تقنيات التعرف الضوئي للحروف
معايير الاختيار، طريقة العمل، الإشكاليات، والآفاق المستقبلية

د. أحمد فرج أحمد
جامعة الإمام محمد بن سعود الإسلامية
كلية علوم الحاسب والمعلومات
قسم دراسات المعلومات

تمثل عملية الرقمنة في المكتبات ومؤسسات المعلومات إجراء يتم من خلاله تحويل نصوص مصادر المعلومات التقليدية إلى شكل إلكتروني مقروء من خلال تقنيات الحاسبات الآلية. وتتم عادة هذه العملية عبر المسح الضوئي للنصوص وذلك استنادا على الماسحات الضوئية المتعددة الأنواع والوظائف، مما يؤدى إلى الحصول على نسخة إلكترونية مطابقة تماما للنص الأصلي وهو ما يطلق عليه الرقمنة في شكل صورة. وأما في حالة الرغبة في الرقمنة في شكل نصي فيتم اللجوء إلى إحدى البرمجيات أو التقنيات المتخصصة في التعرف الضوئي على الحروف «OCR «Optical Character Recognition».

وتختص برمجيات التعرف الضوئي على الحروف بالقيام بمجموعة من العمليات والخطوات التي تستهدف التعرف وقراءة نص معين أثناء إجراء عملية المسح الضوئي له، وتقوم بالتعرف على محتويات النص حرف عبر حرف وكلمة عبر كلمة ومن ثم تحويله إلى ملف نصي يتضمن على بيانات ومعلومات مكودة في شكل معيار (American Standard Code for Information Interchange) ASCII أو معيار Unicode والذي يحتل غالباً مساحة أقل من تلك التي يحتاج إليها ملف الصور.

وقد ظهر الجيل الأول من برمجيات التعرف الضوئي على الحروف «OCR» إلى حيز الوجود في الأسواق التجارية في الخمسينات من القرن الماضي، ولم تكن تلك التطبيقات قادرة إلا على التعرف وقراءة النصوص التي تشتمل على أحجام وأشكال حروف معينة ويتم إعدادها خصيصا لهذا الغرض، وكان يطلق عليها «OCR/A». «OCR/B». وبمرور الوقت ومع التطور التقني ظهرت العديد من البرمجيات التي حققت نجاحات لا بأس بها في التعرف الضوئي لنصوص مصادر المعلومات، حيث توافرت لها القدرة على «قراءة» معظم الحروف الهجائية الشائعة باستثناء بعض الحالات التي يكون فيها حجم الحروف صغير نسبيا (حجم ١٤ أو أقل).

ويكمن الغرض المنشود من وراء تلك البرمجيات في تحويل الصورة النصية إلى ملف نصي يمكن قراءته ومعالجته بواسطة تقنيات

الحاسب الآلي وذلك مع تجنب -بقدر المستطاع- نسبة الأخطاء

معايير اختيار برمجيات OCR

يعتمد اختيار المكتبات ومؤسسات المعلومات لبرمجيات التعرف الضوئي على الحروف في الأساس على مدى صلاحيتها للاستخدام والتطبيق، واشتمالها على قواميس متخصصة غنية وثرية بمفردات المحتوى الموضوعي لنصوص مصادر المعلومات التي يمكن معالجتها بواسطة مثل هذه البرمجيات. وتتضمن معايير الاختيار بالإضافة إلى ذلك مجموعة الإجراءات والآليات المتبعة في التعرف على الحروف والسرعة في قراءتها، وأنواع وأحجام وأشكال الحروف التي يمكن التعرف عليها، واللغات التي تحتويها البرمجيات ويمكن التعامل معها، وأساليب وأنماط عرض وإتاحة النصوص بعد اتمام عملية التعرف عليها، وأن تتضمن تقنيات التعرف الضوئي على بعض المهام المتقدمة منها على سبيل المثال إمكانية التدقيق الإملائي والتوافق مع برمجيات التحرير والنشر وأخيراً يعتبر عامل السعر من العوامل التي يمكن أن تؤخذ في الاعتبار.

ويتوافر في ضوء هذه المعايير فئتين من تطبيقات التعرف الضوئي للحروف والتي يتم استخدامها بشكل مستمر في رقمنة مجموعات مصادر معلومات المكتبات الكبرى خاصة الوطنية وهما برنامج «Omni Page Pro» وبرنامج «Word Scan».

