المراحل التمهيدية وأهميتها في أنظمة التعرف على الكتابة العربية

مصطفى علي أبوزريدة[1]، أكرم محمد زكي[2]، أحمد محمد زكي[3]
[1،2] كلية تقنية المعلومات، الجامعة الاسلامية العالمية ماليزيا
[3] قسم انظمة المعلومات، كلية تقنية المعلومات، جامعة البحرين، مملكة البحرين
abuzaraida@yahoo.com[1]
akramzeki@iium.edu.my[2]
amzeki@uob.edu.bh[3]

الخلاصة: بدأت محاولات التعرف الضوئي على الكتابة والنصوص قبل الحرب العالمية الثانية، لكن لم تكن حتى أوائل 1950 إلا مجرد مشروع تجاري كان من الضروري تحديد ما يبرره لتمويل البحث والتطوير للتكنولوجيا. أصبح هذا المجال فيما بعد ذلك الشغل الشاغل للعديد من المهتمين بالحاسب الآلي والمبرمجين خصوصًا. وذلك لإمكانية الاستفادة من هذا المشروع بشكل كبير في نقل المعلومات والمعارف إلى الحاسب بسرعة ويسر، ولما يمكن أن يوفره من سهولة في التعامل مع الحاسب الآلي. ينقسم نظام التعرف الضوئي على النصوص إلى مجموعة من الخطوات لعل أهمها هي المرحلة التمهيدية. حيث تعتبر الخطوة الأولى للنظام التي يتم فيها تنقيح البيانات المُدخَلة مما يزيد من كفاءة النظام ودقته في التعرف على النص المُدخَل. في هذا البحث تم التطرق للخطوات الأكثر شيوعًا واستخدامًا في هذا النوع من الانظمة وكذلك ملخص لبعض العراقيل في التعرف على النصوص العربية عند إدخال النصوص لهذه الأنظمة.

الكلمات الجوهرية: التعرف المباشر على النصوص، المرحلة التمهيدية، اللغة العربية.

1 المقدمة

زاد اهتمام الباحثين في المشاكل المتعلقة بمحاكاة الآلة للإنسان في عملية القراءة، فبدأت الأبحاث على محمل الجد لتطوير هذا النظام وجُمعت الأبحاث والنظريات فلوحظ الانجذاب الهائل إلى هذا الموضوع بسبب مستوى التحدي الذي تمتاز به المشاكل المصاحبة لهذا المجال، وكان يوصف بالعمل الصعب نوعاً ما لاحتوائه على خاصية التعرف على الحروف المكتوبة بخط اليد، وما يزيد من مدى التعقيد الموجود في هذه الخاصية هو القدرة على التعرف على الأنماط المختلفة للكتابة بخط اليد والتي تعتمد على الطبيعة والشكلية التي يكتُب بها الكاتب بخط يده، فتحليل الحروف المتصلة في النص المكتوب بخط اليد يتطلب فصل هذه الحروف في الكلمة الواحدة عن بعضها البعض ومن ثم التعرف على الحرف الواحد، ولا يُستغرب هذا الأمر في الحاسب الآلي فالبشر الذين يملكون أكثر الأجهزة كفاءة في القراءة (العين) يجدون صعوبة في تمييز بعض الحروف المتصلة مع بعضها البعض حيث تصل نسبة الخطأ عند البشر إلى 4% تقريبا [أمين 2002].

منذ أكثر من ثلاثة عقود والعديد من الباحثين يعملون على تطوير نظام التعرف على الحروف المتصلة في الكلمة المكتوبة بخط اليد، وبالرغم من طول هذه الفترة نسبيا، إلا أن هذا المجال يعتبر صعباً وكل النظم الموجودة تبقى محدودة في تطبيقات معينة وبمحددات Constraints عديدة.

