Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Technology
Department of Computer Science

Hand Written Text Recognition for Security Application

A Thesis
Submitted to the Department of Computer Science of the University of Technology in a Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computer Science

BY
Mustafa Salam Kadhm AL-Shammari

Supervisor
Asst. Prof. Dr. Alia K. Abdul Hassan

2016

Abstract

Most of the governments and organizations have a huge number of handwritten documents generated by their daily processes. It is imperative to use computers to read the generated handwritten texts, and make them editable and searchable. Therefore, handwritten recognition lately became a very popular research topic and the number of its possible applications is very large. It's capable in resolving complex problems and simplify human activities by converting the handwritten documents into digital form and making them suitable for many applications especially for the security one through creating an authenticated recognition. However, the Arabic handwritten text recognition is a complex process compared with other handwritten languages because Arabic handwritten text is of cursive nature. Thus, the task of authenticated Arabic handwritten text documents recognition to be close to human performance is still an open problem.

An accurate and authenticated Arabic handwritten text recognition system based on segmenting the input handwritten text documents into handwritten sub-words is proposed. The system has two main modules that are used, for the recognition of the handwritten text and identifying the document's writer. The first module1 has six stages that work together to recognize the Arabic handwritten text and convert it into editable text. These stages are: image acquisition, segmentation, preprocessing, features base construction, classification and post-processing. The second module2 is identified the desired document's writers through several stages which are: image acquisition, segmentation, preprocessing, features extraction, classification and post-processing. The system proposes an efficient and accurate segmentation algorithm that segments the input handwritten text into a number of handwritten sub-images and each of these segmented sub-images has an Arabic handwritten sub-word. Besides that, an image thresholding algorithm is proposed to convert the sub-images into binary based on using fuzzy c-mean clustering method. Furthermore, the binary sub-images went through proposed noise removal algorithm in order to remove undesired pixels. After

that, two groups of features are extracted from the handwritten sub-images. The first features group that is used for models1 includes structural, statistical, discrete cosine transform (DCT) and proposed modified histogram of gradient (MHOG1) features. However, the second features group which is used for module2 includes proposed MHOG2 and shape features. In addition, best classification results are obtained by using support vector machine (SVM) classifier. An Arabic lexicon is proposed for the first module to convert the classified classes into the Arabic editable text, and a writers' lexicon is proposed too to assign the classified label into the desired writer.

In order to test the system performance, three Arabic handwritten databases are used, which are AHDB database, IESK-arDB database and a proposed Arabic handwritten database. The results obtained from the first module were 96.317% for AHDB, 82% for IESK-arDB and 98% for the proposed database using SVM polynomial kernel. On another hand, the results of the second module using the proposed handwritten database was 85% for handwritten sub-word level and 100% for handwritten text level approaches.

وزارة التعليم العالي و البحث العلمي

الجامعة التكنولوجية

قسم علوم الحاسبات

التعرف على النص المكتوب بخط اليد للتطبيق الأمني

أطروحة مقدمة الى قسم علوم الحاسبات

في الجامعة التكنولوجية كجزء من متطلبات نيل درجة

الدكتوراه فلسفة في علوم الحاسبات

اعدت من قبل

مصطفى سلام كاظم الشمري

بأشراف

أ.م.د. علياء كريم عبد الحسن

١٤٣٨ هـ

٢٠١٦ م

المستخلص

معظم الحكومات والمنظمات لديها عدد كبير من الوثائق المكتوبة بخط اليد الناتجة عن العمليات اليومية. لا بد من استخدام أجهزة الكمبيوتر لقراءة النصوص المكتوبة بخط اليد، وجعلها قابلة للتعديل و البحث. لذلك التعرف على الكتابة اليدوية أصبح في الآونة الأخيرة موضوع بحث جدا شائع وعدد تطبيقاته المحتملة كبيرة جدا. حيث لديه القدرة على حل المشاكل المعقدة وتبسيط الأنشطة البشرية من خلال تحويل الوثائق المكتوبة بخط اليد إلى شكل رقمي وجعلها مفيدة للعديد من التطبيقات خصوصا الامنية منها من خلال خلق تعرف موثوق. ومع ذلك، فإن التعرف على النص العربي المكتوب بخط اليد هو عملية معقدة مقارنة مع أنظمة الكتابة اليدوية للغات الأخرى بسبب طبيعة المزج لكتابة اليد في اللغة العربية. وبالتالي، فإن مهمة التعرف الموثوق على النص العربي المكتوب بخط اليد للوثائق مع ما يقرب من الأداء البشري لا تزال مشكلة قائمة.

