برامج التعرف الضوئي علي الحروف العربية : دراسة تقييمية مقارنة لأغراض استرجاع المعلومات

اعداد
د. عادل نبيل شحات
مدرس المكتبات والمعلومات
كلية الآداب ـ جامعة بنها

الملخص:

كثرت في الآونة الأخيرة مشروعات رقمنة مصادر المعلومات العربية، ونشرها مثل: المستودع الرقمي للرسائل الجامعية المصرية، ومشروع رقمنة مصادر المعلومات بدار الكتب المصرية وغيرها من المشروعات العربية، ولكن تم رقمنة هذه المصادر في شكل صور، ولا يمكن البحث في النص الكامل واسترجاعه إلا من خلال الكلمات الكشفية التي تعد محدودة جدا، وبالتالي ظل النص العربي حبيس هذه الصور، ولا يمكن الاستفادة من نصوصه في علميات البحث والاسترجاع، وحُجبت كثير من المعلومات التي يمكن الاستفادة منها.ومن هنا تحاول هذه الدراسة التعرف علي برمجيات التعرف الضوئي علي الحروف العربية، وخصائصها،ومدى دقتها التي يمكن أن تحل هذه المشكلة .

وتعد أهم مشكلات برمجيات التعرف الضوئي مع الحروف العربية هي: سمات وخصائص هذه الحروف المعقدة سواء في بنيتها، أو في طريقة كتابتها.واعتمدت الدارسة المنهج التقييمي بالاعتماد علي أداة من أدوات جمع البيانات وهي: قائمة مراجعة. وتوصلت الدراسة إلي قلة برمجيات التعرف الضوئي علي الحروف العربية التي انقسمت إلي تجارية، ومفتوحة المصدر،ومتاحة علي الإنترنت، ونجد أن برنامجا واحدا فقط وصل نسبة دقته (١٠٠%) في التعرف علي النصوص العربية وهو برنامج (Google Drive OCR)، وأن عدد (٤) برامج وصلت نسبة دقتها (٩٠%) في التعرف علي النصوص العربية. وأوصت الدراسة مؤسسات المعلومات العربية بتطوير برمجيات للتعرف علي الحروف العربية بدقة عالية، وأن توفر هيئات البحث العلمي ومراكز البحوث ميزانيات لتطوير تقنيات التعرف الضوئي علي الحروف العربية .

الكلمات المفتاحية: برامج التعرف الضوئي علي الحروف ـ رقمنة مصادر المعلومات ـ التعرف الضوئي علي الحروف العربية ـ استرجاع المعلومات

القسم الأول: المقدمة المنهجية:

١/٠ تمهيد:

أثر التطور السريع لتكنولوجيا المعلومات والاتصالات علي أنشطة المكتبات ومراكز المعلومات، وفتحت أفاقا جديدة لرقمنة وتحويل مصادر المعلومات المطبوعة إلي شكل إلكتروني مقروء، يمكن تخزينه وحفظه واسترجاعه، مما دفع المكتبات ومراكز المعلومات باقتناء المصادر الرقمية التي ولدت إلكترونيا أو مسحت ضوئيا للصفحات بالاعتماد علي الماسحات الضوئية scanners التي تعددت أنواعها، ووظائفها حيث سهلت الحصول علي نسخ رقمية مطابقة تماما للنص المطبوع، وهو: ما يطلق عليه الرقمنة في شكل صورة، أما في حالة الحاجة إلي تحويلها إلي شكل نصي text فلابد من الاعتماد علي أحد تقنيات التعرف الضوئي علي الحروف "OCR" Optical character Recognition" التي تقوم بالتعرف الضوئي علي الحروف والنصوص أثناء إجراء عملية المسح الضوئي له، أو بعدها حيث تقوم بالتعرف علي محتويات النص حرف بحرف وكلمة بكلمة ومن ثم تحويله إلي ملف نصي text يتضمن علي بيانات ومعلومات مكودة في شكل معيار ((ASCII) standard code for information interchange أو معيار Unicode والذي يحتل غالباً مساحة أقل من تلك التي يحتاجها ملف الصور. [١]

تاريخ استلام البحث : ٣٠ ابريل ٢٠١٩ ، تاريخ القبول : ١ يوليو ٢٠١٩

وترجع أهمية التعرّف الضوئي على الحروفOCR بشكل عام بتحويل النص من شكل صورة إلى مستند نصي، يمكن إجراء مجموعة من التطبيقات عليه مثل: حفظ وأرشفة مصادر المعلومات وتكشيفها والبحث في النص الكامل لمجموعة متنوعة من الكتب، وتخزين نسخة جديدة من الكتب والأرشيفات ورقمنة المستندات الرسمية في المؤسسات والأقسام الحكومية.

حققت تقنيات التعرف الضوئي علي الحروف اللاتينية درجات دقة ممتازة في التعرف الضوئي علي النصوص اللاتينية، ولكن للأسف، هذه التقنيات واجهت مشكلات وصعوبات مع الحروف العربية حيث ما تزال خوارزميات التعرف الضوئي على الحروف العربية ضعيفة، ومجالها التسويقي منعدم، ويرجع ذلك إلى العديد من الأسباب؛ منها نقص التمويل المطلوب للبحث في التعرف على الحروف العربية بالإضافة إلي تعقيد خصائص الحروف العربية، والوصول إلي نص عربي مرقمن عالي الجودة يتطلب عناية خاصة بالإضافة إلي تجهيزات مادية مثل: أجهزة المسح الضوئي وتقنيات عالية مثل التعرف الضوئي علي الحروف OCR التي تسمح بالبحث والتعديل والسماح بعمليات الترجمة والاقتباس، وإلا تصبح نسخة متطورة من تقنية الميكروفيلم (٢)

وتعود أهمية اللغة العربية إلي الاستخدام الواسع في جميع أنحاء العالم، حيث تعد اللغة الرسمية في عشرين دولة أو أكثر إذ يتحدث بها ما يقرب من ٢٣٤ مليون شخص، بالإضافة إلي أنها لغة القرآن الكريم وجميع المسلمين في العالم يتعلمونها لقراءة القرآن الكريم، بالإضافة إلي أنها مهمة بالنسبة للغات الأخرى التي تستخدم الأبجديات العربية مثل الفارسية والأوردية. (٣)

١/١ مشكلة الدراسة

قامت بعض المكتبات ومؤسسات المعلومات العربية في الآونة الأخيرة برقمنة مصادر معلوماتها المطبوعة، بالإضافة إلى بعض الوثائق التاريخية التي يعود تاريخها إلى ما قبل الكمبيوتر من أجل تسهيل الوصول إليها، ولكن تمت هذه العملية في شكل صور وبالتالي لم تتمكن محركات البحث من البحث في نصوصها الكاملة لعدم توافر برامج للتعرف الضوئي علي الحروف العربية بدقة كاملة.

وعلي الرغم من كثرة تطبيقات التعرف الضوئي علي الحروف اللاتينيةOCR التجارية والمجانية المتاحة علي شبكة الإنترنت والتي تتميز بنسبة دقة ١٠٠% في تحويل النصوص من شكل صورة إلي شكل نص، إلا أن الأمر مختلف بالنسبة لتقنيات التعرف الضوئي علي الحروف العربية، حيث تواجه هذه التقنيات العديد من الصعوبات والمشكلات الخاصة بالتعرف الضوئي علي الحروف العربية من حيث التشابه بين كثير من الحروف وطبيعة كتابة هذه الحروف المتصلة، بالإضافة إلي أن الحرف في اللغة العربية يتغير شكله بتغيير مكانه بالكلمة، وتحاول الدراسة التعرف علي تقنيات التعرف الضوئي علي الحروف العربية ومدى دقتها والمشكلات التي تواجهها.

٢/١ أهمية الموضوع :

اتجهت العديد من المكتبات ومؤسسات معلومات العربية نحو مشروعات رقمنة مصادر المعلومات، ونشرها علي شبكة الانترنت مثل: المستودع الرقمي للرسائل الجامعية المصرية، ومشروع رقمنة مصادر المعلومات بدار الكتب الوثائق المصرية، وغيرها من المشروعات العربية؛ لكن تمت رقمنتها في شكل صور مما أثر بالسلب علي عمليات الاسترجاع والبحث في نصوصها الكاملة، ويمكن تخليص أهمية الموضوع فيما يلي:

- توفر الرقمنة في شكل نصي جهود إعادة إدخال مصادر المعلومات علي الحاسب الآلي وتحريرها بالإضافة إلي توفير الوقت والمال.
- تتيح الرقمنة في شكل نصي الإبحار السهل والسريع بين محتويات النص وفقراته بالاستناد على مجموعة الروابط الفائقة hypertext التي يمكن تضمينها في النص.
- يمكن إجراء عملية البحث عن المعلومات باستخدام اللغة الطبيعية للنص التي تمنح ولوجا مباشرا وسهلا لمصادر المعلومات، دون الحاجة لمعرفة مهارات متعمقة حول استراتيجيات البحث.
- إمكانية القيام بعملية البحث المترابط.
- يمكن ترميز نصوص مصادر المعلومات المرقمنة في شكل نصي وفق المعايير GMLأوXML.

٣/١ أهداف الدراسة:

تهدف الدراسة الحالية التعرف علي تطبيقات التعرف الضوئي علي الحروف العربية وتقييمها والتي يندرج تحتها مجموعة من الأهداف الفرعية منها:

- دراسة أنواع برامج وتقنيات التعرف الضوئي علي الحروف العربية.
- تقييم برامج التعرف الضوئي علي الحروف العربية والمقارنة بينها.
- قياس نسبة دقة تطبيقات التعرف الضوئي علي الحروف العربية .
- التعرف علي أشكال مدخلات ومخرجات برامج تطبيقات التعرف الضوئي علي الحروف العربية .

٤/١ تساؤلات الدراسة:

يمكن بلورة مشكلة الدراسة بالسؤال الرئيس التالي: " ما برامج التعرف الضوئي علي النصوص العربية ؟وما دقتها؟ ويتفرع من هذا السؤال الرئيس مجموعة من الأسئلة الفرعية التالية:

١. ما أنواع تقنيات التعرف الضوئي علي الحروف العربية ؟
٢. ما تقييم تقنيات التعرف الضوئي علي الحروف العربية؟
٣. ما نسبة دقة تقنيات التعرف علي الحروف العربية ؟
٤. ما أشكال الملفات المطلوبة لتقنيات التعرف الضوئي علي الحروف العربية؟

٥/١ منهج الدراسة وأدواتها:

أولاً : المنهج

اعتمدت الدراسة في معالجتها لتقنيات التعرف الضوئي علي الحروف العربية على منهجين : أولهما المنهج المسحي الميداني، وهو المنهج الذي يهدف إلي الكشف عن الأوضاع المتعلقة بظاهرة معينة للوقوف علي ايجابياتها وتدعيمها، والوقوع عند سلبياتها ومحاولة إصلاحها ووضع الخطط البرامج اللازمة لذلك[٤]. باعتباره أكثر المناهج ملائمة لدراسة برمجيات التعرف الضوئي علي الحروف العربية ، وما يتعلق بها من بيانات ومعلومات، وكذلك المنهج الوصفي التحليلي فى فحص ودراسة كل برنامج قيد الدراسة مع تطبيق الأسلوب التقييمى الذي يتم الاعتماد عليه فى دراسة وتقييم برامج التعرف الضوئي علي الحروف العربية .

