التجزئة غير الآنية للكلمات العربية: التحديات والحلول

فاطمة علي المحجوب
كلية التقنية الصناعية، قسم الهندسة الإلكترونية، مصراتة، ليبيا
ffffaaattm@gmail.com

البشير علي الجبو
كلية التقنية الصناعية، قسم الهندسة الإلكترونية، مصراتة، ليبيا
beljabu@gmail.com

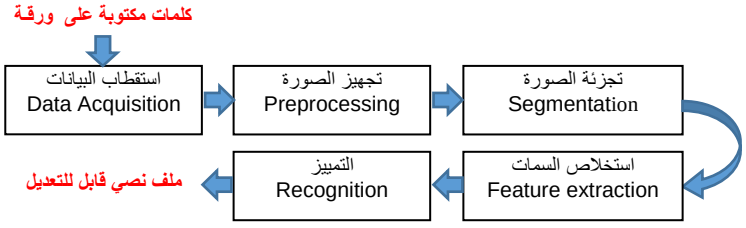
أحمد يوسف بن ساسي
كلية التقنية الصناعية، قسم الهندسة الإلكترونية، مصراتة، ليبيا
dr_ahmed@cit.edu.ly

الملخص— **إن تجزئة صورة الكلمة إلى مكوناتها الأساسية من أهم وأكبر التحديات التي تواجه بناء أنظمة تمييز الكلمات. ويزداد هذا التحدي صعوبة مع الكتابة العربية بسبب طبيعتها المتصلة سواء المطبوعة آلياً أو المكتوبة باليد. ومن خلال الأبحاث والدراسات التي اهتمت بنظام تمييز الكتابة العربية تبين مدى أهمية مواجهة مشكلة التجزئة وإيجاد حلول لها، فلا يمكن الوصول إلى نظام ناجح لتمييز الكلمات ما لم تحل مشكلة تجزئة الكلمة إلى حروف. وتكمن أهمية ذلك في أن أي تحسين صغير في خوارزمية التجزئة ينتج عنه تحسينات كبيرة وواضحة في معدل تمييز النظام. كما أن مرحلة التجزئة تسبب أخطاء أكثر مما يسببه تشوه الشكل الناتج عن جهاز استقطاب البيانات بالإضافة إلى أن إيجاد طريقة ناجحة لتجزئة الكلمات يؤدي إلى تصغير حجم قواعد بيانات التدريب. والتجزئة الجيدة للكلمات تفتح الباب لبناء نظام يمكنه تمييز كلمات غير محدودة. هذه الورقة تناقش أهم التحديات والحلول لتجزئة الكلمات العربية نحو نظام أمثل لتمييز الكلمات العربية.**

الكلمات المفتاحية : تجزئة الكلمات إلى حروف، الكتابة العربية، التجزئة الصريحة، التجزئة الضمنية.

1. المقدمة

منذ أن عرف الإنسان الكتابة أصبحت الوسيلة الأكثر استخداما لتجميع وتخزين وتبادل المعلومات. وبظهور الحاسوب الرقمي وانتشاره ودخوله إلى شتى المجالات لم تعد الكتابة وسيلة للاتصال بين الناس بعضهم ببعض فقط ولكن أيضا هي وسيلة للاتصال أو التخاطب بين الإنسان والآلة. وبرغم وضوح وسهولة تعلم الإنسان الكتابة والقراءة، إلا أن محاكاة الحاسوب لعملية القراءة من المهام البالغة الصعوبة والتي لازالت قيد البحث والتطوير. وتتم عملية المحاكاة هذه باستخدام ما يسمى نظام تمييز الكلمات word recognition system، والذي يتكون بشكل عام من خمس مراحل مبينة في (الشكل 1).



شكل 1. مراحل نظام تمييز الكلمات

ففي مرحلة استقطاب البيانات يتم إدخال النص المراد تمييزه إلى الحاسوب عن طريق الماسح الضوئي أو الكاميرا الرقمية ليحفظ في ملف صورة. ثم يمر بمرحلة التجهيز التي يتم فيها التخلص من الإشارات الضوضائية الناتجة عن استقطاب البيانات، بالإضافة إلى استخدام تقنيات معالجة الصور الرقمية لتحسين الصورة المدخلة. بعد ذلك تبدأ مرحلة تجزئة صورة النص إلى مكونات أصغر، ولها ثلاث مستويات. يتم أولا تجزئة النص إلى سطور ومن ثم يتم تجزئة كل سطر إلى كلمات، وأخيرا يتم تجزئة كل كلمة إلى قطع أوشرائح أو حرف من حروف اللغة، وذلك بما يتماشى مع نظام التمييز المستخدم. تمرر نتائج عملية التجزئة إلى مرحلة استخلاص السمات التي ينتج عنها تمثيل للمدخلات بمجموعة من السمات التي تصف ملامحها، وذلك لتقليل حجم البيانات باستخدام طرق مختلفة مثل التحويلات الرياضية والطرق الإحصائية والهندسية أو المزج بين عدة طرق. تقوم مرحلة التمييز بإستخدام السمات المتحصل عليها لمعرفة الحرف أو الكلمة في الصورة، مستخدمة تقنيات متعددة مثل الشبكات العصبية ونماذج ماركوف المخفية. وبالرغم من أن مرحلة التجزئة تتوسط مراحل نظام تمييز الكلمات، إلا إنها لا تلاقي الاهتمام الكافي رغم تأثيرها الكبير على النظام ككل. لهذا كان تركيز هذه الورقة البحثية على هذه المرحلة لما تحدثه من تأثيرات كبيرة وواضحة على أنظمة تمييز الكلمات ذات الطبيعة المتصلة كالكتابة العربية.

