الحمد لله

الحمد لله اللهم ربنا لك الحمد بما خلقتنا ورزقتنا وهديتنا وعلمتنا ما لم نكن نعلم وزدتنا من العلم بفضللك

مما أعان على إخراج هذا البحث

لك الحمد بالإيمان ولك الحمد بالإسلام ولك الحمد بالقرآن ولك الحمد بالأهل والمال والمعافاة

ومن كل ما سألناك ربنا أعطيتنا

فلك الحمد على ذلك حمداً كثيراً

لك الحمد حتى ترضى ولك الحمد إذا رضيت ولك الحمد بعد الرضى.

وصلي اللهم وسلم على سيدنا محمد وسلم تسليما كثيرا

الآية

وَإِنَّهُ لَتَنزِيلُ رَبِّ الْعَالَمِينَ (192) نَزَلَ بِهِ الرُّوحُ الْأَمِينُ (193) عَلَىٰ قَلْبِكَ لِتَكُونَ مِنَ الْمُنذِرِينَ (194) بِلِسَانٍ عَرَبِيٍّ مُّبِينٍ (195)

الشعراء: ١٩٢ - ١٩٥

الإهداء

إلى القلبان اللذان فاضا حباً وحناناً ودفئاً
والداي
إلى من تربطني بهم أجمل المشاعر وأصدق العواطف
أخوتي
إلى من كانوا خير عوناً وسنداً لي في طريقي
أصدقائي

شكر وعرفان

لله الحمد والشكر أن وفقنا في اخراج هذا البحث , من هنا لابد من رد شكر من بعد الله لكل من ساهم وشارك في إخراج هذا البحث.
فالشكر أجزله للدكتورمحمد الحافظ - عميد كلية علوم الحاسوب وتقانة المعلومات - الذي لم يبخل علينا بفكره وجهده وخبرته وساهم كثيراً في نجاح هذا البحث.

كما نتقدم بالشكر الجزيل للأستاذة الفاضلة نجود محمد التجاني – أستاذة بكلية علوم الحاسوب وتقانة المعلومات- والتي ساعدتنا في بداية مراحل معالجة الصور الرقمية.

ونشكر أيضاً مجموعة التعرف على الأنماط البحثية وأساتذة كلية علوم الحاسوب وتقانة المعلومات والذين ساهموا فكرياً أو معنوياً في انجاز هذا البحث.

نسأل الله أن يعيننا على مكافأتهم بأحسن مما قدموا لنا وجزاهم الله عنا كل خيـــــــــــر.

المستخلص

تعتمد أنظمة التعرف على الكتابة (OCR) (Optical Character Recognition) التي تستخدم الطريقة التحليلية (Analytical Approach) على تجزئة (Segmentation) الكلمة إلى حروف و من ثم تحاول التعرف على كل حرف بمفرده. و من أكثر الصعوبات التي تواجه أنظمة التعرف على الكتابة العربية هى التجزئة (Segmentation)، حيث تتميز الكتابة العربية باتصال حروفها (Cursive) بصورة معقدة و متعددة الاشكال. ومن هنا جاءت الحوجه إلى وجود جهود بحثية كبيرة للوصول لنظام تجزئة فعّال للغة العربية.

هذه الدراسة تقدم نظام لتجزئة الكلمة العربية المكتوبة بخط اليد إلى مكوناتها الحرفية. يقوم نظام تجزئة الكلمات المقترح بتجزئة الكلمات التي يظهر اتصال حروفها بشكل أفقي وعمودي (Overlapped Characters) مثال الكلمات: "حسن" و "محمد" على الترتيب. ويعتمد نظام تجزئة الكلمات المقترح على استخدام السمات الشكلية (Morphological Features) للحرف العربي. ويعتبر هذا النظام جزء من نظام التعرف على الكلمات العربية المكتوبة بخط اليد باستخدام الطريقة الشمولية (Holistic Approach) بحيث يلجأ نظام التعرف إلى نظام تجزئة الكلمات المقترح في حالة عدم التعرف على الكلمة.

مجموعة البيانات المستخدمة هي عبارة عن مجموعة بيانات جديدة للأسماء العربية المكتوبة بخط اليد وتسمى بمجموعة بيانات جامعة السودان للأسماء (SUST-ARG – names) Arabic Recognition Group-Sudan University of Science and Technology - names. وجُهزت مجموعة البيانات (SUST-ARG – names) من قِبل مجموعة التعرف على الأنماط البحثية بجامعة السودان للعلوم والتكنولوجيا.

