التعرف على المحارف العربية الياً ضمن نصوص صور الوثائق متعددة اللغات

أ.د. عبد المنعم صالح رحمة*
جامعة التكنولوجية
بغداد – العراق

أ.د. هلال محمد يوسف
جامعة العلوم التطبيقية
المنامة - البحرين

د.فوزي عبد الواحد العضاض
الجامعة المستنصرية
بغداد – العراق

ملخص

المعالجة الآلية للوثائق المتعددة اللغات تحتوي على عدد من المشاكل الصعبة. لأن تقنيات التعرف على الأحرف ليست متوفرة لكلّ أصناف اللغات. الوثيقة المصوره يجب أن تصنّف طبقا لنوع محار ف اللغة أولا، في هذا البحث تم التعرف على المحارف الرومانية (الانكليزيه) أو العربية. تم احتساب ستة ميزات لتمثيل محارف الوثيقة ولكل ميزه تم احتساب ثلاثة معايير احصائيه هي المتوسط , الانحراف المعياري والانحراف. استخدم البحث مصنف الذي يعتمد على مسافة "موهالمبس(Mahalanobis) " لتصنيف الوثيقة إلى الرومانيه(الانكليزيه) أو العربية .تم الحصول على 100 % دقة تصنيف لوثائق الاختبار..

1. مقدمة

تصنيف المحارف الى اللغة التي تنتمي اليها اصبحت عملية اساسية في المعالجة الالية لصور الوثائق متعددة اللغات.هذه العملية تساعد على الاختيار الامثل لنظام الألي للتعرف على الاحرف.

تعتبر هذه العملية مرحلة اساسية قبل عملية الترجمة خصوصا للأشخاص غير ملمين باللغة العربية، حيث قد تختلط لديهم شكل الاحرف العربية مع اشكال الاحرف في لغات أخرى.

تصنيف محارف اللغة في اللغات التي تكون فيها الحروف مستخدمة فقط من لغة واحدة تعتبر عملية ضمن عملية التعرف الالي على الاحرف. لكن بالنسبة للاحرف العربية والرومانية (الانكليزية) حيث تستخدم هذه الحروف في اكثر من لغة تعتبر عملية تصنيف الاحرف الى اللغة التي تنتمي اليها عملية اساسية اولية قبل عملية التعرف الالي على الاحرف.

في مجال تصنيف المحارف العربية هنالك عدة معوقات اساسية.منها تعقيد النص العربي بسبب الانحناءات المتكررة للاحرف العربية، الاشكال المختلفة للاحرف العربية بحسب مواقعها داخل الكلمة، هذا بالاضافة الى تراكب وترابط الاحرف العربية في الكلمات بحسبب موقعها على خط الكتابة(1).

2. الاعمال البحثية السابقة

ان اهمية هذا الموضوع قد استوجبت الكثير من البحث في هذا المجال. مرشح كابور قد اثبت فعاليته في تصنيف الحرف (2)(3)(4). وقد اعتمدت في هذه الطريقة .حيث تعتمد هذه الطريقة على أستخدام العديد من الميزات للتمييز على أساس الإسقاط الأفقي وتوزيع ارتفاع المكونات المتصلة لتصنيف النصي العربي والروماني. لكن هذه الطريقة قد اثبتت كفاءتها على النصوص المطبوعة الياً ولم تقيس جيداً على النص المكتوب باليد.

علي سلامات (5) اقترح تصنيف الاحرف العربية بالاعتماد على تكرار الاحرف، حيث اعتمدت خوارزمية البحث على اسلوب الرنين المتكيف في المنطق المضبب، التي تنتمي الى اسلوب تعليم الانظمة بالاعتماد على الشبكات العصبية.

* monem.rahma@yahoo.com

بن سالم (6) عرض في بحثه طريقة لتصنيف الحروف العربية والرومانية في الوثائق متعددة اللغات. الطريقة تعتمد اولا على التحويل التشكيلي لصورة خط النص. ثم تحليل خصائص النص المصور باستخراج 12 صفة للنص.

3. تعريف ميزات النصوص والعمليات الحسابية المطبقة عليها

يتم تصنيف النصوص الى اللغة التي تنتمي اليها عن بطريقتين اما عن طريق الصفات العامة أوالخاصة. حيث تستند الميزات العامة على قوام النص ويعتمد على مصفوفات لتحليل المويجات المستخرجة من النص الكتابي وتنفذ ايضا تحليل على مستوى الكلمة بستخدام مرشحات خاصة كمرشح كابور.