ثانيا: أدوات الدراسة :

اعتمد الباحث في الحصول علي البيانات المطلوبة للبحث علي الأدوات التالية:

١- **رصد الإنتاج الفكري في مجال موضوع الدراسة :** حيث تم رصد الإنتاج الفكري المنشور حول تطبيقات التعرف الضوئي علي الحروف العربية من خلال موقع بنك المعرفة المصري وبوابة اتحاد المكتبات الجامعية المصرية.

٢- **قائمة المراجعة :**

قام الباحث بإعداد قائمة مراجعة، كأداة من أدوات جمع المادة العلمية الخاصة بهذه الدراسة وتم تجميعها تحت مجموعة من المحاور والبنود الفرعية كما يتضح من الجدول التالي:

جدول (١) محاور قائمة المراجعة

م	المحور الرئيس	عدد البنود الفرعية
١	برمجيات التعرف الضوئي علي الحروف العربية	٥
٢	مدخلات برامج التعرف الضوئي علي الحروف العربية	٨
٣	معدل دقة برمجيات التعرف الضوئي علي الحروف العربية	٧
٤	تخزين ملفات التعرف الضوئي علي الحروف العربية	٧
٥	مدى التزام البرنامج بالشكل الأصلي للوثيقة	٥
	الاجمالي	٣٢

وقد قام الباحث بتطبيق قائمة المراجعة علي برامج التعرف الضوئي علي الحروف العربية .

ثانياً: إجراءات الدراسة :

طبقاً للمنهج المستخدم فقد تم تحديد مشكلة الدراسة، ومنها اشتقت الأهداف، ووضعت تساؤلات الدراسة، وعلي أساسها تم تجميع البيانات من مصادرها، وتم تبويبها بما يخدم الدراسة.

٦/١ حدود الدراسة :

تمثلت حدود الدراسة في الحدود التالية:

- **الحدود الموضوعية :** وتتمثل في دراسة وتقييم برامج التعرف الضوئي علي الحروف العربية، ومدى دقتها في تحويل النصوص من شكل صور إلي شكل نصي.
- **الحدود الزمنية :**أجريت الدراسة التقييمية علي برامج التعرف الضوئي علي الحروف العربية التي صدرت تجاريا أو مجانيا علي شبكة الانترنت حتى عام ٢٠١٨م.
- **الحدود المكانية:**طبقت الدراسة التقييمية علي البرامج المنشورة سواء التجارية أو المجانية.

٧/١ مصطلحات الدراسة :

التعرف الضوئي علي الحروف (OCR (Optical character recognition :هي مجموعة من التقنيات والبرمجيات التي تسمح بمسح مستند مطبوع، وتحويله إلي مستند نصي يمكن تخزين بياناته في شكل نصي "text file"؛ ليعاد معالجتها باستخدام برامج معالجة الكلمات حيث توفر هذه التقنية الوقت والجهد وتكلفة إعادة إدخال عدد كبير من النصوص وتقليل نسبة الخطأ في الإدخال(٥)

الماسح الضوئي scanner: جهاز يمكنه التعرف علي الصور والنصوص الموجودة علي الورق المطبوع، ثم يحول الورق المطبوع إلي شكل يسمح للكمبيوتر التعامل معه في معالجة البيانات[٦]

ويعرف أيضا بأنه جهاز يقوم بقراءة وتحرير النصوص المكتوبة بخط اليد، أو المطبوعة، أو الرسومات، أو الرموز إلى تنسيق رقمي للمعالجة أو العرض على شاشة الكمبيوتر، دون التعرف فعليًا على المحتوى[٧].

ترميز الحروف:characters encoding: هو الوحدة الأساسية التي تستخدم في تبادل المعلومات، وهي مجموعة مرتبة من الأحرف يتم تشفيرها حتى يستطيع الحاسب الآلي التعامل معها ومعالجتها، ثم إعادة تحويلها بعد ذلك وطباعتها أو عرضها للمستفدين. [٨]

٨/١ الدراسات السابقة والمثيلة:

تم مراجعة عدد من الأدوات الببليوجرافية لحصر الإنتاج الفكري عن موضوع "التعرف الضوئي علي الحروف العربية " وهى :

١- فهارس مكتبات:

أ- فهرس اتحاد مكتبات الجامعات المصرية .

٢- قواعد البيانات : بنك المعرفة المصري

Egyptian knowledge bank (www.ekb.eg)

٣- دليل الإنتاج الفكري العربي في مجال المكتبات والمعلومات / محمد فتحي عبد الهادي بحلقاته المختلفة وهى من الحلقة الأولى (حتى١٩٧٦م) إلى الحلقة (٢٠٠٨ -٢٠٠٩م)

هذا بالإضافة إلى متابعة ما صدر من إنتاج فكرى في الفترة من ٢٠٠٩م حتى ٢٠١٦م من خلال تصفح قاعدة الهادي للإنتاج الفكري العربي في مجال المكتبات والمعلومات المتاحة على موقع الاتحاد العربي للمكتبات والمعلومات ، والتي تتشابك في تغطيتها مع بعض الحلقات السابقة ، إلا أنها تمتد فى تغطيتها إلى عام ٢٠١٦م ، والمتاحة على الرابط التالي :

http://www.arab-fli.org/index.php?page=43&link=92&sub=93

ومن هذه المراجعة لتلك الأدوات الببليوجرافية تم حصر الدراسات التالية :

دراسة Radwan (٢٠١٨) [٩] التي أكدت أن التعرف الضوئي على النص العربي في المستندات المطبوعة، أو المشاهد الطبيعية يواجه مشكلة صعبة مقارنة بنفس التطبيق على اللغات اللاتينية. وبما أن الطباعة العربية أكثر تعقيدا، والاختلاف بين الأحرف قد يكون طفيفا للغاية، لذا تحتاج برامج التعرف الضوئي علي الحروف العربية المزيد من الخطوات اللازمة في مثل هذه الأنظمة للتعرف علي الحروف العربية.

وتقدم هذه الدراسة تطويراً لأنظمة التعرف الضوئي على الأحرف العربية الموجودة علي لوحات سيارة التراخيص العربية، ونقلها من شريط فيديو بالاعتماد علي خوارزمية معينة لاستخراج أرقام لوحة الترخيص من خلال شبكة عصبية عميقة، للتعرف على الحروف الموجودة علي اللوحة، حيث تم اختبار الشبكة العصبية على البيانات الاصطناعية، واختبارها على مثال حقيقي يشرح يدويا نموذج التعرف على الحروف العربية بنسبة ٩٠٪ من الدقة، في حين حقق نموذج التعرف على الأرقام العربية ٩٤٪ من الدقة،

ثم تقدم الدراسة نظام OCR باللغة العربية للتعرف على النص العربي في المستندات الممسوحة ضوئيًا، حيث تم عمل تصوير Scan لوثيقة تحتوي على أسطر نصية ، وتقطيعها في كل سطر بمفردها ثم إلى كلمات فرعية.

وفي حين أكدت دراسة **Ayyash (٢٠١٦)**[١٠] أن هناك تقدماً كبيراً تم تحقيقه في مجال تقنيات التعرف الضوئي علي الحروف في اللغات غير العربية، عكس اللغة العربية، وتعد مرحلة التجزئة للحروف في هذه الأنظمة خاصة في مجال اللغة العربية أحد أهم المراحل، حيث ركزت الدراسة علي مرحلة التجزئة ومعالجة مشاكل مختلفة ومتنوعة، يمكن أن تظهر خلال هذه المرحلة، بدءا من مرحلة تقسيم النص إلي أسطر ثم إلي كلمات وأجزاء الكلمات والتشكيل، انتهاءً بالحصول علي الحروف، واعتمدت الدراسة علي برنامج ماتلاب (MATLAB) لإنجاز هذا العمل وتم الحصول علي نتائج تصل تقريبا إلي ٩٧% لمراحل التجزئة المختلفة.

أما دراسة **يس (٢٠١٥)**[١١] فقد تناولت متطلبات التحول الرقمي لمؤسسات المعلومات العربية من خلال ثلاثة أقسام، الأول: ناقش طرق التحول الرقمي، بشقيه: التحول الرقمي بواسطة المسح الضوئي " scanning"، والتحول الرقمي بواسطة إعادة الإدخال "Re-Entering"، والثاني: عرض لعملية التعرف الضوئي علي الحروف "optical characters recognition (OCR) العربية والأجنبية، أما الثالث والأخير: ناقش ترميز (تكويد) النصوص بغرض التعرف علي ترميز أو تكويد الحروف العربية.

أما دراسة **zaree (٢٠١٢)**[١٢] فترى أن التعرف الضوئي على الحروف (OCR) هو عملية ترجمة للصور النصية المطبوعة أو المكتوبة إلى نص قابل للتحرير آليًا، حيث تحسن أنظمة التعرف الضوئي على الحروف تفاعل الإنسان على الكمبيوتر في العديد من التطبيقات مثل: قراءة العنوان البريدي للخطابات، وقراءة الفحص التلقائي، ومعالجة الآلات للنماذج ورقمنة المكتبات. وأكدت الدراسة أن خاصية تمييز الحروف العربية من خلال التعرف الضوئي (AOCR) تقع ما بين فئتين إما حروف متصلة أو حروف غير متصلة، ولكل منها خوارزميات التعرف الخاصة بها، وسعت الدراسة إلي تطوير نظام للتعرف الضوئي علي الحروف العربية AOCR متعدد الخطوط ، وذلك باستخدام مجموعة من الخصائص الإحصائية.

أما دراسة **فراج (٢٠٠٨)**[١٣] فقد تناولت معايير اختيار برمجيات التعرف الضوئي علي الحروف OCR التي يمكن أن تعتمد عليها المكتبات ومؤسسات المعلومات في اختيار هذه البرمجيات، وطريقة عمل تلك البرمجيات، وتناولت الدراسة الإشكاليات والحلول المقترحة من أجل الاستفادة من هذه البرمجيات.

أما دراسة**McClean (١٩٩٨)**[١٤] فقد استعرضت نقاط القوة والضعف في مجموعة مختارة من أحدث حزم برمجيات التعرف الضوئي على الحروف (OCR) المستخدمة لرقمنة المحتوى الكامل للوثائق الورقية قبل النشر على الإنترنت، حيث تم تحديد خيارات الرقمنة، والمراحل الرئيسية لعملية التحويل الرقمي لمشروع Eurotext ، وهو مشروع مكتبات إلكترونية مقره المملكة المتحدة؛ لتعزيز الوصول إلى مواد التعلم المتعلقة بالاتحاد الأوروبي؛ ولإفادة الطلاب والمحاضرين من خلال تسهيل الوصول إلى مصدر واحد من الوثائق الرسمية الرئيسية في مجموعة واسعة من مجالات المواضيع باستخدام شبكة الويب العالمية.

<u>تعليق على الدراسات السابقة :</u>

توجد دراسات كثيرة تتناول التعرف الضوئي علي الحروف العربية ولكن في مجال الهندسة والحاسبات الآلية حيث تم التعرض لأربع دراسات علي سبيل المثال وليس الحصر تتناول التعرف الضوئي علي الحروف من الناحية البرامجية وتم حصر دراستان تتناول التعرف الضوئي علي الحروف

بشكل نظري من خلال التعرض لآليات اختيار تلك البرامج وطريقة عمل تلك البرمجيات، وقد تم الاستفادة من تلك الدراسات السابقة في تقييم تقنيات التعرف الضوئي علي الحروف العربية ومدى دقتها والمقارنة فيما بينها.