2. دراسة مسحية لاستراتيجيات تجزئة الكلمات

في هذا الجزء من الورقة سيتم القاء الضوء على الاستراتيجيات التي استخدمت في تجزئة الكلمات العربية واللاتينية ، من خلال التصنيفات التي وضعت لتسهيل دراسة هذه الاستراتيجيات. فقد قام أمين [1] بتقسيم طرق تجزئة الكلمات إلى نوعين؛ طرق تحليلية Analytical Approaches وتضم الاستراتيجيات التي تقوم بتجزئة الكلمة إلى حروف، وطرق شاملة Global Approaches وهي التي تتجنب تجزئة الكلمات وتتعرف على الكلمة كاملة. وبشكل مشابه صنف خورشيد [2] طرق التعرف على الكتابة العربية إلى نوعين حسب وجود وعدم وجود تجزئة إلى أنظمة تعرف تقوم بتجزئة الكلمة Segmentation Based وطرق تعرف لا تقوم بتجزئة الكلمة. أما براويز[3] ، فقد ذكر أن عملية تجزئة الكلمة هي جزء من مرحلة التجهيز في نظام التعرف، وبعد أن قام بتحليلها أمكنه تقسيمها إلى ثمانية مجموعات، المجموعات السبع الأولى يمكن أن تندرج من ضمن طرق التجزئة الصريحة والأخيرة هي طريقة التجزئة المعتمدة على التعرف. وبشكل مشابه صنف أحمد زكي طرق تجزئة الكلمات العربية ولكن إلى تسع مجموعات والأخيرة هي طريقة التجزئة المعتمدة على التعرف [4]. وبالمقارنة مع طرق تجزئة الكلمات اللاتينية يمكن الاستفادة من الدراسة المسحية التي قدمها ريتشارد وإيرك [5] والتي تضمنت طرق واستراتيجيات التجزئة غير الآنية للكلمات اللاتينية. ففي هذه الدراسة صنف الباحثان طرق تجزئة الكلمات اللاتينة إلى ثلاث فئات، واعتمدا في هذا التصنيف على مدى التفاعل بين مرحلتي التعرف والتجزئة لنظام التعرف على الكلمات. وقد تضمنت الفئة الأولى طرق التجزئة بالتشريح Dissection Approaches ، وأطلق عليها أيضا التجزئة الصريحة Explicit Segmentation. وهذه الطرق تقوم بتجزئة صورة الكلمة إلى صور أصغر تمثل الحروف من خلال البحث عن نقاط الاتصال بين الحروف أو الواصلة التي تربطها، وتتم تجزئة الكلمة إلى حروف عند نقاط الاتصال أو الواصلات التي تم اكتشافها. تتميز هذه الفئة بأن مرحلة التجزئة ومرحلة التعرف منفصلتان ومتتاليتان ولا تعتمد

إحداهما على الأخرى. أما الفئة الثانية فقد شملت الطرق التي تعتمد فيها عملية التجزئة على عملية التعرف Recognition Based Segmentation، كما سميت أيضاً بالتجزئة الضمنيـة Implicit Segmentation. وفي هذه الطرق يوجد تداخل بين مرحلتي التجزئة والتعرف، فالنظام يقوم أولاً بالتعرف على الحروف منفصلة وتقسيمها إلى أصناف، كل صنف يحتوي على عدة أشكال لإحدى أبجديات اللغة المراد التعرف عليها. وفي مرحلة التجزئة يتم البحث داخل صورة الكلمة بشكل تكراري عن الأشكال التي تتطابق مع إحدى الأصناف التي تعرف عليها النظام مسبقاً. وأخيرا فإن الفئة الثالثة احتوت على الطرق الشمولية Holistic Approaches، وأحياناً تسمى الطرق الخالية من التجزئـة Free Segmentation، ويقصد بها مجموعة الطرق التي تجنبت عملية التجزئة واكتفت بالتعرف على حروف منفصلة أو كلمة كاملة دون تجزئة. وقد استعرضت هذه الدراسة، عدداً من الطرق التي صنفت ضمن كل فئة من الفئات الثلاث. وتجدر الاشارة إلى أن هذا التصنيف قد استخدم في أبحاث عديدة أخرى [6]، [7].

وتلخيصاً لما سبق يمكن تقسيم الاستراتيجيات التي اتبعت في بناء أنظمة التعرف على الحروف إلى نوعين: الاستراتيجية الشمولية والتحليلية.

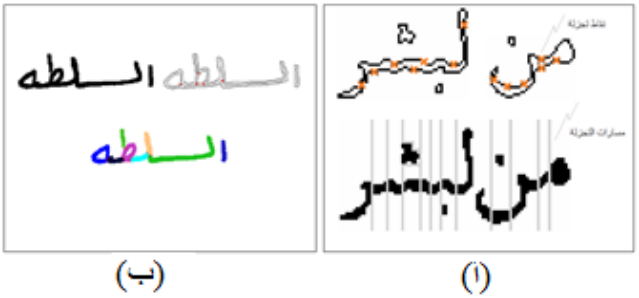
1.2 الاستراتيجية الشمولية Holistic Strategy

تستخدم هذه الاستراتيجية بشكل عام في أنظمة تمييز الكتابة باليد. وتعتبر صورة الكلمة هي الأساس في عملية التمييز، حيث يتم استخلاص السمات من الكلمة دون تجزئتها ومن ثم ارسال متجه السمات إلى مرحلة التصنيف ليتم التعرف على الكلمة .