وقد ساعد برنامج Adobe Capture على زيادة قيمة هذين البرنامجين وذلك لما له من قدرة عالية على معالجة النصوص ذات التنسيق المعقد، ولعل ذلك يأتي نتيجة استخدام شكل (Portable Document Format) PDF الذي يُمكن من الاحتفاظ بالتنسيق الخاص بالنصوص الأصلية.

وإلى جانب مجموعة المعايير السابق ذكرها والخاصة باختيار برمجيات «OCR»، تتوافر سلسلة أخرى من المعايير الهامة والتي يمكن أخذها في الاعتبار ومنها على سبيل المثال:

١. **الدقة**: ويمكن تقدير معدل الدقة في برمجيات «OCR» من خلال التعرف على نسبة الأخطاء الناجمة عن إجراء القراءة الضوئية للنصوص. ويمكن قياس معدل دقة البرنامج عبر النسبة المئوية للكلمات التي يمكن أن يتعرف عليها البرنامج بشكل صحيح، ويمكن التعبير عن ذلك بالصورة التالية: (X ٪). ويشمل حرف «X» نسبة الكلمات التي يتم التعرف عليها بصورة صحيحة في نص معين وأما الرمز ٪ فهو يمثل النسبة المئوية.

2. **التطابق والتوافق مع أجهزة المسح الضوئي المتنوعة**: من المهم التحقق من نوع أجهزة الماسحات الضوئية التي يمكن أن تتعامل معها برمجيات «OCR»، أي ضرورة التوافق بين جهاز الماسح الضوئي المستخدم من ناحية وبرنامج التعرف الضوئي على الحروف من ناحية أخرى، وفي هذه الحالة من الضروري تطابق وتوافق البرنامج مع معيار «TWAIN» المستخدم من جانب غالبية الماسحات الضوئية.

3. **واجهة المستفيد**: تعتبر الأبعاد الشكلية والتصميمية من العوامل التي ينبغي أن تؤخذ في الاعتبار عند اختيار برمجيات «OCR»، حيث يمكن أن يخلق التصميم الشكلي الجذاب نوع من التحاور والتفاعل بين المستخدم من ناحية وبين البرنامج من ناحية أخرى. ومن بين الآليات التي تحتويها تلك البرمجيات بعض الأزرار التي تسمح عند الضغط عليها بواسطة الفارة القيام بالعديد من المهام، وذلك في أثناء عملية التعرف الضوئي للحروف في نص معين.

4. **القدرة على التعرف على الجداول**: تعتبر إمكانية التعرف الشكلي والهيكلي للجداول خاصية مميزة لبعض فئات برمجيات «OCR»، ولهذا السبب من المفيد أن يكون البرنامج قادراً على اكتشاف وجود جداول في النص ومن ثم إعادة صياغتها في إحدى برمجيات معالجة النصوص.

5. **أشكال الحفظ والتخزين**: من المعروف أن تسجيل البيانات والمعلومات المرقمنة في ملف يجب أن يكون وفقاً لشكل أو قالب أو امتداد معين وذلك بهدف إمكانية القراءة والإطلاع عليه في وقت لاحق، ولذلك فإن برمجيات OCR تتضمن أشكالاً مخصصة لتسجيل ومن ثم حفظ واسترجاع المعلومات في الملفات، ومن أهم هذه الأشكال :

- اسم الملف. DOC خاص ببرنامج معالج النصوص Microsoft Word
- اسم الملف. PDF خاص ببرنامج Adobe Reader التابع لشركة Adobe
- اسم الملف. XLS خاص ببرنامج Microsoft Excel
- إلى غير ذلك...

6. **الالتزام بشكل النص الأصلي**: يحقق الالتزام بشكل النص الأصلي إمكانية الحصول في صورة الكترونية على نفس التكوين الهيكلي والتنظيمي للصفحة الأصلية التي يتم رقمنتها من أعمدة، وجداول، ورسوم إيضاحية وجرافيك..... إلى جانب استنساخ نفس الخصائص النصية من نوع الحرف وحجمه وشكله، وجسم النص، والأسلوب إلى غير ذلك.

7. اللغات: كلما زادت وتعددت اللغات التي يمكن التعرف عليها كلما كانت برمجيات «OCR» أكثر قدرة على التعرف على النصوص المكتوبة بتلك اللغات المتنوعة.