تنقسم أنظمة التعرف الضوئي على النصوص المكتوبة إلى نظامين أساسيين وهما: نظام التعرف المباشر (Online) ونظام التعرف غير المباشر (Offline)، تتم عملية التعرف على النصوص في الأول كأشكال ومتجهات vectors كما تمت كتابتها حيث يتم التعرف على الكتابة اليدوية مباشرة أثناء الكتابة باستخدام قلم ضوئي على لوحة إدخال خاصة موصولة بالحاسب. أما بالنسبة للثاني فتتم عملية التعرف فيه بعد إتمام الكتابة أو طباعة النص.

ترتبط مراحل التعرف في النظام الضوئي المباشر ببعض وتعتمد كل مرحلة على نتائج المرحلة التي سبقتها، في هذه الورقة سيتم تناول المرحلة التمهيدية فقط من حيث الأهمية والمشاكل الناتجة عن إدخال النص والخطوات الشائعة المتبعة في أغلب الأنظمة السابقة للتغلب على هذه المشاكل أو التقليل من أثرها.

القسم التالي من هذا البحث يسلط الضوء على التعرف الضوئي على النصوص العربية بشكل عام، يليه سرد لأهم خصائص اللغة العربية التي يجب مراعاتها عند تصميم أنظمة التعرف على النصوص العربية. القسم الرابع يستعرض مراحل أنظمة التعرف الضوئي على النصوص العربية. يلي ذلك عرض تفصيلي للخطوات المتبعة في المرحلة التمهيدية التي تسبق عملية التعرف على النصوص الأولية. يلي ذلك الخاتمة والدراسات المستقبلية المقترحة.

2 التعرف الضوئي على النصوص العربية

تعتبر اللغة العربية من أكثر اللغات انتشارًا حول العالم حيث يزيد عدد الناطقين بها عن 280 مليون شخص [مصطفى ومن معه 2010] ، مما أوجد الدافع لتطوير نظام آلي للتعرف على النصوص العربية. تتعدد اللهجات في البلاد العربية ولكن الكتابة العربية موحدة في مختلف أنحاء العالم العربي ويصطلح عليها باللغة العربية القياسية الحديثة Modern Standard Arabic. كما أن العديد من اللغات الأخرى تستخدم الحروف العربية مثل الفارسية والكردية والأردية، لذا فإن القدرة على تفسير الكلام العربي المكتوب آلياً له فوائد واسعة.

فمنذ ثمانينيات القرن الماضي، والبحث يتزايد في سبيل تطوير أنظمة للتعرف الآلي على الحروف العربية، حيث يعتبر هذا المجال مهما ليس للدول الناطقة بالعربية فحسب، بل وحتى للغات المشابهة مثل الأوردو والفارسية والتي تشترك مع العربية بمجموعة من الحروف. ومن الجدير بالذكر أن التقنيات المستخدمة للتعرف على الحروف اللاتينية والصينية لا يمكن تطبيقها بسهولة على اللغة العربية وذلك لطبيعة الكتابة العربية. يستعرض القسم التالي أهم خصائص الكتابة العربية والتي يجب مراعاتها عند تطوير برامج التعرف الآلي على النصوص العربية.