القسم الثاني : الإطار النظري:

٠/٢ تمهيد:

تختص تقنية التعرف الضوئي على الحروف بقراءة النصوص التي تكون على شكل صورة وترجمتها إلى شكل يمكن للكمبيوتر التعامل معه، وقد بُذلت جهوداً كبيرة علي مدي العقود الماضية لتطوير نظم التعرف الضوئي على حروف اللغات الأجنبية حتى وصلت إلي مستويات دقة ممتازة، وبالتالي لعبت دوراً كبيرا في تحويل مصادر المعلومات المرقمنة في شكل صورة إلي نصوص يمكن تكشيفها واسترجاعها بسهولة، إلا أن الوضع مختلف بالنسبة التعرف الضوئي علي الحروف العربية رغم أهمية اللغة العربية، ويرجع ذلك إلى خصائص الحروف العربية المعقدة والتي تمثل تحدياً كبيراً أمام عملية تطوير نظم التعرف الضوئي علي الحروف العربية[١٥]

وترجع أهمية اللغة العربية إلي كونها واحدة من أقدم اللغات، إذ يتحدث بها كثير من الناس في جميع أنحاء العالم خاصة في الشرق الأوسط وشمال أفريقيا، بالإضافة إلي أنها تعد اللغة الثانية للعديد من الدول الأسيوية التي تعتنق الدين الإسلامي، علاوة على ذلك، وعلى الرغم من هذه الحقائق، لم تحظ البحوث المتعلقة بالتعرف الضوئي علي الحروف العربية بالكثير من الاهتمام؛ إما لصعوباتها أو لنقص التمويل.. الخ، وما إلى ذلك بالإضافة إلي طبيعة الخط العربي وقواعد كتابته. [١٦]

والتعرف الضوئي علي الحروف العربية Arabic optical character recognition (AOCR) هو علم تحويل مستندات الصور النصية العربية المطبوعة أو المكتوبة بخيط اليد إلي نص مشفر يمكن أن تتعرف عليه الحاسبات الآلية ، بهدف مساعدة المؤسسات في رقمنة الأعمال الورقية ليس من أجل الحفظ والتخزين فحسب، ولكن من أجل البحث واسترجاع الكم الهائل من المستندات المخزنة واستبدال البشر في حوسبة الأعمال الورقية وإعادة إدخالها علي الحاسب الآلي مرة ثانية؛ بهدف تسريع وتحسين وخفض التكلفة وتوفير الوقت والجهد [١٧]

١/٢ التعرف الضوئي علي الحروف :

تتيح تقنيات التعرف الضوئي علي الحروف (OCR) إمكانية البحث والتعديل في ملايين الكتب والبحوث إلكترونيا؛ بغرض تسهيل تداولها، بالإضافة إلي إمكانية الربط بين مقاطع النص والوسائط المتعددة الأخرى فيما يعرف بالنص الفائق " Hyper text" .

ويعرف قاموس علم المكتبات والمعلومات علي الخط المباشر " ODLIS" تقنية التعرف الضوئي علي الحروف بأنها "التقنية التي تسمح بالتعرف علي النص المطبوع، أو المكتوب بواسطة الحاسب الآلي بحيث يمكن التعديل والتحرير والبحث فيه".[١٨]

وتعرف أيضا بأنها "برنامج يقوم بتحويل الوثيقة أو الصفحات الممسوحة ضوئيا في شكل صور إلي نص رقمي عن طريق قراءته حرف بحرف وكلمة بكلمة، وتحويلها إلي بيانات أو معلومات يستطيع

الحاسب الآلي التعامل معها، وتكويدها في شكل معيار American standard code for information interchange" (ASCII) بحيث يسهل الوصول إليها(١٩)

ويستخدم التعرف الضوئي على الحروف على نطاق واسع لتحويل أنواع مختلفة من الوثائق مثل: الوثائق الورقية الممسوحة ضوئياً، وملفاتPDF ، أو ملفات النصوص في صور التي تم التقاطها بكاميرا رقمية أو ماسحات ضوئية إلى بيانات قابلة للتعديل، والبحث في نصوص هذه الملفات والتعامل مع تلك النصوص سواء بالنسخ أو التعديل، ومن هنا تبرز أهمية تطبيقات التعرف الضوئي علي الحروف.

٢/٢ أهمية عملية التعرف الضوئي

تحسن نظم التعرف الضوئي على الحروف من كفاءة وفعالية العمل في المكتبات والأرشيفات، خاصة التي تتعامل مع المسح الضوئي للمصادر الكثيرة والكبيرة الحجم، وبالتالي توفر هذه التقنية إعادة كتابة هذه المصادر الكثيرة وتصحيحها، مما يضمن بقاء محتوى المستند سليماً، بالإضافة إلي إمكانية دمجه مع تقنيات أخرى مثل النشر علي الإنترنت وضغط الملفات... وغيرها، ومن هنا تبرز أهمية ومزايا OCR في زيادة تدفق العمل في المكتبات والأرشيفات، إذ توفر علي الموظفين الإدخال والعمل اليدوي، وبالتالي توفير الوقت حيث يمكنهم العمل بشكل أسرع وأكثر كفاءة ويمكن حصر مزايا التعرف الضوئي علي الحروف فيما يلي(٢٠):

١- السرعة:

إن الاعتماد علي التكنولوجيا الحديثة يساعد المؤسسات علي القيام بالمزيد من الخدمات في ظل عدد محدد من الساعات في اليوم، ومنها الاسكانر Scanner وتقنية OCR التي تساعد علي تقليل وقت معالجة المستندات بأكثر من ٨٠٪ من خلال استبدال العمليات اليدوية في إعادة الإدخال ، وزيادة تركيز الموظفين على كفاءاتهم الأساسية.

٢- تقليل التكاليف:

توفر تقنيات OCR تكاليف الأيدي البشرية بالإضافة إلى تكاليف أخرى لا حصر لها يمكن تخفيضها من خلال تطبيق حلول الاسكانر و.OCR وتشمل بعض هذه المجالات: الطباعة والنسخ والمواد الاستهلاكية والصيانة لمعدات المكاتب، وتكلفة المستندات المفقودة، وتكاليف الشحن والتصوير .

٣- الاهتمامات البيئية:

يقلل استهلاك الورق من التأثيرات السلبية علي البيئة، و يساعد في تقليل المخاطر البيئية بطرق لم تكن قد درست من قبل مثل مقدار الوقود المستخدم لشحن الورق إلى المكتبات ؟ ماذا عن نقل المستندات إلى مخازن التخزين خارج المكتبات والأرشيفات؟ ماذا عن إدارة التحكم في المناخ في مناطق تخزين المستندات؟ وغيرها من التأثيرات التي تتعلق بالبيئة .

٤- تقليل الأخطاء:

تزيد نسبة الأخطاء في كثير من الأعمال البشرية اليدوية، سواء كان نسيان الأشياء أو كثرة الأخطاء المطبعية ، أو ضياع وفقدان المستندات. كل هذه أخطاء يمكن أن تحدث للموظفين وغيرها، تكلف المؤسسة تكاليف أخرى عكس الأعمال الآلية التي تقل نسبة الخطأ بها ومنها التعرف الضوئي على الحروف الذي يساعد علي تقليل الأخطاء الكبيرة التي يمكن أن تكلف المؤسسة مئات الآلاف من الأموال سنويًا.

٥- تقليل المساحة:

يستهلك تخزين المصادر الورقية سواء داخل المكتبات أو في مخازن خارجية كثير من المساحة والتكاليف، وبالتالي ضياع كثير من الوقت والجهد في الحفاظ علي تلك المصادر وصيانتها؛ لذا تساعد الرقمنة والتعرف الضوئي على الحروف بتحويل تلك المصادر الورقية إلى معلومات قابلة للاستخدام والتخزين الرقمي وتقليل مساحات التخزين.

٦- توفير الوقت :

تمكن تقنية المسح الضوئي وocr من مسح واستخلاص المعلومات من الوثائق الورقية، وتحويلها إلي شكل قابل للقراءة والتعديل من قبل الحاسبات الآلية بسرعة فائقة، حيث يمكن الحصول علي تلك المعلومات والوثائق في أماكن متعددة ولأنظمة متعددة دون أي تأخير من خلال بضع نقرات بالماوس، بالإضافة إلي توفير المستندات لأولئك الذين يحتاجونها.

٧-الإدارة السهلة:

تساعد تقنية ocr في حل مشكلة تخزين النصوص في شكل صور، وتحويلها إلي محتوى نصي يمكن البحث فيه والاستفادة منه وتخزينه، حيث يمكن أن يتم تلقائيًا تخزين المستندات والاحتفاظ بها بطرق آمنة، وبالتالي لم يعد الموظف مضطرًا إلى تخصيص بعض الوقت لتسجيل الملفات أو حفظها يدويًا في أماكن صالحة بل أصبحت إدارة المستندات سهلة تمامًا.

٨- استرجاع المعلومات

يعد الاستفادة من تنسيقات ملفات pdf التي يمكن البحث فيها بسهولة ميزة أخرى لتقنية ocr ، حيث تستخدم المؤسسات برامج مختلفة لتحويل الصور إلي ملفات نصية بصيغة pdf وكذلك المستندات الورقية إلي ملفات رقمية يمكن البحث فيها بسهولة، بحيث يمكن إدخال تلك الملفات إلي قاعدة بيانات يمكن البحث فيها بسهولة؛ للعثور علي الأسماء والكلمات الرئيسية والعبارات التي يمكن من خلالها الحصول علي المعلومات المطلوبة داخل النص الكامل للملفات.

٣/٢ أنواع أنظمة التعرف الضوئى على الحروف :

انتشرت وتعددت تقنيات التعرف الضوئي علي الحروف، والنصوص المكتوبة بشكل كبير بفضل تطور تكنولوجيا المعلومات والاتصالات، ويمكن تقسيم هذه البرامج وفق ما يلي :

أولاً: من خلال لغة البرنامج:

تنقسم برمجيات التعرف الضوئي علي الحروف إلي قسمين، الأول: برمجيات التعرف علي الحروف الأجنبية، والثاني: برمجيات التعرف علي الحروف العربية .

برمجيات التعرف الضوئي على الحروف الأجنبية:

ترجع بدايات ظهور برمجيات التعرف الضوئي علي الحروف الأجنبية إلي الخمسينات من القرن العشرين، ولم تكن تلك الأنظمة في البداية قادرة علي التعرف علي النصوص إلا التي تشتمل أحجام وأنواع معينة من الحروف يتم إعدادها خصيصا لهذا الغرض، وبمرور الوقت وتطور التقنية ظهرت العديد من البرمجيات التي استطاعت قراءة معظم الحروف الهجائية الشائعة بكفاءة عالية [٢١].

برمجيات التعرف الضوئى على الحروف العربية:

ترجع بدايات التعرف الضوئي علي الحروف العربية إلي بداية التسعينات من القرن العشرين، حيث قام المعهد العالي للعلوم التطبيقية والتكنولوجيا في دمشق عام ١٩٩١م بطلب لتطوير برنامج للتعرف الضوئي علي الحروف العربية إلي المنظمة العربية للتربية والثقافة والعلوم تصل نسبة دقته ٩٩% وقد قامت شركة كولتك "coltec" بتطوير تقنيات التعرف علي الخط اليدوي عام ١٩٩٧ وأطلقت برنامج " القلم الضوئي العربي"

وفي منتصف عام ١٩٩٨م بدأت شركة دلتا "Delta" العمل ببرنامج للتعرف الضوئي علي الحروف العربية، واستخدم البرنامج حلا يرتكز علي تقنية الشبكات العصبية، ومن أشهر الشركات التي عملت في هذا المجال أيضا شركة صخر، التي أصدرت العديد من البرمجيات كان آخرها "Sakhr Arabic OCR" الإصدار الثامن، الذي يستخدم تقنيتين للتعرف الضوئي أولهما: تقنية التعرف العام "Omni Technology " والتي تعتمد علي الذكاء الصناعي؛ للمساعدة في التعرف علي الوثائق مباشرة دون الحاجة إلي تدريب، ثانيهما: تقنية التعليم "Training Technology" لزيادة نسبة الدقة في التعرف علي الحروف العربية .