وبعد هذا العرض لأهم معايير اختيار برمجيات «OCR» يستعرض الجدول التالي أهم المنتجين والمنتجات المتعلقة بهذه التطبيقات:

الجهة المنتجة	المنتج	بيئة التشغيل	عنوان URL على الشبكة العنكبوتية
Caere	Omni Page Pro	Windows 95, 98, XP, 2000 NT, Vista	http://www.caere.com
Xerox	TextBridge pro	Windows 95, 98, XP, 2000 NT, Vista	http://www.textbridge.com
IRIS	Irisgem	Windows 95, 98, XP, 2000 NT, Vista, MAC	http://www.irislink.com
IRIS	Readiris pro	Windows 95, 98, XP, 2000 NT, Vista	http://www.irislink.com
Recognia	Recognia	Windows 95, 98, XP, 2000 NT, Vista	http://www.ocr-systeme.de
Read-It OCR pro 6	Olduvai	Macintosh	http://www.olduvai.com

ويمكن أن نستخلص أن تعدد وتنوع تلك البرمجيات يستلزم إعداد وتبني معايير اختيار خاصة ليكون برنامج «OCR» الأكثر كفاءة وعلى الأخص الأكثر توائماً مع احتياجات المكتبات ومؤسسات>

المعلومات. ويوجد اهتمام بالغ على مستوى منتجي برمجيات «OCR» بإمكانية الوصول إلى المنتج المثالي اي نص خالي من الأخطاء بقدر المستطاع. ومع التطور الذي تشهده هذه التقنيات يمكن أن نفترض بأنه في المستقبل القريب يمكن أن تتمكن هذه البرمجيات من تحسين وتطوير كفاءتها بدرجة مرضية.

طريقة عمل برمجيات OCR

يمر عمل برمجيات التعرف الضوئي على الحروف «OCR» في العادة بخمس مراحل أساسية تتم بشكل تتابعي وذلك عند التعرف الضوئي على نص معين وهي:

المرحلة الأولى: ويتم فيها تميز العناصر النصية عن تلك التي تكون في شكل صور وإيضاحيات، وعادة ما تستخدم غالبية البرمجيات المساحات البيضاء بغرض تحديد مواضع النص في ترتيب يمكن قبوله، ومع ذلك من الضروري تقسيم النص بشكل يدوى إلي مقاطع مع تحديد او ترقيم هذه المقاطع أو العناصر النصية، وبشكل عام لا تهتم برمجيات التعرف الضوئي على الحروف ولا تأخذ في الاعتبار الصور وغيرها من الأشكال الإيضاحية التي تتواجد مع النصوص ففي اغلب الأحوال يتم حذفها.

المرحلة الثانية: وتتمثل هذه المرحلة في التعرف على الحروف، ويطلق على منهج التعرف على الحروف تسمية استخلاص الخصائص الطباعية للحروف، حيث يتم التعرف على كل حرف من الحروف المكونة للكلمات وذلك من خلال التحليل الشكلي له إلي جانب مقارنة السمات والخصائص المميزة له طبقاً لمجموعة من القواعد التي تخص كل حرف أو كل نوع أو شكل حرف.