3 خصائص الكتابة العربية

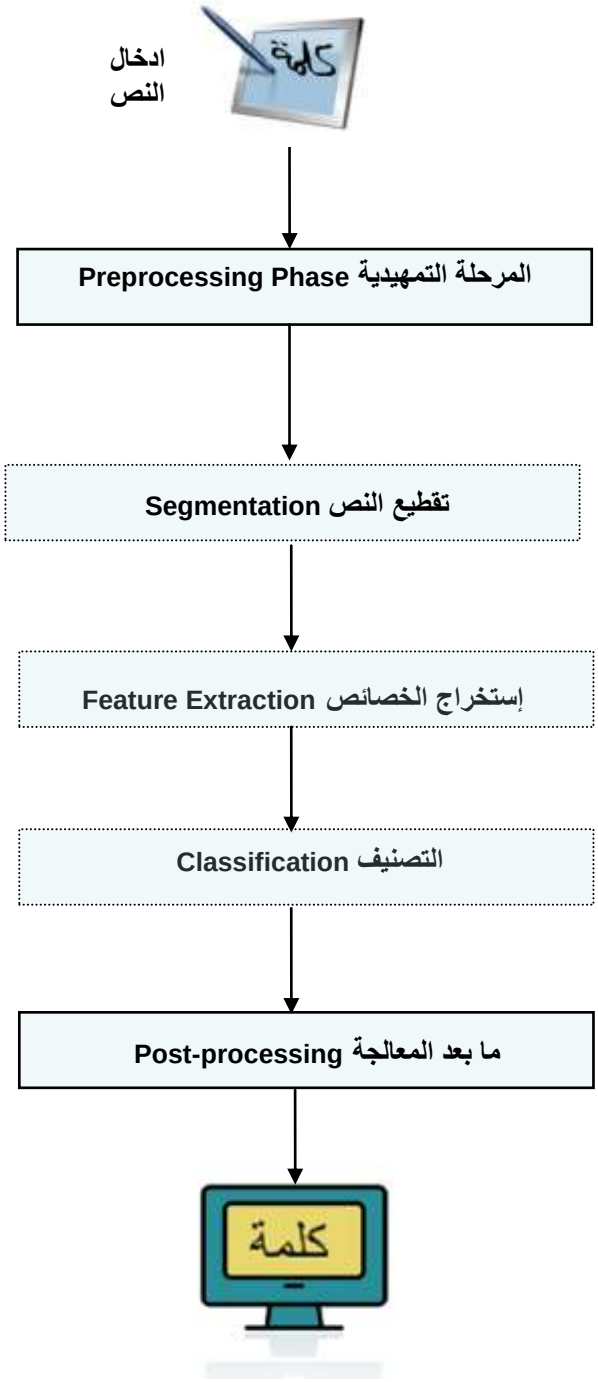
تتكون الكلمات العربية بكتابة حروف متصلة ومرتبطة لتكوين كلمة مفهومة. تتكون الحروف العربية من 28 حرفاً وتكتب من اليمين إلى اليسار. وقد يكون للحرف الواحد أربعة أشكال اعتماداً على موقعه في الكلمة. كذلك كثير من الحروف مكونة من جزأين وهما الجسم وعدد من النقاط أو الأشكال فوقها أو تحتها أو داخلها. ويكون عدد النقاط إما واحدا أو اثنان أو ثلاثة للتمييز بين الحروف المتشابهة شكلا ونطقا. أما الأشكال فهي كثيرة منها الهمزة والمدة والشدة وعلامات الضبط. الجدول 1 يبين قائمة الحروف العربية حسب موقعها من الكلمة.

الجدول 1: قائمة الحروف العربية في الأشكال الأربع المختلفة

الحرف	الشكل المنفصل	الشكل في بداية المقطع	الشكل في منتصف المقطع	الشكل في نهاية المقطع
ألف	ا			ـا
باء	ب	بـ	ـبـ	ـب
تاء	ت	تـ	ـتـ	ـت
ثاء	ث	ثـ	ـثـ	ـث
جيم	ج	جـ	ـجـ	ـج
حاء	ح	حـ	ـحـ	ـح
خاء	خ	خـ	ـخـ	ـخ
دال	د			ـد
ذال	ذ			ـذ
راء	ر			ـر
زاي	ز			ـز
سين	س	سـ	ـسـ	ـس
شين	ش	شـ	ـشـ	ـش
صاد	ص	صـ	ـصـ	ـص
ضاد	ض	ضـ	ـضـ	ـض
طاء	ط	طـ	ـطـ	ـط
ظاء	ظ	ظـ	ـظـ	ـظ
عين	ع	عـ	ـعـ	ـع
غين	غ	غـ	ـغـ	ـغ
فاء	ف	فـ	ـفـ	ـف
قاف	ق	قـ	ـقـ	ـق
كاف	ك	كـ	ـكـ	ـك
لام	ل	لـ	ـلـ	ـل
ميم	م	مـ	ـمـ	ـم
نون	ن	نـ	ـنـ	ـن
هاء	ه	هـ	ـهـ	ـه
واو	و			ـو
ياء	ي	يـ	ـيـ	ـي