ولم يقتصر الأمر علي الشركات العربية بل دخلت شركات أجنبية تعمل في هذا المجال طورت تقنياتها للتعرف علي الحروف العربية، حيث أصدر شركة إيريس "I.R.I.S" برنامج "Readiris" بمختلف اصداراته الحديثة، والذي يدعم ما يصل إلي ١٢٦ لغة ومنها الحروف العربية، وقامت شركة جوجل بتطوير تقنيات للتعرف علي الحروف اللاتينية والعربية وهو متاح علي خدمات جوجل google "drive"[٢٢]

ثانيا": من خلال آلية التعرف على الحروف: [٢٣]

١. **نظم التعرف المباشر (online)** وفيه يتم التعرف علي الحروف والنصوص المكتوبة من خلال أشكال ومتجهات vectors كما تمت كتابتها، حيث يتم التعرف علي الحروف والنصوص والكتابات اليدوية مباشرة أثناء الكتابة، باستخدام قلم ضوئي علي لوحة إدخال خاصة متصلة بالحاسب الآلي.
٢. **نظام التعرف غير المباشر(offline)** وفيه يتم التعرف علي الحروف والنصوص بعد إتمام الطباعة والكتابة في شكل صورة ويتم تحويلها إلي نصوص يمكن التعديل فيها.

ثالثاً: من خلال طرق إتاحة البرنامج:

تتعدد إمكانيات ووظائف كل برنامج من برامج التعرف الضوئي علي الحروف من حيث موافقته نظم التشغيل بكافة إصداراتها، والتعامل معه بسهولة، ومنها ما هو متاح علي شبكة الإنترنت ويعمل بالكامل علي الويب ولا يحتاج إلي تحميل وتثبيت علي الحاسب الآلي ويمكن التعامل معه مباشرة online ويمكن تقسيم هذه البرامج وفق ما يلي :

١- البرامج التجارية للتعرف الضوئى على الحروف:

تتنافس الشركات التجارية في مجال تكنولوجيا المعلومات في إنتاج برامج حاسوبية مميزة من أجل تقديم خدمات كثيرة للمؤسسات والمستفيدين، وأيضا تحقيق ربح مادي يساعدها علي البقاء والاستمرار، خاصة في الآونة الأخيرة، حيث ظهرت برامج تجارية كثيرة، حيث أنتجت شركة Abby software Ltd برنامج للتعرف الضوئي علي الحروف باسم Abby finereader، وقامت شركة صخر الخليج لتقنية المعلومات بإنتاج برنامج للتعرف الضوئي علي الحروف باسم صخرsakhr وغيرها من الشركات .

٢- برامج مفتوحة المصدر للتعرف الضوئى على الحروف open source

يشير مصطلح برمجيات المصدر المفتوح open source software إلي تطبيقات الكمبيوتر، وأنظمة التشغيل التي تم إصدارها بموجب شروط تسمح للمستخدمين باستخدام البرنامج، أو تعديله، أو إعادة توزيعه بأي طريقة يرونها مناسبة، دون مطالبة المستخدمين بدفع رسوم لمنشئي البرنامج، وذلك من خلال توفير شفرة وكود البرنامج التي تجعل وظيفة البرنامج متاحة للفحص والتغيير. (٢٤)

وصدر العديد من أنواع برامج التعرف الضوئي علي الحروف، ويختلف معدل دقة أي برنامج للتعرف الضوئي علي الحروف من ٧١% إلي ٩٨% لكن القليل منها مفتوح المصدر ومجاني مثل: برنامج Tesseract وهو أحد البرامج مفتوحة المصدر ويمكن استخدامه في تطبيقات أخرى (٢٥)

٣- البرامج المباشرة ONLINE المتاحة علي الانترنت

أصبحت بيئة الانترنت بيئة خصبة لتقديم كافة الخدمات للشركات والمستفيدين، منها خدمات ocr عبر الإنترنت التي تتسم بالسهولة بشكل كبير في استخراج النص من الصور وملفات pdf ..الخ، وهي خدمات مجانية من حيث التكلفة، ويمكن استخدام هذه البرامج (OCR) عبر الانترنت دون الحاجة إلي اشتراك.

وظهرت تطبيقات الويب كمصطلح جديد في عالم التقنية، حيث يمكن تشغيل تلك البرمجيات عن طريق الويب عبر تقنيات الويب ٢ التي ظهرت فكرتها مع مطلع العام ٢٠٠٦، حيث تقدم تطبيقات الويب للمستخدم برمجيات صغيرة تقوم بعمل مشابه لعمل البرمجيات الموجودة علي الحاسب الآلي.

٤/٢ العوامل التي تؤثر في عملية التعرف الضوئى على الحروف

توجد مجموعة من العوامل تؤثر في دقة التعرف الضوئي علي الحروف عند القيام بعملية رقمنة للوثائق الورقية، وتحويلها إلي نصوص منها(٢٦):

١ - جودة النص الأصلى:

مما لا شك فيه أن جودة طباعة النص الأصلي، أو مصدر المعلومات تؤثر بشكل كبير علي جودة عمليات الرقمنة، وبالتالي جودة التعرف الضوئي علي الحروف، أما إذا كانت النسخة الأصلية منخفضة الجودة مثل: أن يكون الحبر فاتح جدا، والورق غير مسطح وأبيض، أو يوجد خلفية للنص فإن محرك البحث ocr سيواجه صعوبة في تمييز الحروف.حيث إن جودة الوثيقة التي يتم الحصول عليها بواسطة الكمبيوتر للتعرف الضوئي علي الحروف ocr سيكون لها تأثير علي جودة مخرج ocr إذا كان المستند الأصلي به بعض العيوب مثل :

- تجاعيد أو تمزق أو تضرر الوثيقة الورقية
- تغير لونها
- تشوه النص
- حبرها منخفض التباين أو الملون (اللون الأزرق والأحمر والأرجواني) أما الحبر الأسود يوفر تباين أعلي

٢- جودة المسح الضوئى:

توفر الماسحات الضوئية تمثيلا رقميا للصفحات المطبوعة، وتتوقف جودة مخرجات تعرف الضوئي علي الحروف علي جودة المدخلات لمحرك بحث ocr، وإن كانت الماسحات الضوئية تسمح للمستخدم

بتغيير خصائص المسح الضوئي مثل: درجة السطوع، ودرجة التباين. فإذا كان التباين أقل من ٢٠٠ نقطة في البوصة يؤدي ذلك إلي نتائج غير جيدة، أما إذا كان التباين أعلي من ٦٠٠ نقطة في البوصة يؤدى إلي زيادة حجم الملف المخزن دون الحصول علي نتائج أفضل، لكن من الأفضل أن يكون التباين ٣٠٠ نقطة في البوصة عند المسح الضوئي للنصوص. وتعتمد دقة مخرجات التعرف الضوئي علي الحروف أيضا علي درجة لون مسحها بالإسكانر، فإذا كان اللون الأسود مقابل الأبيض تكون النتائج جيدة، وإن كان غير ذلك فإن النتائج تكون غير مرضية.

٣- محرك بحث ocr

يعد محرك بحث برنامج التعرف الضوئي علي الحروف ocr المسئول عن ترجمة صور النصوص إلي أحرف يمكن الحاسب الآلي أن يتعرف عليها، حيث يتم تزويد محركات بحث ocr بأوامر برمجية من قبل المبرمجين حول كيفية تمييز الحرف الصحيح من مجموعة احتمالات بناءً علي الصورة المقدمة لمحرك ocr وإن جودة هذه الأوامر لها بعض التأثير علي دقة إخراج وتحويل النصوص من الصور .

٥/٢ معايير اختيار برمجيات التعرف الضوئي على الحروفocr

يعتمد اختيار برمجيات التعرف الضوئي علي الحروف علي مدى صلاحيتها للاستخدام والتطبيق واشتمالها علي قواميس متخصصة غنية وثرية بمفردات المحتوى الموضوعي لنصوص مصادر المعلومات التي يمكن معالجتها، ومدى قدرة البرنامج علي التعرف علي الحروف والسرعة في قراءتها، بالإضافة إلي أنواع وأحجام وأشكال الحروف التي يمكن التعرف عليها، واللغات التي تحتويها البرمجيات، وطرق عرض النصوص بعد التعرف الضوئي عليها، وإمكانية التدقيق الإملائي والتوافق مع برمجيات التحرير والنشر. ويمكن تلخيص هذه المعايير في النقاط التالية:

١- **الدقة :** وتعد الدقة في التعرف الضوئي علي الحروف، وتقليل نسبة الأخطاء الناتجة عن القراءة الضوئية من أهم معايير اختيار برمجيات التعرف علي الحروف OCR، ويمكن قياس معدل دقة البرنامج بالنسبة المئوية للكلمات التى يمكن أن يتعرف عليها البرنامج بشكل صحيح.

٢- **التطابق والتوافق مع أجهزة المسح الضوئي المتنوعة :** ويقصد به ضرورة توافق برنامج التعرف الضوئي علي الحروف مع جهاز الماسح الضوئي المستخدم، وفي هذه الحالة من الضروري تطابق وتوافق البرنامج مع معيار " TWAIN" المستخدم من جانب غالبية الماسحات الضوئية.

٣- **واجة المستفيد:** تعد واجهة وتصميم البرنامج من العوامل التي ينبغي أن تؤخذ في الاعتبار عند اختيار برمجيات "OCR"، حيث أن واجهة البرنامج قد يكون لها دور فعال في التفاعل بين المستفيد والبرنامج، بالإضافة إلي إمكانات البرنامج وسهولة استخدام وظائف البرنامج.

٤- **القدرة علي التعرف علي الجداول:** من المميزات لبعض فئات برمجيات **"OCR"** التعرف الشكلي والهيكلي للجداول وإعادة صياغتها في برمجيات معالجة النصوص .

٥- **أشكال الحفظ والتخزين:** من المهم أن يسمح برنامج Ocr أن يحفظ المعلومات المرقمنة والبيانات في ملفات لها امتداد معين، يسمح بإمكانية القراءة والاطلاع عليها في وقت لاحق ومن أهم هذه الأشكال :

- امتداد doc أو docx خاص ببرنامج معالجة النصوص
- امتداد pdf خاص ببرنامج Adobe Reader
- امتداد xls/ xlsx خصا ببرنامج Microsoft excel

٦- **الالتزام بشكل النص الأصلي:** من أهم ما يميز برنامج عن آخر، مدى التزامه بشكل النص الأصلي والحصول علي نفس التكوين الهيكلي، والتنظيمي للصفحة الأصلية التي يتم رقمنتها، إلي جانب استنساخ نفس الخصائص النصية من نوع الحرف وحجمه وشكله وجسم النص والأسلوب إلي غير ذلك.

٧- **اللغات:** يقاس قدرة وكفاءة البرنامج كلما زادت وتعدد اللغات التي يمكن التعرف علي النصوص المكتوبة بتلك اللغات المتنوعة[٢٧].