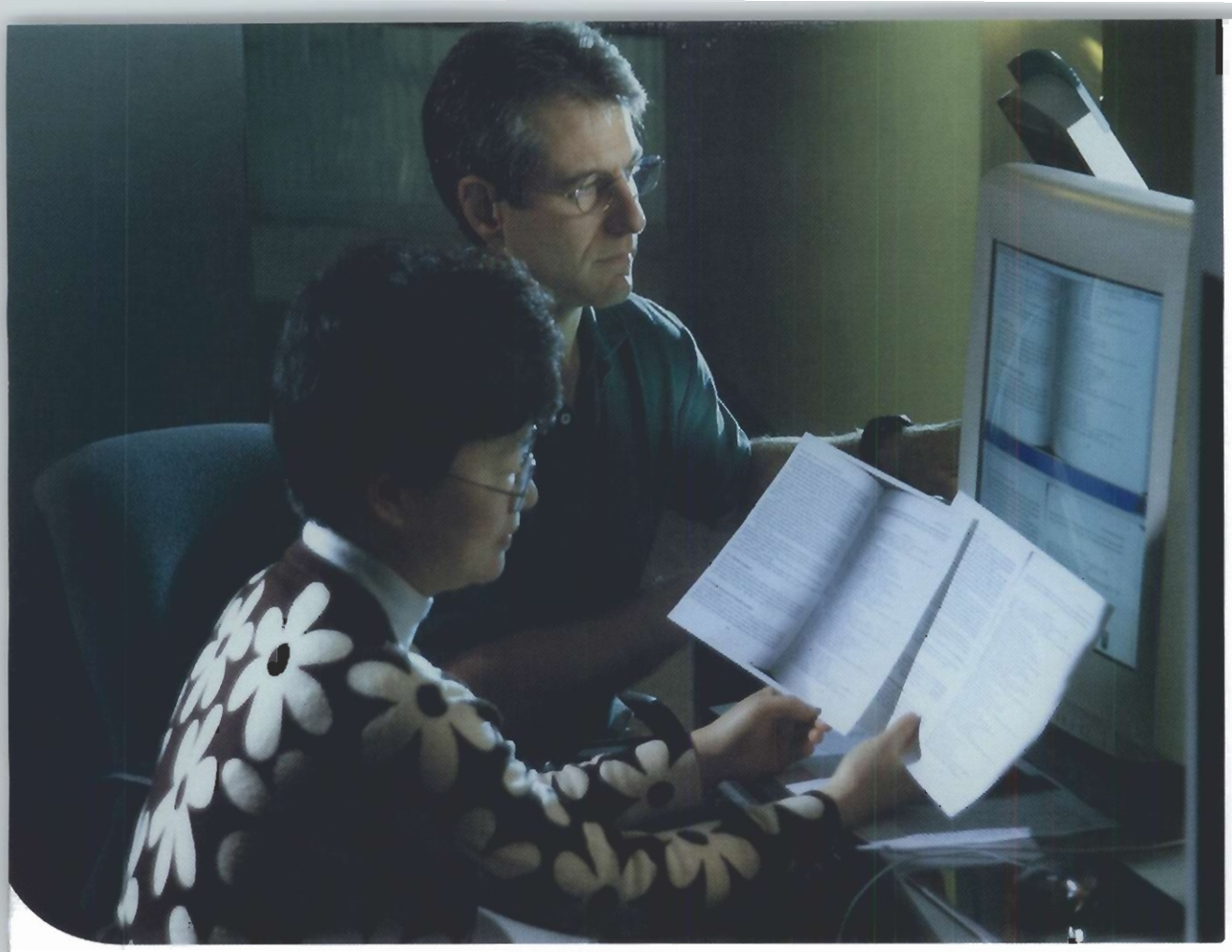
المرحلة الثالثة: ويتم من خلالها التحقق والتعرف على مجموعة الكلمات المكونة للنص الكلمة تلو الأخرى، حيث يتم مقارنة الحروف المكونة للكلمة التي يتم التعرف عليها بكلمات مختزنة داخل قواميس البرنامج وترتبط بنفس لغة النص الأصلي الذي يتم مسحة ضوئياً، ويفضل كذلك أن تكون هذه القواميس متخصصة في المجال الموضوعي للنص الأصلي.

المرحلة الرابعة: وهي تختص بعملية التصحيح، حيث يتم تسجيل نتائج عملية التعرف الضوئي الخاصة ببرمجيات «OCR» في ملف خاص بالبرنامج المستخدم، وبالنسبة للصور والإيضاحات يتم عادة حذفها، ويقوم البرنامج بإبراز الحروف التي لم يتم التعرف عليها إلي جانب سلسلة الحروف المشكوك في درجة صحتها، وهنا تبرز اهمية القيام بعملية التصحيح اليدوي.

المرحلة الخامسة: وترتبط هذه المرحلة بالتنسيق الشكلي للمخرجات، وهنا تقوم برمجيات التعرف الضوئي على الحروف بتسجيل البيانات والمعلومات المتعلقة بالنص المقروء ضوئياً في شكل ملف، ومن أهم الأشكال التي يمكن حفظ الملفات بها: «Word»، «RTF»، «PDF» إلى غير ذلك.