أختبر نظام تجزئة الكلمات المقترح على 287 كلمة تشمل كلمات تحتوي على حروف متداخلة (Overlapped Characters) وغير متداخلة وتم التوصل إلى نسبة تجزئة عامه 67.64% و نسبة 64.41% خاصة للكلمات التي تحتوي على حروف متداخلة (Overlapped Characters). ولأن الكتابة كانت بشكل حر بدون أي قيود ظهرت مشكلتان في النظام المقترح

تحتاجان لمواصلة البحث والتطوير وهما: التجزئة الزائدة (Over Segmentation) والتجزئة الناقصة (Under Segmentation).

ABSTRACT

Most Optical Character Recognition (OCR) systems that using the analytical approaches depend on segmenting the handwritten word into characters and to recognizing each character individually. The greatest difficulty facing the recognition of Arabic handwritten systems is segmentation, because Arabic handwritten is cursive with complex multi-form styles. Hence we need huge research efforts to reach an effective segmentation system for the Arabic handwritten words.

This research presents a system which segment the Arabic handwritten word into characters. The proposed system segments the words that appear alphabet in contact with horizontal and vertical (Overlapped Characters) for examples: "حسن" and "محمد" Respectively. The proposed system depends on the use of Morphological features for Arabic character. This system is the part of the Arabic Words Handwriting Recognition System Using Holistic Approach. The proposed system is activated in the event of unrecognizing the word.

Dataset used is a dataset of new names handwritten Arabic called (SUST-ARG-names) Sudan University of Science and Technology - Arabic Recognition Group – names. The processed dataset (SUST-ARG - names) by pattern recognition research group of the Sudan University of Science and Technology.

The proposed system has been tested on 287 word that include some words which contain overlapped characters and the proposed system attains reached to

a general segmentation 67.64% and 64.41% per special words containing overlapping characters. And because the writing was freely without any restrictions have problems in the proposed system need to continue research and development, namely: Over Segmentation and Under Segmentation).

المصطلحات

Adjustment ……..	التعديل في قيم الصورة .....
Analytical approach ……	الطريقة التحليلية للتعرف على الكتابة.....
Arabic Handwriting Recognition …….	التعرف على الكتابة العربية المكتوبة بخط اليد .....
Artificial intelligent ……..	الذكاء الإصطناعي .....
Artificial Neural Network ……	شبكة عصبية اصطناعية .....
Ascenders characters ……	الحروف العلوية .....
Base line ……...	الخط الأساسي للكلمة .....
Binarization ……..	تحويل الصورة إلى الوضع الثنائي "0" و "1".....
Boundaries ……...	الحدود الخارجية لكائن ما في الصورة .....
Branch point ……	نقطة تفرع يتفرع عندها منحنى الكتابة .....
Bridge ……...	عملية وصل النقاط الصورة الغير متصلة .....
Chain code ……	سلسلة الرموز.....
Character Segmentation ……...	التجزئة لحرف.....
Class ……	مجموعة.....
Connected pixels ……..	مجموعة النقاط المتصلة مع بعضها البعض .....
Contour following algorithm ……...	خوارزمية تمر على كل النقاط المكونة للصورة .....
Cursive script ……...	خاصية اتصال حروف اللغة العربية.....
Descenders characters ……..	الحروف السفلية .....
Digital Images Processing ……	معالجة الصور الرقمية .....
Double local minimas ……..	نقاط دنيا متتالية .....
End point ……..	نقطة نهاية ينتهي عندها اتصال الكتابة .....
Features ……	سمات .....
Features Extraction/Selection ……..	استخلاص أواختيار السمات .....
Fill ……	عملية ملء الفراغات الموجودة في الصورة .....
Gray scale ……	المقياس الرمادي للصورة .....
Holistic approach ……	الطريقة الشمولية للتعرف على الكتابة .....
Horizontal Projection ……...	الإسقاط الأفقي.....
Image processing toolbox …..…….	أدوات معالجة الصور .....
Inner contour ……	الخطوط الكنتورية الداخلية للصورة .....
Input time series ……...	متابعة زمن ادخال البيانات .....
Line Segmentation ……...	التجزئة لسطور .....
Local Minima (LM) ……	نقاط دنيا .....
Loop point ……	نقطة حلقية .....
Median filtering ……	إحدى تقنيات إزالة الشوائب .....
Morphological operations ……	عمليات الأشكال .....
Natural Language Processing ……...	معالجة اللغات الطبيعية .....
Neurons ……	خلايا عصبية .....
Noise removal ……..	إزالة الشوائب .....
Off-line character recognition ……..	التعرف على الكتابة بصورة آجلة .....
On-line character recognition ……..	التعرف على الكتابة بصورة آنية .....