4 مراحل التعرف على النصوص العربية المكتوبة باليد في الأنظمة الضوئية

بالرغم من أن الباحثين حققوا تقدمًا ملحوظًا وطفرة كبيرة من خلال توفر أجهزة الإدخال الضوئي وانتشارها بشكل واسع في الأسواق العالمية، إلا أن هذا المجال لا زال يحظى باهتمام الباحثين إلى اليوم. كذلك أسهمت اللوحات الرقمية والأجهزة الشخصية بشكل كبير في اثراء ودعم هذا الحقل العلمي. تتكون أغلب أنظمة التعرف على الحروف الضوئية المباشرة من مجموعة من المراحل هي: المرحلة التمهيدية Preprocessing، التقطيع Segmentation، استخراج الخصائص Feature Extraction، التصنيف Classification، ما بعد المعالجة Post-processing كما هو موضح بالشكل 1 [مصطفى ومن معه 2012].



شكل 1: مراحل النظام الضوئي المباشر للتعرف على النصوص

تختلف المرحلة التمهيدية بشكل كبير في الانظمة المباشرة عنها في الأنظمة غير المباشرة، حيث تختلف طريقة الإدخال في النظامين بشكل كامل. ففي النظام غير المباشر، من الممكن حصول تشوهات على البيانات المدخلة في الصورة الناتجة من عدة أسباب مثل انخفاض جودة الماسح الضوئي Scanner أو انخفاض جودة الورق أو التشوهات المتكونة على الورق وغيرها. في حين قد تظهر مشاكل التشوهات من خلال عيوب في أجهزة الادخال مثل الالواح الرقمية Digital Tablets في الأنظمة المباشرة. إن الهدف الرئيس من إجراء هذه المرحلة هو التقليل قدر الإمكان من هذه التشوهات ليتم تمرير بيانات ذات جودة ودقة عالية للمراحل التالية مما يساهم في تحسين الدقة النهائية للتعرف على النصوص من قبل النظام. ومن أشهر هذه التشوهات التي قد تنتج عند إدخال النصوص ما يأتي [فادي ومن معه 2006] و[ايمان ومن معه 2010]:

أ- حجم النص المدخل Text Size

حيث قد يختلف حجم النص المدخل عند كل مرة يتم فيها إدخال النص. كذلك يختلف حجم الكتابة من شخص إلى آخر. إن الاختلاف في حجم النص الواحد قد يعطي أبعادًا مختلفة للنص المراد التعرف عليه مما يترتب عليه اختلافًا في خصائصه ومن ثم يتسبب ذلك في عدم الحصول على مخرجات صحيحة من النظام.

ب- موقع النص Text Position

قد تختلف إحداثيات النص المدخل عندما يقوم الكاتب بالكتابة في المساحة المخصصة للكتابة، حيث أن هذا الاختلاف قد يعطي النظام أبعادًا مختلفة للنص الواحد مما قد يقلل من إمكانية التعرف على النص الصحيح.

ج- عدم تساوي المسافات بين النقاط المكونة للنص

نتيجة لسرعة الكتابة على سطح جهاز الادخال، من المحتمل الحصول على نقاط غير موزعة بشكل متساو، مما يقلل من انتظام النقاط المطلوبة للحصول على الشكل المطلوب.

د- تذبذب النص Zigzag Shape

قد يواجه بعض مستخدمي الأنظمة الضوئية صعوبة في التحكم بالقلم والكتابة بالشكل الصحيح كما هو معتاد على الورق، مما ينتج عنه تذبذب وتشوه في الكتابة وهذا بدوره يقلل من دقة النص المدخل ويزيد في عدم وضوحه.

هـ- ميل وانحراف النص المكتوب Slant and Skew of the Handwriting

ميل الكتابة هو الكتابة بشكل مائل على المحور الافقي، أما انحراف الكتابة فهو الكتابة بشكل مائل على المحور العمودي. وهو من الأخطاء والمشاكل الشائعة عند إدخال النصوص اليدوية المباشرة.