وعادة ما يتم استخلاص السمات باستخدام التحويلات الرياضية مثل تحويل فورير وتحويل جيب التمام والعزوم وغيرها. وقد حقق هذا النهج نجاحاً مع بعض التطبيقات مثل قراءة العناوين البريدية والصكوك المصرفية، إلا انها تستخدم مع الأنظمة التي تحتاج قاموس محدود من الكلمات للحصول على أفضل النتائج، فكلما زاد عدد الكلمات المراد التعرف عليها قل معدل التعرف [8]، [9]، [10]، [11]، [12].

2.2 الاستراتيجية التحليلية Analytical Strategy

وتوصف أيضا بالتشريح Dissection، لأنها تعتمد على تحليل محتويات الصورة في تجزئتها إلى أجزاء صغيرة أو قطع، أو ايجاد مسارات أو خطوط لتجزئتها إلى حروف أو أشباه حروف منفصلة. يبين (الشكل 2) أمثلة للاستراتيجية التحليلية لتجزئة كلمة، ففي (الشكل 2. أ) تم استخراج وتحليل المحيط الخارجي للكلمات للبحث عن نقاط تجزئة مؤقتة، بعد ذلك تمت دراسة هذه النقاط للحصول على نقاط التجزئة النهائية [13]. أما المثال المبين في (الشكل 2. ب) فيظهر صورة تحتوي على كلمة "السلطة" وقد تم استخراج المحيط الخارجي لها بالاضافة إلى هيكلها الذي اشتق من خلال تنحيفها وتم الاعتماد عليهما في إيجاد نقاط التجزئة ومن تم تجزئة الكلمة إلى حروف وأشباه حروف [14]. ومن خلال هذين المثالين يتضح أن هذه الاستراتيجية تعتمد على تقنيات معالجة الصور الرقمية في تحليل وتجزئة الصورة. وأهم التقنيات المستخدمة هي: تحليل الاسقاط العمودي والأفقي Vertical and Horizontal Projection، والمدرج التكراري Histogram، وتنحيف الصورة Thinning، واستخراج المحيط الخارجي Contour Extraction ، وتعليم القطع المتصلة Connected Components Labeling وغيرها.



شكل 2. أمثلة على التجزئة التحليلية

وتستخدم التجزئة التحليلية في أنظمة التعرف غير محدودة الكلمات أي تحتوي على عدد كبير جدا من المفردات، بحيث لا يتأثر معدل تمييز النظام بزيادة عدد الكلمات لأنه يعتمد على الحروف أو أجزاء منها وليس الكلمة. كما أنها تتناسب مع أنظمة التعرف على الكتابة المطبوعة بشكل أفضل من الكتابة باليد، فأنظمة التعرف على الكتابة باليد تتطلب تقنيات ذكية، وهذه الاستراتيجية غير مضمونة للحصول على معدل تجزئة عالٍ. وهي تنقسم إلى نوعين: تجزئة صريحة وتجزئة ضمنية.

1.2.2 التجزئة الصريحة Explicit Segmentation

تسعى إلى تجزئة صورة الكلمة إلى صور أصغر تحتوي على حروف منفصلة ليتم إرسالها إلى مرحلة التعرف. وقد أعطت نتائج جيدة مع الكتابة غير المتصلة، مثل الكتابة اللاتينية المطبوعة وكذلك بعض أشكال الكتابة العربية المطبوعة. لكنها لا تتماشى مع الكتابة باليد بسبب تداخل الحروف والتراكب، حيث يصعب فصل الحروف [4]، [14]، [15]، [16]، [17]، [18]، [19].

2.2.2 التجزئة الضمنية Implicit Segmentation

وتعرف أيضاً بالتجزئة المعتمدة على التعرف Recognition Based Segmentation وتركز على كيفية تجزئة صورة الكلمة إلى سلسلة من الوحدات الصغيرة مثل القطع Segments ويتم استخدام هذه الوحدات في التعرف على الكلمة بدلا من الحروف. فكل وحدة هي جزء من حرف، وبالتالي فإن مجموعة الوحدات المتتالية تنتمي إلى حرف واحد. وفي هذا النوع من التجزئة تتداخل مرحلتي التجزئة والتعرف، وبالتالي فإن نتيجة تمييز الكلمات تتأثر بالأخطاء التي قد تسببها عملية التجزئة التي لا يمكن تقييمها. لذلك فإن معدل التمييز فقط هو ما يتم استخدامه لتقييم أداء النظام [1]،[5]، [6]، [7]، [13]، [20]، [21]، [22]، [23]، [24]، [25]، [26]، [27]، [28]، [29]، [30]، [31]، [32]، [33]، [34].

3. التحديات التي تواجه بناء نظام لتجزئة الكلمات العربية

للتمكن من الوقوف على التحديات التي تواجه تجزئة الكلمات العربية تمت دراسة خصائص الكتابة العربية، وكان الهدف من ذلك تحديد العقبات التي تؤثر على بناء نظام تمييز الكلمات بشكل عام وعملية التجزئة بشكل خاص. فتم حصر التحديات الناتجة عن طبيعة وخصائص الخط العربي سواء المطبوع أو المكتوب باليد والمبينة في (الشكل 3) في الآتي:

أ. **تغير شكل الحرف بتغير موضعه من الكلمة:** فقد يأتي الحرف الواحد على أربعة أشكال، شكل في أول الكلمة، وشكل في وسط الكلمة، وشكل في آخرها، وأخير منعزلا عن أي اتصال.