القسم الثالث : الدراسة التقييمية :

١/٣ تقسيمات وأنواع برمجيات التعرف الضوئي على الحروف العربية

تنقسم برمجيات التعرف الضوئي علي الحروف إلي عدة تقسيمات مختلفة، وعند حصر برمجيات التعرف الضوئي علي الحروف العربية تبين أن عددها ٢٠ برنامجا تم تصنيفهم كما يتضح من الجدول التالي:

جدول رقم (٢) أنواع برمجيات التعرف الضوئي علي الحروف العربية

م	نوع البرنامج	العدد	النسبة
١	مباشر علي شبكة الانترنت	٩	45 %
٢	تجاري	٧	35 %
٣	مفتوح المصدر	٤	20 %
		٢٠	100 %

يتضح من الجدول السابق أن برمجيات التعرف الضوئي علي الحروف العربية بلغ عددها (٢٠) برنامجا، وهذا عدد قليل جدا مقارنة ببرامج التعرف الضوئي علي الحروف الأجنبية، وغير مناسب للحروف العربية ويبين الجدول أن هذه البرمجيات انقسمت إلي ثلاثة أنواع، حيث جاءت في المرتبة الأولي برمجيات التعرف الضوئي علي الحروف العربية علي شبكة الإنترنت online بإجمالي عدد (٩) برامج بنسبة ٤٥%، وجاء في المرتبة الثانية برمجيات التعرف الضوئي علي الحروف العربية التجارية بعدد (٧) برامج بنسبة ٣٥% وجاء في المرتبة الثالثة برمجيات التعرف الضوئي علي الحروف العربية مفتوحة المصدر بعدد (٤) برامج بنسبة ٢٠% .

مما سبق يتضح أن النسبة الأكبر من برمجيات التعرف الضوئي علي الحروف العربية متاحة علي شبكة الانترنت، لما تتمتع به من سرعة الانتشار والوصول إلي عدد كبير من المستفيدين، بالإضافة إلي تطور الويب، وتعود قلة عدد البرمجيات التجارية إلي عدم اهتمام مشروعات الرقمنة العربية بتقنيات ocr العربية وعدم سعيهم إلي شراء أو طلب هذه التقنيات والاكتفاء بتحويل تلك المصادر إلي صور مرقمنة بالماسحات الضوئية وحفظها إلكترونيا، أو في نظم إدارة المحتوى الرقمي في أشكال PDF

٢/٣ منصات تشغيل برمجيات التعرف الضوئي على الحروف

يحتاج برنامج التعرف الضوئي علي الحروف إلي نظام تشغيل يتوافق معه، حتى يستطيع المستخدم تشغيله، والجدول التالي يحصر نظم التشغيل التي تتوافق مع برمجيات التعرف الضوئي علي الحروف العربية :

جدول رقم (٣) منصات تشغيل برامج التعرف الضوئي علي الحروف

م	نظام التشغيل	عدد البرامج	النسبة
١	Windows	١١	%٥٥
٢	Web application/Web host	٩	%٤٥
٣	Linux	٥	%٢٥
٤	Mac	٣	%١٥

يقصد بمنصة التشغيل البيئة التي يتم فيه تشغيل البرامج التطبيقية، وقد تكون المنصة عتاداً أو نظام تشغيل أو حتى متصفح ويب أو برمجية أخرى، بمعنى أكثر شمولاً الموقع الذي تعمل فيه البرمجيات، وهو الوسيط بين المستخدم وجهاز الحاسب الآلي، وهو المسئول عن تشغيل البرامج التطبيقية للحاسب الألي، لذا يسعى كل مصممي ومطوري البرامج أن تتوافق برامجهم مع نظم تشغيل الكمبيوتر، حتى يمكن تشغيلها بسهولة.

ونلاحظ من الجدول السابق أن نظام تشغيل ويندوزwindows هو أكثر نظم التشغيل المناسبة لبرمجيات التعرف الضوئي علي الحروف العربية، حيث يتوافق مع عدد (١١) برنامجا ما بين برامج تجارية ومفتوحة المصدر بنسبة (٥٥%)، ويعود ذلك إلي أنه أكثر نظم التشغيل انتشاراً، بالإضافة إلي بيئته الرسومية السهلة والتي لا تحتاج إلي خبرة تقنية في التعامل معه. وجاء في المرتبة الثانية منصات تشغيل مباشرة (الويب) web host أو web application حيث بلغ عدد البرامج التي يتم تشغيلها علي الويب (٩) برامج بنسبة (٤٥%) حيث يمكن تشغيل البرنامج دون الحاجة إلي تحميله وتثبيته علي الحاسب الآلي الشخصي للمستخدم، لكن يمكن تشغيله مباشرةonline. وجاء في المرتبة الثالثة نظام التشغيل linux حيث بلغ عدد البرامج المتوافقة معه (٥) برامج بنسبة (٢٥%). وجاء في المرتبة الأخيرة برنامج التشغيل mac حيث بلغ عدد البرامج المتوفقة معه (٣) برامج، وفيما يلي بيان تفصيلي بالبرمجيات التي تستخدم نظم التشغيل .

١- نظم تشغيل برمجيات التعرف الضوئى على الحروف العربية التجارية :

تم حصر برمجيات التعرف الضوئي علي الحروف العربية التجارية ونظم تشغيلها في الجدول التالي:

جدول رقم (٤) نظم تشغيل برمجيات التعرف الضوئي التجارية علي الحروف العربية

م	اسم البرنامج	بيئة التشغيل
١	Automatic page reader 3.01 form skhar	Windows
٢	Abby finereader	Windows, linux, mac os
٣	Readiris	Window,linux, Mac OSX
٤	Leadtools high lever ocr toolkit	Windows x86/x64, Linux, iOS, macOS, and Android
٥	Sakhr ocr gold edition	Windows 7 & 8 (32 & 64 bit)
٦	Verus professional multilingual ocr	Windows
٧	I.R.I.S.OCR	Windows, linux

يتضح من الجدول السابق أن نظام التشغيل ويندوز windows هو أكثر نظم التشغيل المطلوبة لتشغيل برمجيات التعرف الضوئي علي الحروف العربية التجارية، وهذا أمر طبيعي لأنه أكثر نظم التشغيل انتشارا وسهولة جاء بعده نظام تشغيل لينكس linux حيث نجد عدد (٤) برمجيات متوافقة مع

نظام تشغيل لينكس ويرجع ذلك لصعوبته وعدم موافقته مع جميع الأجهزة وعدم دعمه بعض البرامج التطبيقية مثل برامج معالجة الكلمات. وجاء بعده نظام تشغيل ماك mac حيث نجده متوافق مع عدد (٣) برامج، ويرجع ذلك أن نظام mac غير متوافق مع جميع الأجهزة غير آبل بالإضافة إلي أنه لا يدعم كافة التطبيقات مثل ويندوز.

٢- نظم تشغيل برامج التعرف الضوئي علي الحروف العربية مفتوحة المصدر open source

تم حصر برمجيات التعرف الضوئي علي الحروف العربية مفتوحة المصدر open source ونظم تشغيلها في الجدول التالي:

جدول رقم (٥)نظم تشغيل برمجيات التعرف الضوئي علي الحروف العربية مفتوحة المصدر

م	اسم البرنامج	نظام التشغيل
١	Tesseract ocr	Windows, Linux, OS/2
٢	Arabic ocr image segmentation	Windows
٣	GOCR	Windows, Linux, OS/2
٤	Cuneiform	Windows

يتضح من الجدول السابق أن كل برمجيات التعرف الضوئي علي الحروف العربية مفتوحة المصدر تتوافق مع نظام تشغيل ويندوز windows ثم يليه نظام تشغيل لينكس linux

٣- الانترنت منصة تشغيل برامج التعرف الضوئي على الحروف العربية:

توفر الانترنت بيئة تشغيل تناسب البرامج التطبيقية حيث توفر الخدمات الأساسية، مثل واجهة المستخدم الرسومية، ونظام الملفات الافتراضية(virtual file system)، والوصول إلى الرقابة الإدارية والإمكانيات لتطوير ونشر تطبيقات الإنترنت. ولأنه نظام تشغيل على الإنترنت، فتنفيذه يتم ضمن متصفح ويب، وهو ليس نظام تشغيل حقيقي ولكنه المدخل لتطبيقات الويب المختلفة.

جدول رقم (6) برامج التعرف الضوئي علي الحروف العربية

م	البرنامج	بيئة التشغيل	الموقع
1	OCR Online Using The Google Drive	الويب	https://www.google.com/drive/
2	Newocr.Com	الويب	https://www.newocr.com/
3	Finereaderonline.Com	الويب	https://finereaderonline.com/en-us/Tasks/Create
4	OCRconvert.Com	الويب	https://www.ocrconvert.com/
5	I2ocr.Com	الويب	http://www.i2ocr.com
6	Totext.Net	الويب	http://www.to-text.net/
7	Verypdf.Com	الويب	http://www.verypdf.com/online/ocr-converter.php
8	Online OCR By A9t9.Com	الويب	https://ocr.space/
9	ABBYY FineReader Online	الويب	https://finereaderonline.com/en-us/Tasks/Create

يتضح من الجدول السابق أن عدد برمجيات التعرف الضوئي علي الحروف العربية التي يمكن تشغيلها علي الإنترنت بلغ (٩) برامج تستخدم منصة تشغيل الانترنت لبث خدمات البرنامج، حيث يمكن الوصول إلي البرنامج من خلال موقع البرنامج .

٣/٣ أنواع الملفات التي تدعمها برامج التعرف الضوئي علي الحروف العربية

يقصد بنوع الملف بأنه امتداد الملف الذي يحدد البرنامج التطبيقي المطلوب لتشغيله، والجدول التالي يوضح أنواع الملفات التي تدعمها برمجيات التعرف الضوئي علي الحروف العربية :

جدول رقم (7) أنواع الملفات التي تدعمها برامج التعرف الضوئي علي الحروف العربية

م	نوع ملف النصوص	عدد البرامج التي تدعمه	النسبة
١	Image	١٥	%٧٥
٢	PDF image	٩	%٤٥

يتضح من الجدول السابق أن نسبة (٧٥%) من برمجيات التعرف الضوئي علي الحروف العربية تدعم ملفات النصوص في شكل صور، سواء كان امتدادها JPEG, JFIF, PNG, GIF, BMP, PBM, PGM, PPM, PCX وتتعرف علي نصوصها، وتعد الصور الرقمية أساس التحول الرقمي وبدايات الكتب الرقمية، وأن معظم مشاريع الرقمنة اعتمدت علي صيغ الصور في رقمنة مصادر المعلومات. ونلاحظ أن نسبة (٤٥%) من برامج التعرف الضوئي علي الحروف العربية تدعم ملفات النصوص في شكل PDF Image ، وتتعرف علي نصوصها التي كانت في الأساس ملفات صور تم تحويلها إلي شكل PDF.

جدول رقم (8)الملفات التي تدعمها برامج التجارية للتعرف الضوئي علي الحروف العربية

م	اسم البرنامج	ملفات النصوص في شكل PDF image	ملفات النصوص في شكل صورة IMAGE
١	Abby finereader14	√	√
٢	Readiris.16	√	√
٣	Sakhr ocr gold edition	√	√
٤	Automatic page reader 3.01 form skhar	√	√
٥	Leadtools high lever ocr toolkit	√	√
٦	Verus professional multilingual ocr	√	√
٧	I.R.I.S.OCR	√	√

نلاحظ من الجدول السابق أن كل برمجيات التعرف الضوئي علي الحروف العربية التجارية تدعم صيغ الملفات في شكل صورة Image وتدعم الملفات PDF Image، ويعود ذلك أن هذه البرامج لمؤسسات تجارية تنافسية، لديها من الكوادر البشرية والتمويل ما يسمح لها بتطوير برامجها؛ لتناسب كل الاحتياجات من أجل الربح والمنافسة والاستمرار.