تم اقتراح نظام دقيق وموثوق للتعرف على النص المكتوب بخط اليد للغة العربية على أساس تجزئة المدخلات من نصوص الوثائق المكتوبة بخط اليد إلى كلمات فرعية مكتوبة بخط اليد. النظام يحوي اثنين من الاجزاء (modules) الأساسية المستخدمة للتعرف على النص المكتوب بخط اليد وتحديد كاتب الوثيقة. الجزء الاول (module1) له ست مراحل والتي تعمل معا للتعرف على النص العربي المكتوب بخط اليد وتحويله إلى نص قابل للتعديل. وهذه المراحل هي: اكتساب الصور، التجزئة، التجهيز، بناء قاعدة الميزات, التصنيف ومرحلة ما بعد المعالجة. في حين أن الجزء الثاني (module2) يقوم بتحديد الكتاب المطلوب للوثيقة من خلال عدة مراحل والتي هي: اكتاس الصور، التجزئة، التجهيز، استخراج الميزات, التصنيف ومرحلة ما بعد المعالجة. اقترح النظام خوارزمية تجزئة فعالة ودقيقة والتي تجزء النص المكتوب بخط اليد المدخل إلى عدد من الصور الفرعية المكتوبة بخط اليد وكل صورة فرعية تحوي على كلمة فرعية من اللغة العربية. بالاضافة الى ذلك، تم اقتراح خوارزمية صورة العتبة لتحويل الصور الفرعية إلى صورة ثنائية باستخدام دالة التجميع (fuzzy c-mean). وعلاوة على ذلك، تمر الصور الفرعية الثنائية من خلال خوارزمية مقترحة لإزالة الضوضاء من أجل إزالة البكسلات غير المرغوب فيها. بعد هذا، مجموعتان من الميزات يتم استخراجها من الصور الفرعية المكتوبة بخط اليد. المجموعة الأولى من الميزات التي تستخدم ل(models1) تضم الهيكلي، الإحصائي، discrete cosine transform (DCT) و modified histogram of gradient (MHOG1) المقترحة. من جهة اخرى، فإن مجموعة الميزات الثانية التي تستخدم ل(module2) يشمل modified histogram of (MHOG2) gradient المقترح وميزات الشكل. وبالإضافة إلى ذلك، تم الحصول على أفضل نتائج التصنيف من خلال استخدام المصنف support vector machine (SVM). وتم اقتراح معجم عربي ل(module1) لتحويل المسميات المصنفة الى نص عربي قابل للتعديل ، ومعجم للكتاب اقترح أيضا لغرض تعيين المسمى المصنف إلى الكاتب المنشود.

من أجل اختبار أداء النظام، تم استخدام ثلاثة قواعدة بيانات للغة العربية المكتوبة بخط اليد والتي هي قاعدة بيانات AHDB، قاعدة بيانات IESK-arDB وقاعدة بيانات مقترحة للغة العربية المكتوبة بخط اليد. وكانت النتائج التي تم الحصول عليها من الجزء الأول (module1) 96.317٪ لAHDB، 82٪ لل IESK-arDB و 98٪ لقاعدة البيانات المقترحة باستخدام SVM نواة متعدد الحدود. من جهة أخرى، كانت نتائج الجزء الثاني (module2) باستخدام قاعدة البيانات المقترحة 85٪ لطريقة مستوى الكلمات الفرعية المكتوبة بخط اليد و 100٪ لطريقة مستوى النص المكتوب بخط اليد.