ب. **تداخل الحروف Overlapping:** يحدث التداخل بسبب الحروف مثل و، ز، ر. فطبيعة هذه الحروف كونها تتصل من جهة واحدة فقط تاركة فراغ بعدها يتيح للحرف التالي لها أن يكتب بحيث يوضع أعلى أو أسفل منها دون أن يلمسها.

ت. **تراكب الحروف Ligature:** بعض الحروف تتراكب أثناء الكتابة بحيث يصعب تجزئتها حيث تصبح كأنها حرف واحد كبير يخفي خط الأساس الذي يربط الحروف معا. ومثال للتراكب، اتصال حرفي اللام والألف ليكونا (لا)، واتصال حرفي الميم والباء ليكونا الشكل (بمـ). ويظهر التراكب عند الكتابة باليد بشكل أكثر من الكتابة المطبوعة آلياً.

ث. **اختلاف أحجام الحروف:** بعض الحروف العربية أعرض من بعض. فمثلاً حرف التاء (ت) ليس بعرض حرف الألف (ا)، فمكان حرف التاء يمكن أن يتسع لكتابة حوال 4 أشكال لحرف الألف .

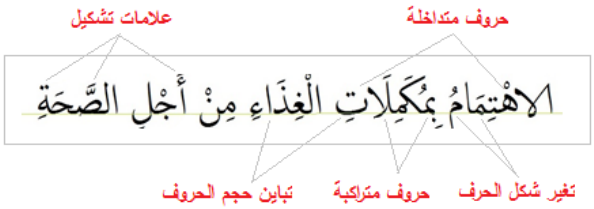
ج. **علامات التشكيل:** تتكون الحروف العربية من جزأين، الشكل الأساسي للحرف، وعلامات التشكيل التي تأخذ أشكال عدة، كالنقطة (.)، والهمزة (ء)، والمد (~). و تحتوي معظم الحروف على نقطة واحدة أو اثنتين أو ثلاث. وفي بعض أنواع الخطوط كالرقعة يتم كتابة النقاط على شكل خطوط. وتوضع علامات التشكيل أعلى أو أسفل أو في وسط الحرف.

بالإضافة إلى ما سبق تواجه أنظمة تمييزالكتابة باليد تحديات أخرى مثل:

- تلامس الحروف أثناء الكتابة.

- تنوع رسم الحرف فوجود العديد من الخطوط العربية سمح برسم الحرف بأكثر من شكل.
- تشوه شكل الحرف بسبب رداءة خط الكاتب.

كما يجب الأخذ في الأعتبار المشاكل الناتجة عن تشوه شكل الحرف بسبب التشويش والضوضاء الناتجة عن جهاز استقطاب البيانات.



شكل 3. التحديات التي تواجه عملية التجزئة

4. الحل : نموذج لخوارزمية تجزئة الكلمات

لقد أمكن الاستفادة من دراسة استراتيجات تجزئة الكلمات في تحديد المكونات اللازمة لبناء نموذج خوارزمية لتجزئة الكلمات العربية. فعملية التجزئة تبحث في كيفية تفكيك صورة كلمة إلى متسلسلة من الصور تمثل حروف هذه الكلمة أو أجزاء منها. لذلك فهي عملية اتخاذ قرار. أي تعتمد على الطريقة التي يتخذ بها قرار بأن الشكل الذي تم فصله هو قرار صحيح أم خاطئ. وكلما كان هذا القرار صائباً قل معدل الخطأ لنظام التجزئة. وهنا يتبين أن الأساس لمعرفة كيفية تجزئة الكلمة تكمن في الإجابة عن السؤالين المهمين التاليين:

- كيف سيتم تمثيل الحرف؟ أو ممّ يتشكل الحرف؟
- كيف سيُقَرَّر بأن شكلٍ ما هو الذي يجرى البحث عنه؟

إن الباحثين الذين حاولوا إيجاد خوارزمية للإجابة عن هذه الأسئلة وجدوا أنفسهم في مصيدة. فصورة الحرف هي شكل يشبه صورة إحدى حروف أبجدية النظام المصمم لتمييزها. ولكن لتحديد أيٍّ منها يشبهه أكثر يجب أولاً تقطيع الحرف من الكلمة، ولتقطيع الحرف من الكلمة يستلزم أولاً معرفة الشكل الذي يجب تقطيعه. أي أن تحديد الشبه وتقطيع الحرف هما عمليتان تعتمد كل منهما على الأخرى. ومن غير المعقول أن يتم البحث عن الشكل الذي يطابق إحدى حروف أو رموز أبجدية النظام دون دمج تفاصيل تركيب هذه الرموز في العملية. علاوة على ذلك، حتى لو كانت مطابقة عدة حروف متتالية صحيحة قد يحكم على الكلمة الناتجة بأنها غير صحيحة إذا وجد خطأ في إحدى هذه الحروف. لذلك فإن قرار التجزئة لا يرتبط مع مطابقة الشكل المتحصل عليه مع إحدى الحروف فقط بل يعتمد أيضاً على قرارات أخرى عامة مثل موافقة الشكل المختار لسياق الكلمة. فمثلاً، يزيد إحتمال معرفة الحرف الثاني من الكلمة إذا تمت معرفة الحرف الذي يسبقه. وبناء على ما سبق، تم التوصل إلى أن الاجابة عن السؤالين السابقين يتضمن حل المسائل الأربعة التالية:

- المسألة الأولى: ما هو شكل المخرجات المتوقعة من عملية التجزئة؟
- المسألة الثانية : كيف يمكن الحصول على المخرجات المطلوبة؟
- المسألة الثالثة : كيف يتم إتخاذ قرار التجزئة؟
- المسألة الرابعة: كيف يتم تقييم أداء عملية التجزئة؟

1.4 المسألة الأولى: ما هو شكل المخرجات المتوقعة من عملية التجزئة؟

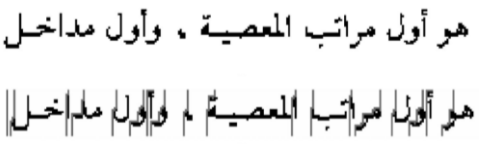
هذه المسألة تناقش الاجابة على السؤال التالي: ممّ تتشكل الكلمة؟، والذي يمكن صياغته بشكل آخر: ما هي الوحدة الأساسية التي تتشكل منها الكلمة؟. ويقصد بالوحدة الأساسية أصغر صورة مجزأة من الصورة الأصلية للكلمة المراد التعرف عليها. من خلال دراسة الأبحاث المختلفة التي اهتمت بتجزئة الكلمة وبالتحديد التجزئة التحليلية تبين أن شكل مخرجات مرحلة التجزئة لم يخرج عن واحدة من الأشكال التالية:

- شبه كلمة.
- حـــــرف.
- شرائـــح.
- قطـــــع.

وهذا الترتيب يبين أن شكل المخرجات متدرج من حيث تعقيد مستوى التجزئة. فالفئة الأولى أقل تعقيداً من حيث التجزئة، والوحدة الأساسية هي الأكبر حجما من باقي وحدات الفئات الأخرى. بينما الفئة الأخيرة هي الأكثر تعقيداً ووحدتها الأساسية هي الأصغر حجماً.

1.1.4 تجزئة الكلمة إلى شبه كلمة

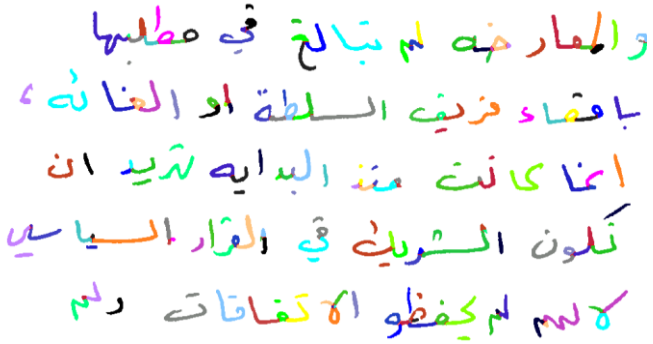
شبه الكلمة Pseudo Arabic Word (PAW) هي المقاطع التي تتكون منها الكلمات العربية. فبالرجوع إلى خصائص الكتابة العربية نجد أنها تتكون من مقاطع، والمقطع هو مجموعة حروف متصلة قد تشكل كلمة كاملة أو جزءاً منها، وسبب انقطاع الكلمة وجود بعض الحروف. وقد تمت الاستفادة من وجود المقاطع في جعل الكلمات العربية قصيرة، أي تحتوي حروف قليلة وهذا يسهل عملية تجزئة الكلمة وكذلك التعرف عليها. ولتجزئة الكلمة إلى شبه كلمة أو مقاطع اتبع الباحثون اسلوبين؛ الأول يعتمد على تحليل الاسقاط العمودي لصورة الكلمة للبحث عن الفراغات بين الحروف، والاسلوب الآخر هو طريقة تعليم القطع المتصلة Connected Components Labeling. وقد استخدم تشينج طريقة تعتمد على ماسك لإيجاد مسارات التجزئة التي تفصل صورة كلمة إلى مقاطع. وهذه الطريقة أوجدت حلاً لمشكلتي التداخل والتراكب التي تتصف بها الكتابة العربية. يبين (الشكل 4) توضيحا لطريقته. وقد أعطت هذه الطريقة نتائج جيدة [4].



شكل 4. مثال على تجزئة نص إلى أشباه كلمات

2.1.4 تجزئة الكلمة إلى حروف

لقد انقسمت الأبحاث التي تهتم بدراسة أنظمة التعرف على الحروف إلى قسمين: القسم الأول ضم مجموعة من الأبحاث التي تخطت عملية تجزئة الكلمة إلى حروف واكتفت بإدخال حروف مجزأة مسبقا إلى نظام التعرف بعضها حروف مطبوعة [15]، [16] وأخرى حروف مكتوبة باليد [14]، [18]. فهي بذلك ركزت على مرحلة استخلاص سمات الحروف ولم تكن التجزئة هدفها. أما القسم الآخر فهي أبحاث جعلت من تجزئة الكلمة إلى حروف هدفا لها. لذلك استخدمت طرق تحليلية للبحث عن النقاط أو الواصلات التي تربط الحروف لتشكل الكلمة. لذلك استخدمت تقنيات تنحيف الكلمة وتقنيات استخلاص محيط الكلمة وتحليل خط الأساس في البحث عن الواصلات التي تربط الحروف. يبين (الشكل 5) نتيجة تجزئة كلمات نص إلى حروف في البحث الذي اعتمد على كلا من تنحيف الكلمة واستخراج المحيط الخارجي [35].