جدول (9)الملفات التي تدعمها برامج مفتوحة المصدر للتعرف الضوئي علي الحروف العربية

م	اسم البرنامج	ملفات النصوص في شكل PDF image	ملفات النصوص في شكل صورة IMAGE
١	Tesseract ocr	×	√
٢	Arabic ocr image segmentation	×	√
٣	**GOCR**	√	√
٤	**Cuneiform**	×	√

نلاحظ من الجدول السابق أن كل برمجيات التعرف الضوئي علي الحروف العربية مفتوحة المصدر تدعم صيغ الملفات في شكل صور Image. أما ملفات النصوص في شكل PDF Image فلا يدعمها إلا برنامج واحد فقط وهو برنامج "Arabic ocr image segmentation وهو تابع لموقع sourceforge هو موقع مركزي لمشاريع البرمجيات الحرة والمفتوحة المصدر بدعم من شركة VA Software الأمريكية وهي تدعم مشروعات المفتوحة المصدر. (٢٨)

جدول (10) أنواع الملفات التي تدعمها البرامج المباشرة للتعرف الضوئي علي الحروف العربية

م	البرنامج	النصوص في شكل PDF image	النصوص في شكل صورةIMAGE	نوع الصورة Image format
1	OCR Online Using The Google Drive	×	√	JPEG, .PNG, .GIF
2	Newocr.Com	√	√	JPEG, JFIF, PNG, GIF, BMP, PBM, PGM, PPM, PCX
3	Finereaderonline.Com	√	√	JPEG, PNG, GIF, BMP, TIFF ,djvu
4	OCRconvert.Com	√	√	GIF, BMP, JPEG, PNG
5	I2ocr.Com	×	√	JPG - PNG - BMP - TIF - PBM - PGM – PPM
6	Totext.Net	√	√	JPG, JPEG, BMP, TIFF, GIF
7	Verypdf.Com	×	√	JPEG, PNG, GIF, BMP, TIFF.
8	Online OCR By A9t9.Com	√	√	(.png or .jpg
9	Abby finereader online	√	√	GIF, BMP, JPEG, PNG

يتضح من الجدول السابق أن كل برمجيات التعرف الضوئي علي الحروف العربية المباشرة online تدعم الملفات في صيغ صورimage ، في حين أن هناك عدد (6) برامج مباشرة تدعم الملفات في صيغ PDF image

٤/٣ أحجام الملفات التى تدعمها برامج التعرف الضوئى على الحروف العربية

يحدد كل برنامج من برامج التعرف الضوئي علي الحروف العربية الحد الأقصى لحجم الملف المطلوب تحويله إلي نص، والجدول التالي يوضح حجم الملفات المطلوبة لكل برنامج :

جدول (11)أحجام الملفات التي تدعمها برامج للتعرف الضوئي علي الحروف العربية التجارية

م	اسم البرنامج	الحد الأقصى للملفات المرفوعة (الحجم بالميجا)
١	Abby finereader14	غير محدد
٢	Readiris.16	غير محدد
٣	Sakhr ocr gold edition	غير محدد
4	Automatic page reader 3.01 form skhar	غير محدد
5	Leadtools high lever ocr toolkit	غير محدد
6	Verus professional multilingual ocr	غير محدد
7	I.R.I.S.OCR	غير محدد

نلاحظ أن البرامج التجارية لا تشترط حجما معينا للملف، وإن كان ذلك ينعكس علي وقت تحويل الصور إلي نصوص، فكلما كبر حجم الملف استغرق وقتا أطول في التحويل والقراءة والعكس، وينطبق ذلك أيضا علي البرامج مفتوحة المصدر كما يتضح من الجدول التالي:

جدول (12)أحجام الملفات التي تدعمها برامج للتعرف الضوئي علي الحروف العربية مفتوحة المصدر

م	اسم البرنامج	الحد الأقصى للملفات المرفوعة (الحجم بالميجا)
١	Tesseract ocr	غير محدد
٢	Arabic ocr image segmentation	غير محدد
٣	**GOCR**	غير محدد
٤	**Cuneiform**	غير محدد

يتضح من الجدول السابق أن البرمجيات مفتوحة المصدر لا تحدد حد أقصى للملف المراد تحويله إلي نص وأما برمجيات التعرف الضوئي علي الحروف العربية علي الإنترنت فتحدد الحد الأقصى للملف المراد تحويله إلي نصوص كما يتضح من الجدول التالي:

جدول (13) أحجام الملفات التي تدعمها البرامج للتعرف الضوئي علي الحروف العربية المباشرة

م	البرنامج	الحد الأقصى لحجم للملفات المرفوعة (الحجم بالميجا)
1	Newocr.Com	غير محدد
2	I2ocr.Com	غير محدد
3	Finereaderonline.Com	١٠٠
4	Verypdf.Com	10
5	OCRconvert.Com	5
6	Totext.Net	5

م	البرنامج	الحد الأقصى لحجم للملفات المرفوعة (الحجم بالميجا)
7	Online OCR By A9t9.Com	٥
8	OCR Online Using The Google Drive	2
9	Abby fine reader online	2

نلاحظ من الجدول السابق أن موقعين (Newocr.com وI2ocr.com) لا يحددان حجما محددا للملف المرفوع المراد تحويله إلي نصوص، أما موقع (finereadeonline.com) فيحدد الحد الأقصى للملف المرفوع المراد تحويله إلي نصوص ألا يتجاوز حجمه (١٠٠ ميجا)، أما مواقع (verypdf.com) فيحدد الحد الأقصى لحجم الملف المرفوع ألا يتجاوز حجمه (١٠ميجا)، أما المواقع الثلاثة (ocrconvert.com وtotext.net و online ocr bya9t9.com) يشترطوا ألا يتعدى حجم الملف المراد تحويله (٥ ميجا)، أما موقع ocr online google فيحدد الحد الأقصى لحجم الملف المراد تحويله إلي نصوص بألا يتعدى حجمه (٢ميجا)، ونلاحظ أن معظم البرامج المباشرة للتعرف الضوئي علي الحروف العربية تحدد أحجام الملفات المراد تحويلها، حتى يمكن السيطرة علي كمية البيانات والملفات المرفوعة إلي البرنامج حيث أنه متاح مباشر لكل المستفيدين، وبالتالي الضغط عليه سيكون كبيراً يؤثر ذلك علي أداء البرنامج.

٥/٣ أنواع مخرجات ملفات برامج التعرف الضوئى على الحروف العربية

تخزن برامج التعرف الضوئي علي الحروف العربية مخرجات النصوص التي يتم تحويلها من الصور إلي عدة تنسيقات، حتى يمكن حفظها ونسخها وتعديلها حسب حاجة المستفيد، وتختلف تنسيقات ملفات التخزين من برنامج إلي آخر، ويمكن حصر هذه التنسيقات في الجدول التالي:

جدول (14) أنواع ملفات تخزين لبرامج التعرف الضوئي علي الحروف العربية

م	نوع ملف التخزين	عدد البرامج التي تدعمه	النسبة
١	Text	١١	%٥٥
٢	Doc/ docx	٩	%٤٥
٣	PDF Searchable	٨	%٤٠
٤	Html	٦	%٣٠

يتضح من الجدول السابق أن نسبة (٥٥%) من برامج التعرف الضوئي علي الحروف العربية تخزن مخرجات التعرف الضوئي علي الحروف في شكل ملفات text، وأن نسبة (٤٥%) من البرامج تخزن مخرجات عملية التعرف الضوئي علي الحروف في شكل ملفات doc/docx، وأن نسبة (٤٠%) من برامج التعرف الضوئي علي الحروف العربية تحفظ مخرجات ocr في شكل ملفات pdf searchable وأن نسبة (٣٠%) من البرامج تخزن مخرجاتها في شكل html

جدول (15) أنواع ملفات تخزين لبرامج التعرف الضوئي علي الحروف العربية التجارية

م	اسم البرنامج	نوع الملف File format			
		PDF searchable	Doc/docx	Text	Html
١	Abby finereader14	√	√	×	×
٢	Readiris.16	√	√	√	√

م	اسم البرنامج	نوع الملف File format			
		PDF searchable	Doc/docx	Text	Html
٣	Sakhr ocr gold edition	√	√	×	√
٤	Automatic page reader 3.01 form skhar	√	√	×	×
٥	Leadtools high lever ocr toolkit	√	√	√	√
٦	Verus professional multilingual ocr	√	×	√	×
٧	I.R.I.S.OCR	√	√	√	×
	النسبة المئوية	%١٠٠	%٨٥,٧	%٥٧,١	%٤٢,٩

يتضح من الجدول السابق أنه يمكن تخزين مخرجات ملفات برمجيات التعرف الضوئي عل الحروف العربية التجارية تكون بصيغة PDF Searchable قابل للبحث، وهذه الصيغة هي الصيغة المعيارية للحفظ الرقمي للملفات والوثائق حيث تحافظ علي الشكل الأصل للوثيقة ولا يمكن التعديل فيها. ونسبة ٨٥,٧% من البرامج تخزن مخرجات التعرف الضوئي بصيغة doc/docx بحيث يمكن تعديلها وتحريرها بسهولة ويسر، ونجد أن نسبة ٥٧,١% من البرامج تخزن الملفات بصيغة text نص، أما أنواع ملفات التخزين للبرامج مفتوحة المصدر فيمكن حصرها في الجدول التالي:

جدول (16) أنواع ملفات تخزين لبرامج التعرف الضوئي علي الحروف العربية مفتوحة المصدر

م	اسم البرنامج	نوع الملف File format			
		PDF searchable	Doc/docx	Text	Html
١	Tesseract ocr	×	√	√	×
٢	Arabic ocr image segmentation	×	√	×	×
٣	GOCR	√	×	√	√
٤	Cuneiform	×	×	√	√
	النسبة المئوية	%٢٥	%٥٠	%٧٥	%٥٠

يتضح من الجدول السابق أن أكثر صيغ حفظ ملفات التعرف الضوئي علي الحروف العربية للبرمجيات مفتوحة المصدر هي صيغة نص text ثم يليه صيغ doc/docx وصيغ html أما صيغة pdf searchable لا يقوم به غير برنامج Arabic ocr image segmentation هذا بالنسبة للبرمجيات مفتوحة المصدر أما البرمجيات المباشرة علي الانترنت فيتضح من الجدول التالي:

جدول (19) أنواع ملفات تخزين لبرامج التعرف الضوئي علي الحروف العربية المباشرة

م	البرنامج	نوع الملف File format			
		PDF searchable	Doc	Text	Html
1	OCR Online Using The Google Drive	×	√	×	×
2	Newocr.Com	√	√	√	×
3	Finereaderonline.Com	√	√	√	√
4	OCRconvert.Com	×	×	√	×
5	I2ocr.Com	√	√	√	√

م	البرنامج	نوع الملف File format			
		PDF searchable	Doc	Text	Html
6	Totext.Net	×	×	√	×
7	Verypdf.Com	×	×	√	×
8	Online OCR By A9t9.Com	√	×	√	×
	النسبة المئوية	%٥٠	%٥٠	%٨٧,٥	%٢٥

يتضح من الجدول السابق أن نسبة ٨٧,٥% من برمجيات التعرف الضوئي علي الحروف العربية المباشرة علي الانترنت تتيح تخزين مخرجات الملفات بصيغة نص TEXT واقل صيغة هي HTML .