وهناك من يشير إلى أن عملية التعرف الضوئي على الحروف تمر بسلسة من الخطوات والمراحل والتي منها:

1. **التعرف على نوع النصوص:** وذلك بهدف المساعدة في تبني أسلوب للتعرف الضوئي يتماشى مع طبيعة كل نوع من أنواع مصادر المعلومات (جرائد، ومقالات إلي غير ذلك).
2. **تحليل التنسيق الشكلي:** وذلك من خلال تقسيم النص إلي قطاعات أو مناطق مستقلة عن بعضها البعض مثل الأعمدة والصور والمواد الإيضاحية إلي غير ذلك.
3. **التعرف على الأشكال:** وذلك يتكون من تنظيم وترتيب الصور والمواد الإيضاحية في أنواع متنوعة، وهذا النوع يمكنه التعرف وتحديد نوع النص الموجود بجانب الرسم أو الإيضاح وذلك بهدف



توجيه عمليات التعرف نحو نوع الكلمات الأكثر تخصصاً وتحديداً.

4. **تعيين روابط منطقية بين القطاعات المختلفة**: وتسمح هذه المرحلة بإنشاء رابطة تربط بين الصورة والنص المتعلق بتلك الصورة. كما تُمكن من الربط بين نصين من قطاعين مختلفين. وتحليل هذا الربط ينقسم إلى ثلاثة أجزاء: البحث بالكلمات الدالة أو المفتاحية على كل قطاع، وتحديد وتعيين نوع الجرافيك، وأخيراً إنشاء وتصميم الروابط.

5. **التعرف على الحروف**: في أثناء تلك المرحلة يتم تحرير العلامات (الحروف والكلمات) التي تم الحصول عليها في المراحل السابقة ويتم بناء على ذلك عملية التحرير النصي.

وتلعب البرمجيات من نوع «OCR» دوراً هاماً في تحديد الحروف التي يتضمنها نص معين، حيث يشرع البرنامج في البداية بالتعرف على كل حرف بشكل منفصل فردي، ثم يقوم باختيار الشكل الصحيح للكلمة المكونة من مجموعة من الحروف بواسطة تقنيات التدقيق اللغوي للكلمات التي تشتمل على الحروف. ويجب أن تكون هذه البرمجيات قادرة على التمييز بين أنواع وأشكال الحروف وذلك مع مختلف الأساليب والأشكال الطباعية.

الإشكاليات والحلول المقترحة

من المشكلات التي ترتبط باستخدام تقنيات التعرف الضوئي على الحروف، تلك التي تنشأ من النصوص التي تتضمن تباين ضعيف في الألوان كونتراست أو تباين في حروف الطباعة، أو تلك التي تحتوي على متغيرات طباعيه بمعنى استخدام عدة أنواع من الخطوط في نفس النص أو أشكال حروف نادرة أو بعدة لغات أو مخطوطة بخط اليد أو تلك الحروف التي تأخذ أشكال ورسومات جرافيك والتي يكون من الصعب على برمجيات «OCR» التعرف عليها.

ويمكن الجزم - من خلال التجارب العملية - بأن مجموعات الوثائق التاريخية، والصحف والجرائد، والمخطوطات، وأوائل المطبوعات.... يكون من الصعب الحصول على نتيجة مسح ضوئي على درجة عالية من الكفاءة والدقة باستخدام تقنيات «OCR». حيث تشير بعض الإحصاءات إلى أن أقصى معدل لجودة وصلاحية التعرف الضوئي على مثل هذه المصادر يتراوح ما بين (95% إلى 98%) مما يعني أن من (2% إلى 5%) من عمليات التعرف الضوئي على الحروف تحتوي على أخطاء. ونتيجة لذلك، يجب أن يتم مراعاة تصحيح النصوص التي يتم مسحها ضوئياً بطريقة يدوية وبعناية فائقة. ولكن تؤدي تلك العملية إلى رفع تكلفة عملية الرقمنة، حيث أنه إذا كان النص يشتمل على أخطاء فإن جميع الكشافات المرتبطة به سوف تنطوي هي الأخرى على مجموعة من الأخطاء.

وعند استعراض الإشكالية المتعلقة بالصلاحية والدقة المتوقعة من عملية التعرف الضوئي على الحروف والمناهج والأساليب المطبقة بهدف تقويم النتائج التي يتم الحصول عليها، نشير إلى أن برمجيات التعرف الضوئي تقوم بالعديد من العمليات والإجراءات منها التقسيم أو العزل، والتعرف والتدقيق والتحقق اللغوي.