Optical Character Recognition (OCR) ….	التعرف على الكتابة ……
Outer contour ……...	الخطوط الكنتورية الخارجية للصورة ……...
Over segmentation ……..……	تجزئة زائدة ……
Overlapped characters ……..	الحروف المتداخلة …….
Pattern Recognition ……..	التعرف على الأنماط …….
Pixels/Image elements ……..……	نقاط الصورة الرقمية …….
Postprocessing ……..	مابعد المعالجة ……..……
Preprocessing ……...	المعالجة الأولية ……
Recognizer ……...	المتعرف ……..……
Segmentation ……	التجزئة ……..……
Segmentation of Arabic word ……..	تجزئة الكلمات العربية……
Shrink ……..	انكماش …….
Skeleton ……	الهيكل الداخلي للكائن ……..……
Spur ……...	الزوائد……
Start point …….	نقطة بداية الكتابة ……..……
Statistical Pattern Recognition ……	الطريقة الإحصائية للتعرف على الأنماط ……..……
Structural (or syntactic) Pattern Recognition …….	الطريقة الهيكلية أو النحوية للتعرف على الأنماط .....
SUST-ARG ……..	مجموعة التعرف على الأنماط بجامعة السودان .....
Thinning ……...	ترقيق الصورة ……..……
Threshold …….	عتبة ……..……
Touching characters …….	حروف ملتصقة مع بعضها ……..……
Touching connected components …….	أجزاء كلمات ملتصقة مع بعضها .....
Turning points ……..	نقاط التفاف ……..……
Under segmentation …….	تجزئة ناقصة ……
Vertical Projection ……...	الإسقاط العامودي ……..……
Word Segmentation ……..	التجزئة لكلمة ……

فهرس الأشكال

رقم الشكل	موضوع الشكل	رقم الصفحة
1.1	مثال لتجزئة كلمة مكتوبة بخط اليد إلى مكوناتها الحرفية	3
1.2	مراحل نظام التعرف على الأنماط (Pattern Recognition)	6
2.2	بعض الحروف العربية التي يتم تمييزها بعدد وموضع النقاط	10
1.3	اجراء عملية الترقيق (Thinning) للصورة	13
2.3	الصورة بعد عملية ازالة الزوائد (Spur)	13
3.3	الحصول على الشكل الحلقي لصورة	14
4.3	ايجاد مجموعة النقاط المتصلة مع بعضها البعض (Connected Pixels)	14
5.3	حساب الإسقاط العمودي (Vertical Projection) لصورة	15
6.3	حساب الإسقاط الأفقي (Horizontal Projection) لصورة	15
7.3	النقاط الثمانية المجاورة لنقطة معينة	16
8.3	مثال يوضح حساب سلسلة الرموز باستخدام الاتجاهات الثمانية للنقطة	16
9.3	النقاط الأربعة المجاورة لنقطة معينة بالإضافة إلى توضيح للإتجاهات الأربعة لأي نقطة وقيمة الاتجاه	17
1.4	مراحل نظام التعرف على الكتابة	19
2.4	طريقة التجزئه لسطور (Line Segmentation)	20
3.4	تجزئة الكلمات في السطر الواحد (Word Segmentation)	20
4.4	طريقة الإسقاط العمودي(Vertical Projection) في تجزئة الكلمة إلى مكوناتها الحرفية	21
5.4	الخطوط الكنتورية الخارجية (Outer Contour) والداخلية (Inner Contour) للصورة	22
6.4	كتابة الاسم على خط أفقي يسمى بالخط الأساسي (Base Line)	22
7.4	مناطق الصورة الثلاثة والخط الأساسي (Base Line)	23
8.4	نقاط التجزئة المقبولة والمرفوضة في أجزاء كلمات مختلفة	25
9.4	مثال لحروف متداخلة (Overlapped Characters)	25
10.4	مثال لحروف ملتصقة (Touching Characters) وكلمات ملتصقة (Touching Connected)	25
1.5	الهيكل العام لنظام تجزئة الكلمات المقترح	28
2.5	مراحل المعالجة الأولية (Preprocessing)	29
3.5	الحصول على الأجزاء المتصلة وهي تشمل أجزاء الكلمة ونقاط الكلمة والهمزات	30
4.5	ضم النقاط لجزء الكلمة الأقرب لها	31
5.5	أجزاء كلمة "حامد"	31
6.5	الأجزاء المتداخلة في كلمة "مزمل"	31
7.5	الأجزاء منفصلة في كلمة "مزمل"	31
8.5	مثال يوضح السمات الأساسية لأي نقطة (Pixel) مكونة لكلمة "يوسف"	32
9.5	بعض النماذج الخاصة بنقطة النهاية (End Point)	34
10.5	بعض النماذج الخاصة بنقطة التفرّع (Branch Point)	34
11.5	حساب قيمة الإسقاط الأفقي (Horizontal Projection) للصورة وايجاد موقع الخط الأساسي (Base Line) لها	35
12.5	مدخلات ومخرجات خوارزمية التجزئة	36