و- الزوائد عند بداية ونهاية الكتابة Hooking

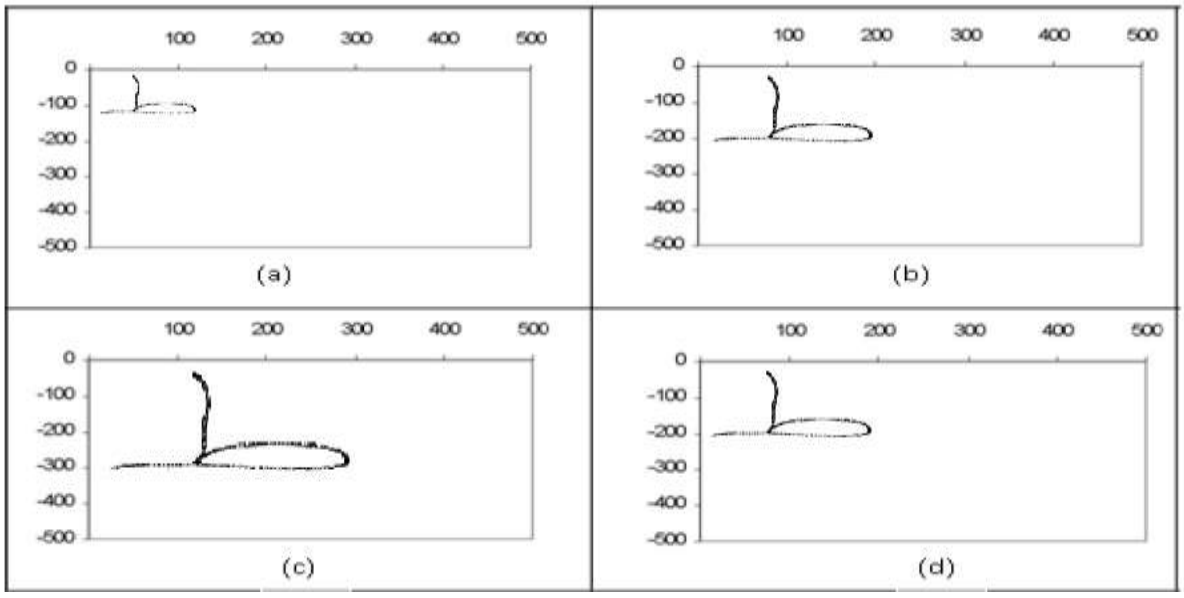
تعتبر الزوائد إحدى المشاكل التي تقلل من كفاءة النظام. تنتج هذه الزوائد عند البدء في الكتابة على السطح الضوئي بدون التركيز على بداية رسم النص، كذلك قد تنتج عند رفع القلم عن السطح حيث تسمى كلابات Hooks.

5 الخطوات المتبعة في المرحلة التمهيدية:

كما أشير سابقًا فإن المرحلة التمهيدية من أهم مراحل نظام التعرف الضوئي مباشرًا كان أو غير مباشر، حيث تعتبر هذه المرحلة مرحلة إعداد وتهيئة للنص لتمريره للمراحل التالية في تشكيل ملائم وموحد لتسهل عملية التعرف على النص الصحيح من قبل النظام. تختلف خطوات مرحلة التمهيد من نظام إلى آخر بناءا على نوع النص واللغة وكذلك تعتمد على طريقة التصنيف Classification Method. ومن أشهر الطرائق المتبعة في المرحلة التمهيدية في النظام المباشر[مصطفى ومن معه 2013]:

أ. تغيير حجم النص Text Resizing

أحد وظائف مرحلة التهيئة هو إحداث تناغم بين أحجام النصوص. وتظهر أهمية هذه الخطوة في الكتابة اليدوية حيث تكون النصوص غير متطابقة في حجمها نتيجة لطبيعة الكتابة اليدوية. يتم في هذه الخطوة تغيير حجم النص إما بتصغيره أو تكبيره إلى حجم واحد لتسهل عملية استخراج الخصائص من النص وحتى تكون هذه الخصائص متشابهة للنص الواحد. الشكل 2 يبين مثالا على تغيير حجم حرف وذلك لتوحيد الحجم لجميع النصوص المدخلة.