وفيما يتعلق بالتقسيم العزل، فيسمح هذا الإجراء بفصل وعزل العناصر النصية من كلمات وحروف تمهيداً لإجراء عملية التعرف>

الضوئي عليها، ويكون هذا الفصل محدداً على أساس الفراغات أو المساحات البيضاء التي تتواجد بين السطور وكذلك بين الحروف المكونة للكلمات.
ويؤدى تنوع أشكال وأحجام الحروف والتنسيق المعقد إلى عدم السماح بترسيخ البدايات والاستهلاليات المتعلقة بمبدأ الفصل، مما يكون له الأثر في إظهار مساحات بيضاء غير متواجدة بالنص الأصلي الذي يتم رقمنته أو بالعكس إهمال مساحات بيضاء عازلة وفاصلة بين الكلمات تكون متواجدة في النص الأصلي، وتجدر الإشارة إلى أن أي تطبيق غير مناسب لأسلوب ومنهج العزل أو الفصل من الممكن أن يؤدى إلى مجموعة من المشكلات والتي من بينها:

• الدمج الأفقي لمناطق نصية: ويؤدي هذا الخطأ إلى الإدماج بين مجموعة من السطور المتجاورة والمتلاصقة التي تنتمي إلى أعمدة مختلفة مما يؤدى إلى إحداث تأثيرات مباشرة في ترتيب عملية القراءة.
• الدمج الرأسي لمناطق نصية: ويقود ذلك إلى تجميع فقرتين ولكن دون أي تعديل في ترتيب القراءة، ولكن عملية التصحيح لهذا الخطأ تعتبر جوهرية وذلك في حال تصنيف وترتيب النص.
• انفصال أفقي لمناطق نصية: وينتج عنه صعوبات ومشكلات في ترتيب عمليات القراءة.
• انفصال رأسي لمناطق نصية: وهذه الحالة تتشابه إلى حد ما مع الحالة السابقة، ولا ينتج عنها أخطاء بالغة، ويرجع السبب في ذلك إلى أن ترتيب عناصر القراءة يكون متشابها.
• منطقة لم يتم اكتشافها: ويشير هذا الخطأ إلى منطقة نصية لم يتم اكتشافها مثل أشكال الجرافيك والجداول والإيضاحات.....
• جرافيك مختلط مع نص: تقوم بعض برمجيات «OCR» بالتعامل مع الجرافيك على اعتباره نص، وينتج عن هذا الخطأ مجموعات ولقطات غير متجانسة من الحروف داخل النص ونفس الشيء من الممكن حدوثه مع المعادلات الرياضية.
• الاندماج الأفقي مع جرافيك: وذلك يقود مثل الحالة السابقة إلى مجموعة من الحروف غير المقروءة داخل النص.
• الاندماج الرأسي مع جرافيك: وهذه الحالة مشابهة تماماً للحالة السابقة ولكن بطريقة رأسية.

ويمكن القول أن الحصول على نص مقروء ضوئياً بواسطة برمجيات «OCR» صحيح بنسبة 100٪ - أي بدون أية نسبة أخطاء – ما زال هدف ومطلب يرجى الوصول إليه، وتحقيقه لا يعتمد فقط على كفاءة برمجيات التعرف الضوئي للحروف ولكن أيضاً يقتضى في الواقع بعض الاحتياطيات التي ينبغي أخذها في الاعتبار أثناء القيام بهذه العملية ومنها:

• التهيئة والإعداد المسبق للنص المراد رقمنته من أجل الحصول على قراءة جيدة له بواسطة البرنامج ومن بين هذه الإعدادات نذكر ضبط الإضاءة والكونتراست، تباين الألوان مع ضرورة أن يكون مصدر المعلومات كذلك سهل المعالجة.
• إمكانية اللجوء إلى مساعدة برمجيات «OCR» في اختيار المقاييس والخصائص وذلك بهدف التأقلم الجيد وفقاً لطبيعة المحتوى المراد معالجته، مع الأخذ في الاعتبار جودة النص من الناحية الطباعية والورق، ولغة النص، وطبيعة التنسيق المستخدم إلى غير ذلك.
• إمكانية الربط والتوحيد بين العديد من محركات قراءة النصوص المتعلقة ببرمجيات التعرف الضوئي مما يسمح باتخاذ القرار المناسب وذلك بناءا على النتائج المتمخضة لكل منها.