٦/٣ معدل دقة برامج التعرف الضوئي على الحروف العربية

يعد معدل الدقة من أساسيات تقييم برامج التعرف الضوئي علي الحروف سواء الأجنبية أو العربية، وهي الوظيفة الأساسية لمحرك بحث البرنامج، ويتوقف اختيار برنامج التعرف الضوئي علي الحروف علي مدى دقته في التعرف علي الحروف والكلمات، ويوضح الجدول التالي معدلات دقة برامج التعرف الضوئي علي الحروف العربية محل الدراسة :

جدول رقم (20) يوضح معدل الدقة لبرامج التعرف الضوئي علي الحروف العربية

م	معدل دقة التعرف علي الحروف العربية	عدد البرامج	النسبة %
١	معدل دقة ١٠٠%	١	%٥
٢	معدل دقة من ٩٠- ٩٩%	٣	%١٥
٣	معدل دقة ٨٠- ٨٩%	١	%٥
٥	معدل دقة أقل من ٨٠%	١٥	%٧٥

يتضح من الجدول السابق أن برنامجا واحدا وصل معدل دقته في التعرف علي الحروف العربية إلي (١٠٠%) أما البرامج التي معدل دقتها من ٩٠-٩٩% بلغ عددها (٣) برامج وبلغ عدد البرامج التي معدل دقتها من ٨٠-٨٩% برنامجاً واحداً أما البرامج التي معدل دقتها أقل من ٨٠% قد بلغ عددها (١٥) برامجاً.

وبنظرة عامة علي معدل دقة برامج التعرف الضوئي علي الحروف العربية نلاحظ أن برنامجا واحدا وصل معدل دقته ١٠٠% مما يدل علي قلة البرامج وندرتها التي تتعرف علي الحروف العربية بدقة عالية ويرجع ذلك إلي قلة الأبحاث والتمويل علي دعم مشاريع التعرف الضوئي علي الحروف العربية بالإضافة إلي طبيعة اللغة العربية وحروفها حيث تتميز بالصعوبة وكثرة التشابهات بين الحروف

يمكن حساب معدل دقة برامج التعرف الضوئي علي الحروف من خلال عدد الكلمات الصحيحة التي تعرف عليها البرنامج والجدول التالي يوضح مدى دقة برامج التعرف الضوئي علي الحروف العربية حيث تم مسح وثيقة بالماسح الضوئي بالاسكانر بمعدل وضوح ٣٠٠ نقطة للحصول علي نتائج جيدة وبلغ عدد كلمات الوثيقة ١٩٥ كلمة وتم حساب نسب الدقة بعدد الكلمات الصحيحة التي تعرف عليها البرنامج مقابل عدد الكلمات الخاطئة

جدول رقم (21) يوضح نسب دقة التعرف علي الحروفِ العربية

م	اسم البرنامج	عدد الكلمات	دقة التعرف علي الحروف			
			عدد الكلمات الصحيحة	النسبة	عدد الكلمات الخاطئة	النسبة
١	OCR Online Using The Google Drive	195	١٩٥	%١٠٠	٠	%٠
٢	Online OCR By A9t9.Com	195	١٩٢	%٩٨,٥	٣	%١,٥
٣	Sakhr ocr gold edition	١٩٥	١٨٨	%٩٦,٤	٧	%٣,٦
٤	Abby finereader14	١٩٥	١٨٧	%٩٥,٩	٨	%٤,١
٥	Finereaderonline.Com	195	١٧٣	88.7%	٢٢	%١١,٣
٦	Verypdf.Com	195	١٤٧	%٧٥,٤	٤٨	%٢٤,٦
٧	Readiris.16	١٩٥	١٢٧	%٦٥,١	٦٨	%٣٤,٩
٨	Newocr.Com	195	١٠٨	55.4%	٨٧	%٤٤,٦
٩	Totext.Net	195	٩٣	47.7%	١٠٢	%٥٢,٣
١٠	OCRconvert.Com	195	٨٦	%44.1	١٠٩	%٥٥,٩
١١	Tesseract ocr	١٩٥	٧٠	٣٥,٩	١٢٥	٦٤,١
١٢	Arabic ocr image segmentation	١٩٥	٦٤	٣٢,٨	١٣١	٦٧,٢
١٣	GOCR	١٩٥	٥٥	٢٨,٢	١٤٠	٧١,٨
١٤	Cuneiform	١٩٥	٤٠	٢٠,٥	١٥٥	٧٩,٥
١٥	I2ocr.Com	195	٠	%٠	١٩٥	%١٠٠

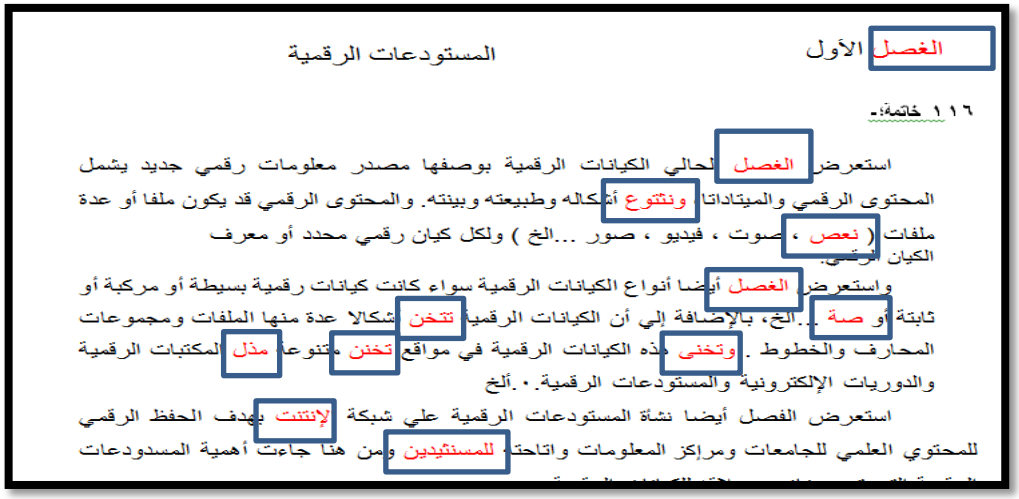
يتضح مما سبق أن موقع google drive ocr تعرف ضوئيا علي الكلمات والحروف العربية للوثيقة بنسبة ١٠٠%، ولم تنتج أخطاء في عملية التعرف الضوئي للنص العربي، ولكنه يشترط أن يكون الملف المراد التعرف عليه في صيغة صورة image حيث يتعرف علي صورة واحدة في كل عملية ocr ولا يقبل ملف مكون من عدة صور، أو ملف بصيغة pdf image، وهذا الموقع تابع لشركة جوجل العالمية وهي شركة متخصصة في تكنولوجيا المعلومات ورقمنة مصادر المعلومات والمسئولة عن مشروعgoogle books

صورة بعد التعرف الضوئي	صورة قبل التعرف الضوئي

شكل رقم(1) يوضح التعرف الضوئي علي مستند بواسطة Google drive ocr

أما البرامج التي تراوحت نسبة دقتها ما بين ٨٠%-٨٩% وتمثلت أخطاء التعرف علي الحروف العربية فيما يلي:

١. لم تستطع تلك البرامج من التفرقة بين حرف التاء(ت) وحرف النون (ن) وحرف الثاء (ث) حيث تعرفت عليهم علي أنهم حرف واحد.
٢. أخطاء التعرف علي حرف (ص) حيث تم التعرف عليه بأنه حرف (س)
٣. أخطاء التعرف علي حرف (ف) حيث تم التعرف عليه بأنه حرف (غ)
٤. أخطاء التعرف علي حرف (ط) حيث لا تفرق بينه وبين حرف (ظ)
٥. أخطاء التعرف علي حرف (ت) حيث لا تفرق بينه بين حرف (ث)
٦. أخطاء التعرف علي حرف (ع) حيث لا تفرق بينه وبين حرف (غ)



شكل رقم (2) نموذج لأخطاء التعرف الضوئي علي الحروف العربية ببرنامج abby finereader

٧/٣ معدل دقة برمجيات التعرف الضوئى على الأرقام العربية :

يعد التعرف علي الأرقام العربية داخل النصوص من الأمور الهامة حيث لا تخلو النصوص العربية من الأرقام، وإن كان التعرف الضوئي علي الأرقام العربية أسهل بكثير من التعرف علي الحروف والكلمات والجدول التالي يوضح معدل دقة التعرف الضوئي علي الأرقام العربية .

جدول (22) معدل دقة التعرف الضوئي علي الأرقام العربية

م	البند	عدد البرامج	النسبة %
١	التعرف علي الأرقام العربية	٢٠	100

يتضح من الجدول السابق أن كل برمجيات التعرف الضوئي علي الحروف العربية تمكنت من التعرف الضوئي علي الأرقام العربية بنسبة (١٠٠%) ومعدل دقة (١٠٠%) ويعود ذلك إلي أن الأرقام العربية لها شكل ثابت ولا تكتب متصلة وبالتالي يسهل للبرمجيات التعرف عليها بشكل دقيق.

٨/٣ الالتزام بشكل النص الأصلي

لعل ما يميز أي برنامج من برامج التعرف الضوئي علي الحروف عن الآخر، هو المحافظة علي الشكل الأصلي للنص، أي استنساخ نفس التكوين الهيكلي للصفحات من أعمدة، وجداول، ورسوم جرافيك وغيرها، بالإضافة إلي نفس الخصائص النصية من نوع الحرف وحجمه وشكله وجسم النص، والقدرة علي التعرف الشكلي والبنائي والهيكلي للصفحات التي تجري لها عمليات التعرف الضوئي، من حيث التعرف علي الجداول، بحيث يكون قادراً علي اكتشاف الجداول في النص، ومن ثم أعادة صياغتها في برامج معالجة النصوص، والجدول التالي يوضح مدى التزام برمجيات التعرف الضوئي علي الحروف العربية بشكل النص الأصلي.

جدول رقم (23) مدى التزام برمجيات التعرف الضوئي علي الحروف العربية بشكل النص الاصلي

م	بنود الإلتزام بشكل النص الأصلي	عدد البرامج	النسبة %
١	التعرف علي الفقرات	٥	%٢٥
٢	التعرف علي التشكيل	١	٥
٣	التعرف علي الجداول	٠	0
٤	التعرف علي الصور داخل النص	٠	0
٥	التعرف علي النصوص في أعمدة	٠	0

يتضح من الجدول السابق أن برمجيات التعرف الضوئي علي الحروف العربية عند تحويل صور النصوص العربية إلي نصوص، لم تلتزم هذه البرمجيات بشكل النص الأصلي المحول حيث نجد أن عدد (٤) برامج فقط بنسبة (٢٥%) من إجمالي برمجيات التعرف الضوئي علي الحروف العربية هي التي تعرفت وحافظت علي فقرات النصوص كما وردت بالشكل الأصلي.

وتبين أن برنامجا واحدا هو الذي تمكن من التعرف علي التشكيل، ولكن ليس بنسبة دقة عالية، إلا انه تعرف علي بعض علامات التشكيل، ويرجع ذلك لعدم اهتمام مطوري نظم التعرف الضوئي علي الحروف العربية بتطوير تقنيات التعرف علي التشكيل؛ لصعوبة التعرف عليها من ناحية وكثرتها من ناحية أخرى بالإضافة إلي ندرة الوثائق العربية المكتوبة بالتشكيل .

