ABSTRACT

The researches in Optical Character Recognition (OCR) area by using Hidden Markov Models (HMMs) are continuing until this moment.

The work presented in this thesis is proposed for recognition of offline isolated Arabic handwritten characters using HMMs as classifier and Freeman chain code as feature extraction method.

we use scheme to recognize the bodies of the characters only since many characters might share the same body. Characters with similar bodies are grouped as one class, then each class represented by one-character body.

This scheme decreases the number of the characters in dataset from 34 characters to 18 characters.

The dataset used in this thesis is isolated Handwritten Arabic Characters (IHAC) dataset, which collected by Arabic Language Technology Research Group at Sudan University of Science and Technology.

Most systems attempt to segment characters into sub-characters however, segmenting handwritten characters is very difficult. So, to avoid this, characters are treated without segmentation.

Moreover, this work is divided into three main phases to provide a recognition system. The first phase is the preprocessing, which applies efficient preprocessing methods which are essential for optical character recognition. In this phase, methods for normalization and digitization are implemented. Then dots are removed from some characters, then characters' bodies are thinned.

The second phase is feature extraction. This phase makes use of the thinned images to extract features that are essential in recognizing the images. Features are extracted by implementing Freeman chain code (FCC) method, then it normalized to 10 digits for each sample.

The third and final phase is the classification of characters' bodies by Hidden Markov Models (HMMs) classifier. 25179 samples from SUST/ALT dataset are used for training (70%), and testing (30%). Several experimental were examined and a best recognition performance of 59.39% is achieved for testing dataset and 80.86% for training dataset. The results were acceptable.

One of the important finding of these set of experiment is the high confusion between some classes, this due to the variation of writers' styles which case similarity between characters. Moreover, error may occur due to inadequate capability with the features used.

The proposed system has been implemented and tested on MATLAB R2010b environment.

المستخلص

مازالت البحوث مستمرة في مجال هذا البحث وذلك باستخدام نماذج ماركوف المخفية حتى يومنا هذا.

تعمل الدراسة في التعرف على حروف اللغة العربية غير المتصلة المكتوبةيدوياً باستخدام نماذج ماركوف المخفية للتصنيف واستخلاص الخواص باستخدام سلاسل فريمان.

في هذه الأطروحة تم للتعرف على اجسام الحروف، بما أن معظم الحروف لها اجسام مشابهة، تم استخدام نهج لتجميع الحروف المتشابهة الأجسام في قسم واحد وكل قسم تم تمثيله بحرف واحد. بهذه الطريقة عدد الحروف تناقص من 34 حرف الي 18 حرف.

البيانات المستخدمة في هذه الأطروحة هي عبارة عن حروف اللغة العربية المكتوبة بخط اليد التي تم الحصول عليها بواسطة مجموعة بحث تقنية اللغة العربية في جامعة السودان للعلوم والتكنلوجيا.

تحاول معظم الانظمة إلى تقسيم الحروف الى أجزاء ومع ذلك نجد أن هنالك صعوبة في تقسيم الحروف المكتوبة بخط اليد، ولتفادي هذه المشكلة، تمت معالجة الحروف بدون تقسيم في هذه الأطروحة.

تم تقسيم هذا العمل إلى ثلاثة مراحل اساسية وذلك للتعرف على النظام.

المرحلة الاول وهي التجهيز، والتي تقوم على تطبيق طرق اساسية للتعرف على الحروف. في هذه المرحلة يتم تطبيق أحرف خاصه للتطبيع والترقيم. ومن ثم ازالت النقاط من بعض الحروف وتنحيف أجسام الحروف.

المرحلة الثانية استخلاص الخواص. هذه المرحلة تقوم باستخدام الحروف ذات الصور غير الواضحة أو المبهمة لاستخلاص السمات أو خواص اساسية لإدراك هذه الصور.

ثم استخلاص الخواص أو الميزات وذلك عن طريق تطبيق خوارزمية فريمان ومن ثم تطبيع هذه الخواص الى عشرة ارقام لكل نموذج.

المرحلة الثالثة والاخيرة هي عبارة عن التعرف على الحروف باستخدام نماذج ماركوف المخفية (HMMs).

25179 عينة من قاعدة البيانات استخدمت 70% للتمرين، و30% للاختبار ثم أجراء العديد من الاختبارات وأفضل اختبار أدرك أداء 59.39% ثم تحقيقه في اختبار البيانات و80.86% لتمرين البيانات وكانت النتائج مقبولة من اهم الاكتشافات في هذه التجارب وجود التباس في التعرف على بعض المجموعات، هذا بسبب الاختلاف في طريقة الكتابة للكاتبين مما يؤدي الي التشابه بين الحروف بالإضافة الي ذلك الأخطاء قد تحدث بسبب عدم دقة الخواص المستخدمة.

تم تطبيق هذا النظام المقترح على بيئة الماتلاب R2010b.