuscrits. Ainsi, nous représentons chaque point d'intérêt détecté à l'aide de l'algorithme de caractéristiques visuelles invariantes à l'échelle. Pour remédier au problème du stockage en mémoire, nous représentons les régions de l'image par des histogrammes en utilisant un modèle de sac de mots visuels. Le choix de la meilleure taille du codewords (nombre de mots visuels) est dicté par l'analyse de la courbe du fléau de la dimension. Enfin, nous proposons d'utiliser une distance de similarité variable afin de surmonter le problème lié au seuil variable de chaque mot.

Mots–Clés : Manuscrits Arabes, indexation, prétraitement, segmentation, points d'intérêts, sac-de-mots, SVM, HOG, SIFT, Harris, BOW.