شكل 5. مثال تجزئة نص إلى حروف

3.1.4 تجزئة الكلمة إلى شرائح

تندرج هذه الفئة تحت استراتيجية التجزئة الضمنية وهي الأكثر استخداما، وفيها يتم تجزئة الكلمة إلى شرائح Slides أونوافذ Windows أوإطارات Frames. ويمكن ملاحظة أن جميع الأبحاث ضمن هذه الفئة استخدمت نماذج ماركوف المخفية Hidden Markov Models (HMM) في التعرف على الشرائح المتتالية التي جُزِّئت إليها الكلمة. كما أن هذه الفئة موجهة لأنظمة التعرف على الكتابة باليد. ولقد اعتمدت على الاسقاط العمودي وتحليل خط الأساس والخصائص التركيبية للشرائح لاستخلاص السمات. وقد تباينت في كيفية تقسيم صورة الكلمة بين شريحة بعرض بكسل واحدة وارتفاع الصورة، أو شريحة بثلاثة اطارات أي بعرض 3 بكسلات، بعضها متداخل وأخرى غير متداخل، أو نوافذ باتساع ثابت أو متغير. يبين (الشكل 6) أمثلة لتجزئة صورة إلى شرائح، حيث جزئت الصورة (شكل 6. أ)، إلى شرائح متساوية بينما جزئت الكلمة في (شكل 6. ب) إلى شرائح متغيرة الاتساع، وتم الاعتماد على الاسقاط العمودي لتحديد اتساع النوافذ في (الشكل 6. ج).

(أ) (ب) (ج)

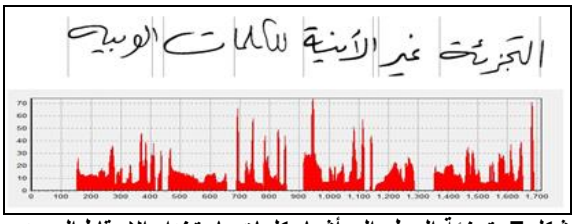
شكل 6. أمثلة على تجزئة الكلمة إلى شرائح

4.1.4 تجزئة الكلمة إلى قطع

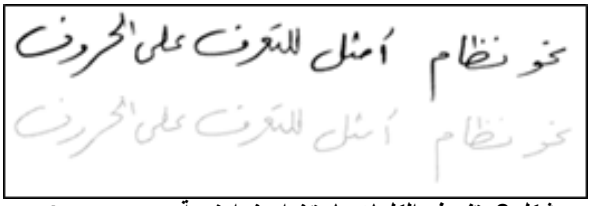
إن تجزئة الكلمة إلى قطع Segments هو الخيار الآخر الأكثر استخداما في الأبحاث في السنوات الأخيرة بعد الشرائح لتجزئة الكتابة باليد. وقد تم استخدام أساليب متنوعة في تحديد شكل القطع التي تتكون منها الكلمة، كما تنوعت طريقة الوصول واستخلاص هذه القطع. ويتم الحصول على هذه القطع عن طريق تجزئة المحيط الخارجي للكلمة أو ايجاد التقاطعات في هيكل الكلمة بعد تنحيفها [5]، [6]، [21]، [22]، [31].

2.4 المسألة الثانية : كيف يمكن الحصول على المخرجات المطلوبة؟

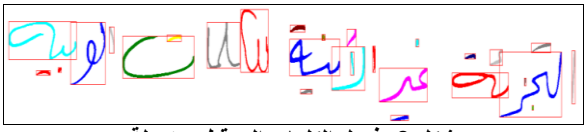
يتم اختيار طريقة التجهيز التي تساعد على الحصول على المخرجات المطلوبة. وأكثر الطرق الشائعة للتجهيز هي: الاسقاط العمودي (شكل 7) والأفقي Vertical/ Horizontal Projection، وتنحيف صورة الكلمة Thinning (شكل 8)، وتتبع المحيط Contour tracing (شكل 9)، وفصل محتويات الصورة إلى قطع المتصلة Connected components Labeling (شكل 10).



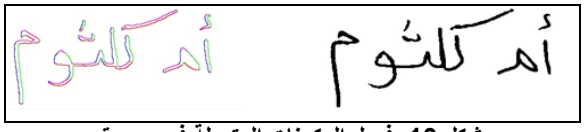
شكل 7. تجزئة السطر إلى أشباه كلمات باستخدام الإسقاط العمودي



شكل 8. تنحيف الكلمات باستخدام خوارزمية Zhang Suen



شكل 9. فصل الكلمات إلى قطع متصلة



شكل 10. فصل المكونات المتصلة في صورة

3.4 المسألة الثالثة : كيف يتم إتخاذ قرار التجزئة؟

بعد الحصول على مخرجات التجزئة باتباع إحدى الأساليب المبينة في الفقرة السابقة، يصبح من المهم الاجابة عن السؤال التالي: هل ما تمت تجزئته صحيح أم لا؟. ويتم الإجابة عن هذا السؤال بإحدى الطرق التالية:

- قوانين تجريبية.
- بناء نموذج باستخدام نماذج ماركوف المخفية.
- استخدام الشبكات العصبية.
- طرق هجين من الطرق السابقة.