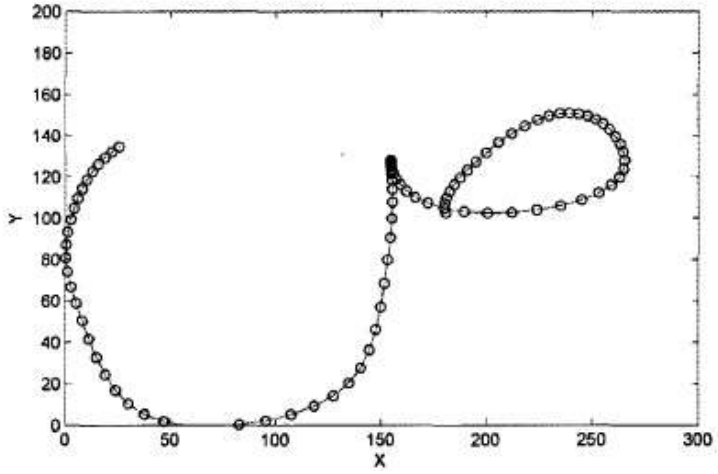


شكل 2: تغيير حجم النص

ب. مركزية النص Text Centering

تعتبر من الطرق الشائعة خلال المرحلة التمهيدية، حيث يقوم النظام بإزاحة النص ليلامس المحورين العمودي والأفقي الرئيسين عند أصغر إحداثيين للنص، وبذلك يضمن المبرمج أن جميع النصوص لها إحداثيات متشابهة مهما

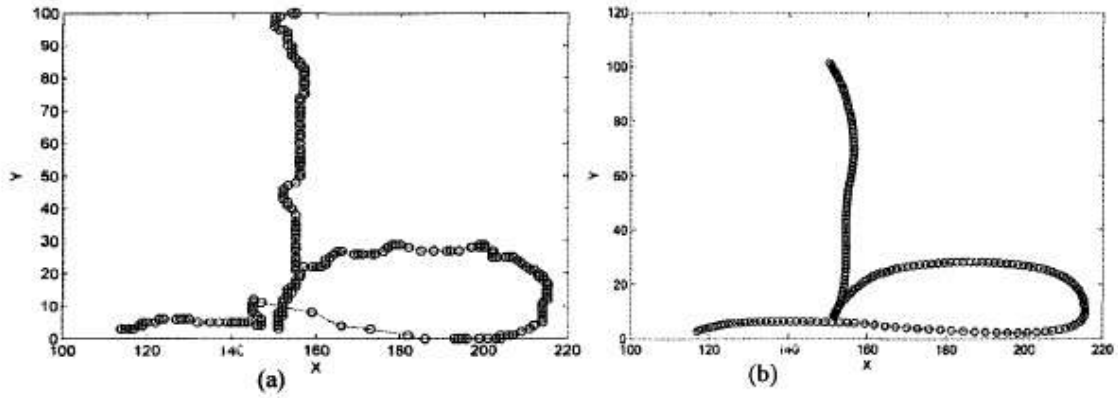
اختلفت أماكن كتابتها على مساحة الكتابة. الشكل 3 يبين مثالا على عملية إزاحة النص إلى نقطة الأصل. ومن الطرائق المشابهة لهذه الطريقة هي إحاطة النص بإطار واعتبار المحورين العمودي والأفقي هما المحورين الرئيسين.