ونلاحظ أن كل برامج التعرف الضوئي علي الحروف العربية لم تتعرف علي الجداول أو الصور داخل النصوص، ولم تعرف علي النصوص المكتوبة في شكل أعمدة، حيث لم تستطع تلك البرامج من قراءة كل عمود علي حدة بل تعرفت علي النصوص والسطور كاملة مما غير المعنى

جدول رقم (24) يوضح مدى الالتزام بشكل النص الأصلي للنظم التجارية

م	اسم البرنامج	التعرف علي الفقرات	التعرف علي الجداول	التعرف علي الصور	التعرف علي علامات التشكيل	التعرف علي الأرقام العربية
١	Abby finereader14	√	×	×	×	√
٢	Readiris.16	√	×	×	×	√
٣	Sakhr ocr gold edition	√	×	×	√	√
٤	Automatic page reader 3.01 form skhar	×	×	×	×	√
٥	Leadtools high lever ocr toolkit	×	×	×	×	√
٦	Verus professional multilingual ocr	×	×	×	×	√
٧	I.R.I.S.OCR	×	×	×	×	√

جدول رقم (25) يوضح مدى الالتزام بشكل النص الأصلي للنظم مفتوحة المصدر

م	اسم البرنامج	التعرف علي الفقرات	التعرف علي الجداول	التعرف علي الصور	التعرف علي علامات التشكيل	التعرف علي الأرقام العربية
١	Tesseract ocr	×	×	×	×	√
٢	Arabic ocr image segmentation	√	×	×	×	√
٣	GOCR	×	×	×	×	√
٤	Cuneiform	×	×	×	×	√

جدول رقم (26) يوضح مدى الالتزام بشكل النص الأصلي لبرامج البث المباشر online

م	البرنامج	التعرف علي الفقرات	التعرف علي الجداول	التعرف علي الصور	التعرف علي علامات التشكيل	التعرف علي الأرقام العربية
1	**OCR Online Using The Google Drive**	√	×	×	×	√
2	**Newocr.Com**	×	×	×	×	√
3	**Finereaderonline.Com**	×	×	×	×	√
4	**OCRconvert.Com**	×	×	×	×	√
5	**I2ocr.Com**	×	×	×	×	√

م	البرنامج	التعرف علي الفقرات	التعرف علي الجداول	التعرف علي الصور	التعرف علي علامات التشكيل	التعرف علي الأرقام العربية
6	Totext.Net	×	×	×	×	√
7	Verypdf.Com	×	×	×	×	√
8	Online OCR By A9t9.Com	×	×	×	×	√

القسم الرابع : النتائج والتوصيات

١/٤ النتائج:

خرجت الدراسة الوصفية لبرامج التعرف الضوئي علي الحروف العربية والمقارنة بينها بمجموعة من النتائج التي انقسمت إلي :

أ. برمجيات التعرف الضوئى على الحروف العربية :

١. قلة برمجيات التعرف الضوئي علي الحروف العربية وانقسمت ما بين برمجيات تجارية ومفتوحة المصدر ومتاحة علي الإنترنت .

٢. أكثر برمجيات التعرف الضوئي علي الحروف العربية متاحة علي الإنترنت حيث بلغت نسبتها (٤٥%) ثم تلتها البرمجيات التجارية بنسبة (٣٥%) ثم برمجيات مفتوحة المصدر بنسبة (٢٠%)

٣. يتوافق نسبة (٥٥%) من برمجيات التعرف الضوئي علي الحروف العربية مع نظام تشغيل ويندوز windows ثم يليه نظام تشغيل الويب web host حيث يتوافق مع نسبة (٤٥%) من البرمجيات ثم يليه نظام تشغيل لينكس LUNIX حيث يتوافق مع نسبة ٢٥% من تلك البرمجيات

٤. تدعم نسبة (٧٥%) من برمجيات التعرف الضوئي علي الحروف العربية ملفات النصوص علي شكل صورة image وتدعم نسبة (٤٥%) من تلك البرمجيات صيغ ملفات النصوص في شكل PDF

٥. لم تحدد نسبة (٦٠%) من برمجيات التعرف الضوئي علي الحروف العربية حجم الحد الأقصى للملفات التي يقبلها البرنامج للتعرف علي النصوص العربية.

٦. تعد الملفات في صيغ وشكل TEXT أكثر مخرجات برمجيات التعرف الضوئي علي الحروف العربية بنسبة (٥٥%) ثم تلتها الملفات في صيغة DOC/DOX بنسبة (٤٥%) ثم تلتها ملفات في صيغة PDF Searchable بنسبة (٤٠%)

ب. معدل دقة برمجيات التعرف على الحروف العربية :

١- بلغت نسبة دقة برنامج Google drive doc (١٠٠%) في التعرف الضوئي علي الحروف العربية.

٢- بلغ عدد البرامج التي وصل معدل دقتها ٩٠-٩٩% في التعرف الضوئي علي الحروف العربية (٣) برامج بنسبة (٢٠%)

٣- تمركزت أخطاء برمجيات التعرف علي الحروف العربية في الحروف العربية المنقوطة والحروف المتشابهة .
٤- تعرفت كل برمجيات التعرف الضوئي علي الحروف العربية علي الأرقام العربية .
٥- حافظ عدد (٤) برامج علي شكل فقرات النصوص التي تم التعرف الضوئي عليها.
٦- لم تتمكن برمجيات التعرف الضوئي علي الحروف العربية التعرف علي الجداول داخل النصوص.

٢/٤ : التوصيات:

توصى الدراسة أن تقوم مؤسسات المعلومات العربية بتطوير برمجيات للتعرف علي الحروف العربية بدقة عالية، وأن توفر هيئات البحث العلمي ومراكز البحوث ميزانيات لتطوير تقنيات التعرف الضوئي علي الحروف العربية .

مصادر الدراسة:

١- أحمد، أحمد فرج. "تقنيات التعرف الضوئي للحروف: معايير الاختيار ، طريقة العمل ، الاشكاليات ، والافاق المستقبلية ".*مجلة المعلوماتية: وزارة التربية والتعليم - وكالة التطوير والتخطيط* ع ٢١ (٢٠٠٨).-ص ص ٢٦-٣١ (ص ٢٦.) مسترجع من http://search.mandumah.com/Record/30459

٢- يس، نجلاء أحمد. "متطلبات التحول الرقمي لمؤسسات المعلومات العربية ".*مجلة المكتبات والمعلومات: دار النخلة للنشر* ع١٣ (٢٠١٥).- ص ص٢٧ – ٩٠ .(ص٢٧) مسترجع من http://search.mandumah.com/Record/781047

3- Qartameez, Mohammad mahmmoud ali. "Arabic printed text analysis and character recognition". Thesis (M.S) - yurmouk university. Faculty of science, 2010.

٤- عبد الهادي، محمد فتحي . البحث ومناهجه في علم المكتبات والمعلومات .القاهرة : الدار المصرية اللبنانية، ٢٠٠٣. ص ١٠٣.

5- Hosch, willaim L. “OCR technology:. *Encyclopedia Britannica*. Chicago: Encyclopaedia Britannica. available at <]http://08107ypq6.1104.y.https.academic.eb.com.mplbci.ekb.eg/levels/collegiate/article/OCR/472978 > (1/12/2017)

6- Reitz, Joan M., odlis : Online Dictionary for Library and Information Science. available at . https://www.abc-clio.com/ODLIS/odlis_A.aspx (1/9/2017)

7- Optical scanner technology . *Encyclopædia Britannica*.2019 Available at http://08107ypii.1104.y.https.academic.eb.com.mplbci.ekb.eg/levels/collegiate/article/optical-scanner/473252 (15/11/2017)

٨- أرمز، وليام. المكتبات الرقمية . ترجمة جبريل بن حسن العريشي، هاشم فرحات سيد. الرياض : مكتبة الملك فهد الوطنية ، ٢٠٠٦. ص ٣٣٣.

9- Radwan,Mohamed Atia Mohamed, deep learning approaches in Arabic ocr. Supervised By Hazem M. Abbas,Mahmoud I. Khalil. thesis(M.S.) - ain shams

university.faculty of engineering.Computer Engineering and Systems Dep, 2018

10- Ayyash, muna abd al Fattah.Arabic optical character segmentation stage. Advisor mohammad khader.these (ms) Birziet university, faculty of engineering and technology.2016] <http://search.mandumah.com/record/850876>

١١-يس، نجلاء أحمد . مرجع سابق. ص. ٤٨.

12- Zaree, marwa Rashad.Arabic optical character recognition. Supervised by mohiy Mohamed hadhoud, weil shawky elkilani, khaled Mohamed amin. Theses (m.s) menofia university. Faculty of computers and information, department of information technology,2012

١٣-أحمد، أحمد فرج . مرجع سابق . ص ١٠

14- M. McClean, Clare. Digitization of Full-Text Documents Before Publishing on the Internet: A Case Study Reviewing the Latest Optical Character Recognition Technologies. **Library Software Review**. 17. September 1998

15- Hamouda ،Eslam Fouad Mohamed. **Optical character recognition algorithms using evolutionary computations** . Supervised by Taher Tawfik Hamza ،El-Sayed Fouad Radwan, 2011

16- Zaree, Marwa Rashad. **Arabic Optical character Recognition** .supervisors Mohiy Mohamed Hadhoud, Wail Shawky El-kilani . Thesis (M.S) - Menofia university. Faculty of computers and information.Department of Information systems, 2012.

17- Alkholy, Mohamed Dahi Abdel-Zaher. **Arabic Optical character Recognition Using Local Invariant Features** .supervisor Mohiy M.Hadhoud, Noura Semary . thesis(m.s) -Minoufiya UniversitY.Faculty of Computers and Information.Department of Information Technology,2016.

18- Joan M. Reitz , **ODLIS — Online Dictionary For Library and Information Science**.[2018] < https://www.abc-clio.com/ODLIS/odlis_A.aspx>

١٩-يس، نجلاء أحمد . المصدر السابق . ص ٦٢

20- Bowker,lynne. **Computer-aided translation technology: a practical introduction**. Canada: university of Ottawa press, 2002. P28

21- Eikvil, line, **OCR-Optical character recognition** , 1993.available at <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.25.3684> . [2/4/2018]

٢٢-- يس، نجلاء أحمد ، مرجع سابق ، ص ٦٧

٢٣- أبو زريده، مصطفى علي ، أكرم محمد زكي ، أحمد محمد زكي . "المراحل التمهيدية وأهميتها في أنظمة التعرف علي الكتابة العربية ". **المجلة الدولية للتطبيقات الإسلامية في علم الحاسب والتقنية** .. مج١، ع٢ (سبتمبر ٢٠١٣) . ص ص ١٤-٢٣

24- Puckett, Jason. **Open source software and library values. Library and information science commons**. 2018. P.159, [8/6/2018] <https://scholarworks.gsu.edu/univ_lib_facpub>

25- Patel, chirag. Atul patel, Dharmendra Patel .”optical character recognition by open source ocr tool tesseract : a case study” . **international journal of computer applications** . v55, n10 ,(october 2012) . P.50

26- Holley, Rose . “How Good Can It Get? : Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs” . D-Lib Magazine (Mrch/April 2009).V15. N3/4

٢٧- أحمد، أحمد فرج . مرجع سابق. ص٢٧.

28- Sourcforge: the complete open-source and business software platform. available at < https://sourceforge.net/> [7/8/2018]