4.4 المسألة الرابعة: كيف يتم تقييم أداء خوارزمية التجزئة؟

يتم تقييم أداء خوارزمية التجزئة من خلال مجموعة من المقاييس التي تظهر مقدار خطأ أو صحة عملية التجزئة، ويعتمد التقييم على حجم ونوع قاعدة البيانات المستخدمة للاختبار. ويظهر التقييم مدى كفاءة الطريقة في حل أهم مشاكل الكتابة، مثل تداخل الحروف وتراكبها وتلامسها وغيرها. وبشكل عام، هناك ثلاثة أنواع من الأخطاء مبينة في (الشكل 11)، يمكن استخدامها لقياس أداء نظام التجزئة.

1.4.4 خطأ بسبب التجزئة الزائدة

الخطأ الناتج عن التجزئة الزائدة Over-Segmentation Error يقصد به تمثيل حرف واحد وكأنه أكثر من حرف، مثل حرف السين (س) حيث يتم تقطيعه إلى أكثر من جزء، ويحدث عندما يكون عدد المقاطع المحسوبة من خوارزمية التجزئة أكبر من العدد الحقيقي.

2.4.4 خطأ بسبب التجزئة الناقصة

يحدث خطأ ناتج عن التجزئة الناقصة Under-Segmentation Error عندما يكون عدد المقاطع المحسوبة من خوارزمية التجزئة أقل من العدد الحقيقي. ويظهر هذا الخطأ بسبب تداخل حرفين أو أكثر فتظهر كأنها حرف واحد مثل لم أو لم أو مجموع . تحدث هذه المشكلة في الكتابة العربية بسبب خاصية تداخل أو تراكب الحروف، فيحدث أن عدد نقاط القطع الصحيحة أكبر من المحسوبة.

3.4.4 خطأ التجزئة في غير مكانها

يحدث خطأ التجزئة في غير مكانها Misplaced Segmentation Error عندما يكون عدد المقاطع المحسوبة من خوارزمية التجزئة صحيحاً ولكن حدود المقاطع ليس في مكانه الصحيح.

شكل 11. أخطاء التجزئة الثلاثة

الخلاصة

إن الكثير من الأبحاث تشير إلى مرحلة التجزئة كجزء من نظام متكامل لتمييز الكلمات، لذلك يوجد القليل جداً من التقارير التي تقيم مدى كفاءة هذه المرحلة، أو مدى تأثيرها على النظام ككل. لهذا، فإن هذا البحث ركز على هذه المرحلة الهامة من النظام بشكل دقيق ومفصل. حيث قدم دراسة مسحية لطرق تجزئة الكلمات العربية، تضمنت تقسيم لأهم استراتيجيات تجزئة الكلمات، لتسيهم دراستها والمقارنة بينها وتطويرها. ومن خلال التعرف على هذه الاستراتيجيات تم بناء نموذج لخوارزمية تجزئة يعتمد على تحديد إجابة لأربع مسائله تشمل شكل مخرجات عملية التجزئة، وكيفية الحصول على هذه المخرجات، وكيفية اتخاذ قرار التجزئة وأخيرا كيفية تقييم نتيجة التجزئة. من خلال هذه الدراسة نوصي بتكثيف الأبحاث والتطبيقات العملية التي تدرس تجزئة الكلمات في أنظمة تمييز الكلمات العربية، كما نوصي ببناء منظومات برمجية تستخدم نموذج بناء خورازمية لتجزئة الكلمات العربية الذي تم التوصل إليه في هذه الدراسة.

المراجع

[1] A. Amin, "Recognition Of Printed Arabic Text Based On Global Features And Decision Tree Learning Techniques", The journal of Pattern Recognition society, Val 33,Issue 8, August 2000, P. 1309-1323.

[2] M.S. Khorsheed, "Recognizing Handwritten Arabic Manuscripts Using A Single Hidden Markov Model", Pattern Recognition Letters, Volume 24 Issue 14, October 2003.

[3] C. Leila, B. Mohammed, " Art Network For Arabic Handwritten Recognition System", International Conference on Applied Informatics (ICAI),Vol. 15,Issue 17, November 2009, P. 61-66.

[4] A. M. Zeki, "The Segmentation Problem in Arabic Character", 1st International Conference on Information and Communication Technologies (ICICT 2005), 2005.

[5] R. G. Casey, Eric Lecolinet, "Survey Of Methods And Strategies In character Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, Issue 7, 2002.

[6] A. AL-Shatnawi, F. AL-Zawaideh "Offline Arabic Text Recognition - An Overview", World of Computer Science and Information Technology Journal, Vol. 1, No. 5, P. 184-192, 2011.

[7] R. Al-Hajj, C. Mokbel, "HMM Based Arabic Handwritten Cursive Recognition System", 9th International Conference on Document Analysis and Recognition (ICDAR), Vol. 2, 2009.

[8] L. Souici-Meslatim, M. Sellami, "A Hybrid Approach For Arabic Literal Amounts Recognition", The Arabian Journal for Science and Engineering, Vol. 29, No. 2, October 2004.

[9] A. Cheung, M. Bennamoun, N.W. Bergmann, "An Arabic optical character recognition system using recognition-based segmentation", The journal of Pattern Recognition society, Vol. 34, Issue 8, November 2001, P. 512-233.

[10] A. Benouareth, A. Ennaji, M. Sellami1, "Arabic Handwritten Word Recognition Using HMMs with Explicit State Duration", EURASIP Journal on Advances in Signal Processing, Article ID 247354, 2008.