شكل 3: إزاحة النص إلى المحور الرئيس

ج. صقل النص Smoothing

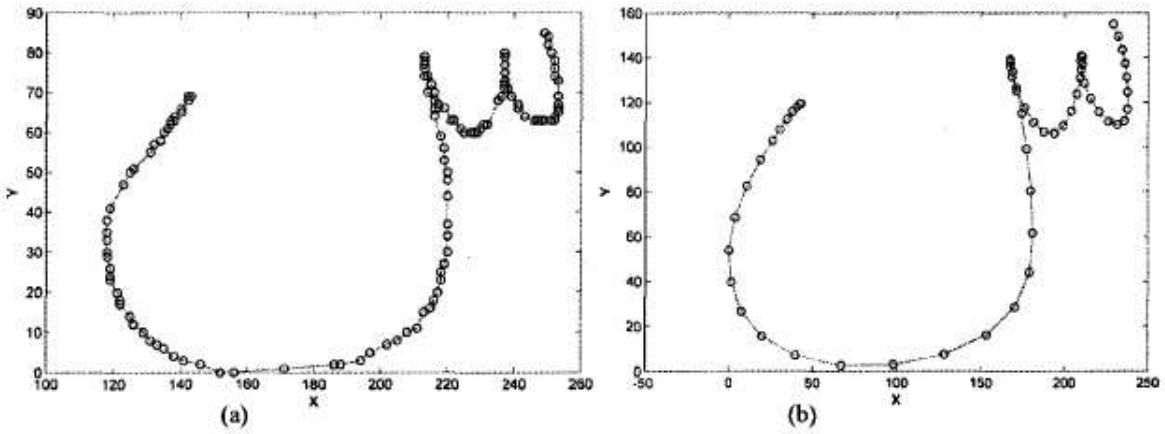
نتيجة للتذبذب الحاصل من الكتابة اليدوية على السطح الأملس وكذلك لعدم انتظام النقاط المكونة للنص، فعليه يتوجب تصحيح هذه العيوب في شكل النص وذلك بتطبيق بعض الخوارزميات التي تسهل وتساهم في صقل النص. ومن أشهر الخوارزميات المستخدمة لصقل النص هي خوارزمية لويس LOESS كما هو مبين في الشكل 4.



شكل 4: صقل النص وتقليل التشوه الحاصل في عملية الإدخال

د. توزيع نقاط الكتابة بشكل متساو Re-sampling

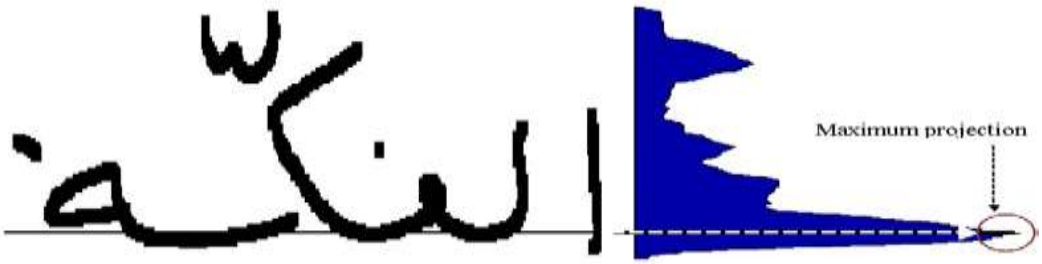
من الحلول المطبقة لجعل النص أكثر وضوحا ودقة هو عملية توزيع نقاط الكتابة بشكل متساو، ويتم ذلك إما بإلغاء النقاط أو بإضافة نقاط جديدة للخط المرسوم. في هذه الطريقة يتم اعتماد معامل قياسي للمسافة بين كل نقطتين على خط الكتابة، بحيث يتم تحوير مواقع النقاط لتنتج مسافات متساوية بين النقاط كما هو مبين بالشكل 5.



شكل 5: إعادة توزيع النقاط بشكل متساو

هـ. تحديد السطر المرجعي Baseline Detection

من أشهر الطرائق المستخدمة لمعالجة مشكلة ميل النص هي تحديد السطر المرجعي Baseline. توجد أكثر من طريقة لتحديد السطر المرجعي لعل من أنجحها فيما يتعلق باللغة العربية هي طريقة الخطوط البيانية Histogram حيث يتم حساب عدد النقاط في كل خط أفقي ويكون السطر المرجعي هو الخط الأفقي صاحب أكبر عدد من النقاط للسطر الواحد. يوضح الشكل 6 مثالاً على استخلاص السطر المرجعي.