وفي إطار ما سبق عرضه يمكن ملاحظة أن جودة وكفاءة برمجيات «OCR» يمكن أن تعتمد على ما يلي:-
• جودة النص: حيث أن النصوص التي تنطوي على جودة غير كافية مثل اشتمالها على ضعف في تباين الألوان (الكونتراست)، ووجود بقع على النص بشكل ملحوظ، أو تتضمن على تنسيق غاية في التعقيد مثل (عدة أعمدة، ووجود إيضاحات في أماكن غير منتظمة، والحروف تكون في أحجام صغيرة للغاية إلى غير ذلك...)، فكافة هذه العوامل لها تأثير سلبي على عملية التعرف الضوئي على الحروف.
• الاحتياطيات الواجب أخذها في الاعتبار أثناء المسح الضوئي: ويتمثل ذلك في ضبط وضع الصفحة على السطح الزجاجي لجهاز الماسح الضوئي، واختيار مستوى تباين الألوان الكونتراست ودرجة الإضاءة، ودرجات ومستويات وضوح النص، حيث أن هذه العوامل تؤثر بشكل كبير ومباشر على نسبة التعرف الضوئي، وبصفة عامة فإن مقدار التعرف الذي يتم تحقيقه يمكن أن يكون أقل من ٦٠٪ وينتج عن ذلك حجم كبير من البيانات الواجب تصحيحها، وهذا المقدار من الممكن تحسينه من خلال العمل على تطوير وتنمية المواصفات الخاصة بالماسح الضوئي الجيد إلى جانب استخدام بعض التقنيات المتقدمة مثل Intelligent Character Recognition (ICR) وتعمل تلك التقنيات عبر وسائط تتحكم في عملية المسح الضوئي وتسمح بالمساعدة في رفع مستوى درجة التعرف الضوئي للحروف وذلك عن طريق الاستعانة بقواميس متخصصة ولوغاريتمات ترتبط ارتباطا مباشرا بالمحتوى الموضوعي للنص الذي يتم مسحه إلكترونياً.
وبعد هذا العرض للمحاور الرئيسية المتعلقة ببرمجيات التعرف الضوئي على الحروف، يبدو من الأهمية تتبع الآفاق المستقبلية المنظورة والخاصة بتلك التقنيات، ولذا يمثل هذا الهدف نقطة المعالجة الرئيسية التالية.

الآفاق المستقبلية

وكما سبقت الإشارة، الملف النصي الذي يتم إنشاءه بواسطة برمجيات التعرف الضوئي على الحروف يجب في جميع الحالات تصحيحه بطريقة يدوية، وذلك بهدف ضمان الحصول على نص بدون أي أخطاء، وهذا العمل يحتاج إلى فترات طويلة ومجهود للقيام به مما يخلق نوعاً من التأثر السلبي نحو الانجذاب لتطبيق تلك التقنيات مقارنة بالإدخال اليدوي للنصوص أو حتى الرقمنة في شكل صورة، ونتيجة لذلك فإن هناك بعض المشروعات الخاصة بالرقمنة في الولايات المتحدة الأمريكية مثل " JSTOR و «TULIP» والتي لجأت إلى تطبيق معايير وأساليب جديدة من خلال التزويد بنص وفقاً لمعيار «ASCII» غير مصحح ومصحوبا بنصوص في شكل صور مع استخدام وتطبيق برمجيات ذات كفاءة عالية في

التعرف الضوئي على الحروف حيث أن نسبة القراءة الصحيحة للنص قد تصل إلى 90%، 99 في مشروع «ISOTR» حيث انه كل (2000) رمز او حرف يحدث خطأ واحد فقط على الأكثر. وهذا الأسلوب يربط بين ميزتين تتعلق الأولى بالتعرف الضوئي الجيد للنص والثانية ترتبط بالتقديم الشكلي المطابق للنص الاصلي وذلك بفضل استخدامه كصورة وتطبيق الإمكانيات الخاصة بالبحث في النص الكامل.

وتهدف الأعمال الجارية إلى الوصول إلى عملية تعرف ضوئي أكثر كفاءة وفعالية خاصة فيما يتعلق بالبناء الهيكلي والمنطقي لنصوص مصادر المعلومات.