[11] B.Vaseghi, S. Hashemi, "Farsi Handwritten Word Recognition Using Discrete HMM and Self-Organizing Feature Map", Academic Journal International Proceedings of computer Science & Information Tech, Vol. 38, P. 123, 2012.

[12] J. Chen, H. Cao, R. Prasad, "Gabor Features for Offline Arabic Handwriting Recognition", the 9th IAPR International Workshop on Document Analysis Systems, P. 53-58, June 2010.

[13] L. Rothacker, S. Vajda, G. A. Fink, "Bag-of-Features Representations for Offline Handwriting Recognition Applied to Arabic Script", International Conference on Frontiers in Handwriting Recognition, P. 149-154, 2012.

[14] M. Dehghan, K Faez, "Off-line Unconstrained Farsi Handwritten Word Recognition Using Fuzzy Vector Quantization and hidden markov word Models", The Journal of Pattern Recognition society, Vol. 34, Issue 5, May 2001, P. 1057-1065, 2001.

[15] A. Onat, F. Yildiz, M. Gündüz, "Ottoman Script Recognition Using Hidden Markov Model", World Academy of Science, Engineering and Technology, Vol 14, P. 72, 2006

[16] S. Touj, S. Touj, H. Amiri, "Two approaches for Arabic Script recognition-based segmentation using the Hough Transform", 9th International Conference on Document Analysis and Recognition (ICDAR), Vol. 2, P. 654-658, 2009.

[17] J. Alkhateeb, "Word-Based Handwritten Arabic Scripts Recognition Using Dynamic Bayesian Network", 5th International Conference on Information Technology, 2011.

[18] F. Samoud, S. Maddouri, N. Ellouze, "A Hybrid Method for Three Segmentation Level of Handwritten Arabic Script", The International Arab Journal of Information Technology (IAJIT), Vol. 9, Issue 2, March 2012.

[19] J. AlKhateeb, J. Jiang, J. Ren, S. Ipson, "Component-based Segmentation of Words from Handwritten Arabic Text", International Journal of Computer Systems Science and

Engineering, Vol. 5, N0. 1, 2009.

[20] S. Abdulla, A. Al Nassiri, "Off-Line Arabic Handwritten Word Segmentation Using Rotational Invariant Segments Featuers", The International Arab Journal of Information Technology (IAJIT), Vol. 5, No 2, April 2008.

[21] T. Sari, M. Dellami, "Overview Of Some Algorithms Of Offline Arabic Handwriting Segmentation", The International Arab Journal of Information Technology, Vol.4, No. 4, October 2007.

[22] H. Aljuaid, Z. Muhammad, M. Sarfraz, "A Tool to Develop Arabic Handwriting Recognition System Using Genetic Approach", Journal of Computer Science, Vol 6, Issue 6, P. 619-624, 2010.

[23] N. Zermi, M. Ramdani, M. Bedda," Arabic Handwriting Word Recognition Based on a Hybrid HMM ANN Approach", International Journal of Soft Computing Vol 2, Issue 1, P> 5-10, 2007.

[24] A. Alnsour, L. M. Alzoubady, "Arabic Handwritten Characters Recognized by Neocognitron Artificial Neural Network ", University of Sharjah Journal of Applied Sciences, Vol 3, No. 2, 2006.

[25] S. Touj, N. Ben Amara, H. Amiri, "Arabic Handwritten Words Recognition Based on a planar Hidden Markov Model", The International Arab Journal of Information Technology, Vol.2, No. 4, October 2005.

[26] I. A. Jannoud, "Automatic Arabic Hand Written Text Recognition System", American Journal of Applied Sciences, Vol 4, No 11, P. 857-864, 2007.

[27] R. Al-Hajj, L. Sulem, "Combining Slanted-Frame Classifiers for Improved HMM-Based Arabic Handwriting Recognition", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 31, No. 7, July 2009.

[28] M. Pechwitz, V. Maergner, H. El Abed, "Comparison of Two Different Feature Sets for Offline Recognition of Handwritten Arabic Words", 10th International Workshop on Frontiers in Handwriting Recognition, 2006.

[29] A. M. Elgammal, M. A. Ismail, “A Graph-Based Segmentation and Feature Extraction Framework for Arabic Text Recognition", 6th International workshop on document analysis systems, P622-626, 2001.

[30] M. Dehghan, K. Faez, M. Ahmadi, M. Shridhar, " Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM", Pattern Recognition Society, Vol. 34, Issue 5, May 2001, P. 1057-1065, 2001.

[31] M. Pechwitz, V. Maergner, "HMM Based Approach For Handwritten Arabic Word Recognition Using The IFN/ENIT – Database", Proceedings of the Seventh (ICDAR), 2003.

[32] A. El affar, "Krawtchouk Moment Feature Extraction For Neural Arabic Handwritten Words Recognition", IJCSNS International Journal of Computer Science and Network Security, Vol.9 No.1, January 2009.

[33] J. H. Alkhateeb, "Multiclass Classification of Unconstrained Handwritten Arabic Words Using Machine Learning Approaches", The Open Signal Processing Journal, 2009.

[34] J. Alkhateeb, "Offline Handwritten Arabic Cursive Text Recognition Using Hidden Markov Models And Re-Ranking", Pattern Recognition Letters, Vol. 32, 2011.

[35] M. Omidyeganeh, "A New Segmentation Technique For Multi Font Farsi/Arabic Texts", 13th International Multimedia Modeling Coference, Vol. 1, P. 670-696, 2007.