شكل 6: تحديد السطر المرجعي للنص المدخل

يعتبر التعرف الضوئي المباشر للنصوص العربية أحد مواضيع الدراسة في حقل التعرف الضوئي للأشكال والقوالب في الآونة الأخيرة. نظرا لصعوبة التعرف على الكتابة العربية وعدم تشابهها مع اللغات الاخرى، استوجب تطوير واستحداث اساليب وطرق جديدة مختلفة عن الطرق المستخدمة في اللغات الاخرى. من هذه الطرائق خطوات المرحلة التمهيدية في التعرف الضوئي المباشر للنصوص العربية. في هذا البحث تم التطرق للخطوات الاكثر شيوعا واستخداما في هذا النوع من الانظمة وكذلك ملخص لبعض العراقيل في التعرف على النصوص العربية عند ادخال النصوص لهذه الانظمة.

من خلال الدراسات السابقة في مجال التعرف الضوئي المباشر للنصوص العربية ومقارنتها مع الكم الهائل من البحوث التي أنجزت على اللغات الاخرى، يستوجب بذل مجهود مضاعف من قبل الدارسين والبحاث للوصول إلى المستوى المأمول في مجال التعرف الضوئي المباشر للنصوص العربية المكتوبة بخط اليد. ومن هذا المنطلق سيتم في القريب العاجل استكمال مشروع لنظام ضوئي مباشر للكلمات العربية، نأمل من الله عز وجل التوفيق والسداد فيه.

6 المراجع

[أمين 2002]

Amin, A, "Structural description to recognising Arabic characters using decision tree learning techniques". Joint IAPR International Workshops on Structural, Syntactic, and Statistical Pattern Recognition (SSPR'02 and SPR'02). pp. 152-158. Windsor, Ontario, Canada. 6-9 August.

[مصطفى ومن معه 2010]

M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Segmentation techniques for online Arabic handwriting Recognition: A survey," in 3rd International Conference on Information and Communication Technology for the Moslem World: ICT Connecting Cultures, ICT4M 2010, Jakarta, Indonesia, 2010, pp. D37-D40.

[مصطفى ومن معه 2012]

M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Recognition Techniques for Online Arabic Handwriting Recognition Systems," in International Conference on Advanced Computer Science Applications and Technologies (ACSAT2012), Kuala Lumpur, Malaysia, 2012.

[مصطفى ومن معه 2013]

M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Problems of writing on digital surfaces in online handwriting recognition systems," in Information and Communication Technology for the Muslim World (ICT4M), 5th International Conference on, 2013, pp. 1-5.

[فادي ومن معه 2006]

F. Biadsy, J. El-Sana and N. Habash, "Online Arabic Handwriting Recognition using Hidden Markov Models," in 10th International Workshop on Frontiers in Handwriting Recognition, La Baule (France), 2006, pp. 85-90.

[ايمان ومن معه 2010]

I. Khodadadzadeh, "Recognition of Handwritten Arabic Characters," Master's thesis, Department of Electrical and Computer Engineering, The University of Windsor, Windsor, Faculty of Graduate Studies, Ontario, Canada, 2010.

7 الخلاصة باللغة الانجليزية

The Importance of Preprocessing Phase in Online Arabic Character Recognition System

Mustafa Ali Abuzaraida[1], Akram M. Zeki[2], Ahmed M. Zeki[3]

[1,2] Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia

[3] Department of Information Systems, College of Information Technology University of Bahrain, Sakhir, Kingdom of Bahrain

[1] abuzaraida@yahoo.com, [2] akramzeki@iium.edu.my, [3] amzeki@uob.edu.bh

Abstract:

The typical online recognition system contains of main four phases which are: Preprocessing, Feature extraction, Recognition, and post-processing phases. Preprocessing phase aims to reduce or remove imperfections caused by acquisition devices, and to minimize handwriting variations irrelevant for pattern classification which may exist in the acquired data. The preprocessing phase has a great influence on subsequent processing, and a real impact on the recognition rate. This paper presented and highlighted the problems those may appear during capturing the entered Arabic handwritten text and also presents the steps of the preprocessing phase with the importance of each step.