وهناك بعض البرمجيات المتخصصة التي استطاعت التعرف على مختلف القطاعات التي يتضمنها التقنين الدولي الموحد للكتاب «International Standard Book Description» «ISBD» في البطاقة التقليدية للفهرسة وذلك بفضل الحروف الطباعية المعيارية. وبناء على ذلك فإن النص الذي يتم الحصول عليه لم يعد مجرد نص وفق معيار «ASCII» ولكنه ايضا نص يمكن إثرائه من خلال إمكانية اشتماله على بعض اللغات الخاصة بتكويد وترميز النصوص مثل «SGML». «XML». «ODA»

وتجدر الإشارة إلى أن جيل برمجيات OCR والمتمثلة في برمجيات «Pix Tex» وهي مُنتجة بواسطة مؤسسة Excalibure Technologies". ويقوم هذا البرنامج بتطبيق تقنية يطلق عليها التعرف على الأشكال حيث يمكنه التعرف على المتنوعات والعلاقات داخل الأشكال مثل الخصائص او المواصفات للنص والتي تعتمد على استخدام بعض التقنيات المتقدمة مثل ICR «Intelligent Character Recognition». كما أن استعادة المصطلحات الخاصة بالبحث تعتبر من ضمن القدرات المتعلقة بتلك التقنية والتي تم تجريبها بالمكتبة الوطنية البريطانية وتعتبر النتائج الأولية مرضية ولكن بالرغم من ذلك فإن تلك التقنية قد فشلت في أثناء محاولة معالجة التشويشات الناتجة عن رقمنة نص مطبوع بشكل رديء.

المصادر

1. BASAR Ivan. Publications en série électroniques : état d'avancement du catalogue. Nouvelle de la Bibliothèque nationale. Vol. 29, no 5, Mai, 1997. http://www.nlc-bnc.ca/publications/2/p205--9705-f.html 192006/09/.

2. BEQUET Gaëlle. La numérisation des documents patrimoniaux. http://www.culture.fr/culture/conservation/fr/preventi/documents/c13.pdf 142007/04/.

3. Bibliothèque nationale de France. BnF- Information professionnelle La numérisation à la BnF. http://www.bnf.fr/pages/infopro/sib/num_pro.htm 062007/05/

4. BLAIR Rolande. Le programme de numérisation de la Bibliothèque nationale du Canada. Nouvelles de la Bibliothèque nationale. Vol. 28, no. 5, Mai 1996. http://www.nlc-bnc.ca/92//p211--9605-f.html 062005/05/

5. BURESI. Charlette A propos de la numérisation : Notions et conseils techniques élémentaires. (sous la direction des bibliothèques et de la documentation DES/B3. édition imprimée décembre 1998, mise à jour novembre 1999. http://www.sup.adc.education.fr/bib/acti/Num/num2.htm 242007/03/.

6. FRASLIN Jean-Jacques (AMGIT Web). OCR (Optical Character Recognition) : reconnaissance de caractère imprimés, 271999/11/. http://www.amgitweb.com/utilitaires/ocr.htm 242006/03/.

7. HAIGH Susan. L'initiative canadienne sur les hibliothèques numériques a les yeux rivés sur l'avenir numérique des bibliothèques. Nouvelle de la Bibliothèque nationale. Vol. 30, no 6, juin,1998. http://www.nlc-bnc.ca/nl-news/p209--9806-f.html 062006/06/.

8 HAIGH Susan. La reconnaissance optique de caractères (ROC) en tant que technologie de numérisation. Flash Réseau n°37 ISSN 1200- 5304, le 15 novembre 1996. http://www.nlc-bnc.ca/publications/1/p1236--f.html Visité le 2202/07/

9. Ministère de la culture et de la communication. (Fiches de la direction du livre et de la lecture) Numérisation des bibliothèques : L'image numérique : acquisition : Les caractéristiques du fichier numérique. http://www.culture.fr/culture/mrt/numerisation/fr/dll/compress.html 242007/03/.

10. WINSTON Iris. La numérisation des collections : le partenariat entre le Bibliothèque nationale et Industrie Canada. Nouvelle de la Bibliothèque nationale. Vol. 28, no 9, juillet- août, 1996. . http://www.nlc-bnc.ca/nl-news/p217--9609-f.html 182005/06/