13.5	أمثلة لبعض الكلمات والتي ظهر فيها الحرفان "حـ ، جـ " متداخلان مع تحديد لشكل الحرف عند التداخل	37
14.5	أمثلة لبعض الكلمات والتي ظهر فيها الحرف "مـ " متداخلاً مع تحديد لشكل الحرف عند التداخل	37
15.5	تحديد النقطة المقبولة (Accept Point) للكلمات التي تظهر فيها الحروف "خـ،، حـ ، جـ " متداخلة	38
16.5	تحديد النقاط المقبولة (Accept Points) للكلمات التي يظهر فيها الحرف "مـ " متداخلاً	38
1.6	أمثلة لبعض العينات من كلمة "محمد" والتي لم تُجزأ كل حروفها (Under Segmentation)	45
2.6	أمثلة لبعض الكلمات جُزأت حروفها لأكثر من جزء (Over Segmentation)	45

فهرس الجداول

رقم الجدول	موضوع الجدول	رقم الصفحة
1.2	مثال لبعض أشكال الحروف المختلفة في أكثر من موضع في الكلمة	10
1.6	نتائج أداء الخوارزمية المقترحة	41
2.6	نتائج أداء الخوارزمية المقترحة بالنسبة للكلمات التي تحتوي على حروف متداخلة (Overlapped Characters)	41
3.6	نتائج أداء الخوارزمية المقترحة بالنسبة للكلمات التي تحتوي على حروف غير متداخلة (Non-overlapped Characters)	42
4.6	بعض من الكلمات التي تحتوي على حروف غير متداخلة (Non-overlapped Characters) تم اختبارها على الخوارزمية المقترحة	43
5.6	بعض من الكلمات التي تحتوي على حروف متداخلة (Overlapped Characters) تم اختبارها على الخوارزمية المقترحة	44
6.6	نتائج أداء الخوارزمية المقترحة على الكلمات التي تحتوي على حروف متداخلة (Overlapped Characters) بالنسبة للكلمات "أحمد" و "محمد" و "عمر"	45

فهرس المحتويات

الباب - الفصل	الموضوع	الصفحة
الباب الأول	**المقدمة**	
1.1	مقدمة	2
2.1	مشكلة البحث وأهدافه	3
3.1	أهمية البحث	3
4.1	فروض البحث	4
5.1	أبواب البحث	4
الباب الثاني	**التعرف على الكتابة**	
1.2	التعرف على الأنماط	6
1.1.2	مراحل التعرف	6
2.1.2	تطبيقات التعرف على الأنماط	7
2.2	التعرف على الكتابة	8
1.2.2	التعرف على الكتابة العربية المكتوبة بخط اليد	9
3.2	خصائص الكتابة العربية	9
الباب الثالث	**معالجة الصور الرقمية**	
1.3	مقدمة	12
2.3	مكونات الصورة الرقمية	12
3.3	عمليات الأشكال	12
4.3	الإسقاط الأفقي والعمودي	14
1.4.3	طريقة حساب الإسقاط العمودي	14
2.4.3	طريقة حساب الإسقاط الأفقي	15
5.3	الطرق المستخدمة لوصف كائن معين من الصورة	15
1.5.3	سلسلة الرموز	16
الباب الرابع	**تجزئة الكلمات العربية**	
1.4	مقدمة	19
2.4	مراحل التجزئة	20
1.2.4	التجزئة لسطور	20
2.2.4	التجزئة لكلمات	20
3.2.4	التجزئة لحروف	21
3.4	بعض الطرق المستخدمة في تجزئة الكلمة العربية لحروف	21
الباب الخامس	**نظام تجزئة الكلمات المقترح**	
1.5	مقدمة	27
2.5	المعالجة الأولية	28
3.5	تقسيم الكلمة إلى أجزاء	30
4.5	استخلاص السمات لكل جزء من أجزاء الكلمة	31
5.5	تجزئة أجزاء الكلمة إلى حروف	36
1.5.5	خوارزمية تجزئة الحروف المتداخلة	37
الباب السادس	**التجارب والنتائج**	
1.6	مجموعة البيانات المستخدمة	40
2.6	التجارب	40
3.6	نتائج الدراسة	41

الباب السابع **الخلاصة والمهام المستقبلية**
1.7 خلاصة البحث 47
2.7 المهام المستقبلية 47

المراجع 50