اما المجموعة الثانية من الميزات المحلية معضمها يعمل على سمات تستخلص على مستوى الكلمة في النص والمكونات المتصلة. الاساليب هنا كثير وتقسم الى اربع فئات:سمات هيكلية أو هندسية ، الخصائص التشكيلية ، الميزات الإحصائية و انتشار الميزات المكانية (7).

على المستوى الميزات المحلية عند النظر الى صورة الوثائق تحول الصورة الى مجموعة من الكائنات المترابطة لكل كائن مجموعة من النقاط تكون هذه المجموعة من 8 نقاط مترابطة حيث تم تعريف مجموعة من الصفات المحلية التي استخدمت في هذا البحث.

- **متجه النقاط الوسطى (Centroid) :**متجه يحتوي على الاحداثي السيني و الصادي لمركز الكتلة في المنطقة.
- **المنطقة (Area):** العدد الفعلي للنقاط في المنطقة.
- **الصندوق المحيط (Bounding_box):** هو أصغر مستطيل يمكن أن يحتوي على الكائن. و يمثله متجه يحتوي على إحداثيات الزاوية العلوية اليسرى من الصندوق ، و العرض والارتفاع للصندوق.
- **مساحة الصندوق المحيط (Area of Bounding_box)** = العرض * الارتفاع من مربع الاحاطة.
- **المدى (Extent):** نسبة النقاط في مربع الاحاطة والصندوق.
- **المحور الأكبر(Major Axis length):** طول (ويقاس بعدد النقاط) المحور الاكبر للشكل البيضوي للكائن
- **المحور الاصغر (Minor axis length):** طول (ويقاس بعدد النقاط) المحور الاصغر للشكل البيضوي للكائن
- **نسبة الارتفاع الى العرض (Aspect ratio)**: هي نسبة ارتفاع Bounding-box الى عرضه.
- عدد اولير **(Euler_number)** والفجوات **(holes):** عدد اولير هو عدد الكائنات في المنطقة مطروح منه عدد الفجوات ,ففي منطقة box_Bounding التي تحتوي على كائن واحد من النقاط المترابطة الفجوات تكون مساوية ل holes = 1- Euler_number
- **الانحراف (Eccentricity):** انحراف الشكل البيضوي للكائن ، الانحراف هو نسبة المسافة بين بؤر الشكل البيضوي و طول المحور الرئيسي له . القيمة تكون بين 0 و 1 . (تكون القيمة 0 كلما كان الشكل البيضوي اقرب الي الشكل الدائري، في حين تكون القيمة 1 عندما يكون الشكل البيضوي قطعة مستقيمة) . يمكن حساب استدارة الكائن بالعلاقات الرياضية التالية (8): -

$$\text{eccentricity}(ecc) = 2 \times \frac{\sqrt{\left(\frac{major}{2}\right)^2 - \left(\frac{minor}{2}\right)^2}}{major} \quad \ldots\ldots\ldots\ldots\ldots(1)$$

حيث ان:

$$major = 2 \times \sqrt{2 \times (uxx + uyy + common)} \quad \ldots\ldots\ldots\ldots\ldots(2)$$
$$minor = 2 \times \sqrt{2 \times (uxx + uyy - common)} \quad \ldots\ldots\ldots\ldots\ldots(3)$$
$$common = \sqrt{(uxx - uyy)^2 + 4 \times uyy^2} \quad \ldots\ldots\ldots\ldots\ldots(4)$$
$$uxx = \frac{1}{N} \sum_{i=1}^{N} x_i^2 + \frac{1}{12} \quad \ldots\ldots\ldots\ldots\ldots(5)$$
$$uyy = \frac{1}{N} \sum_{i=1}^{N} y_i^2 + \frac{1}{12} \quad \ldots\ldots\ldots\ldots\ldots(6)$$

علما ان قيمة (N) تمثل عدد النقاط في الكائن , و(x) (y) هي القيم المحسوبة ومخزونة في متجه النقاط الوسطى.

لو افترضنا مثلا ان صورة ثنائية (binary image) التالية (اي ان قيم النقط في الصورة اما 0 أو 1).

$$\text{binary image} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

وكانت قيم:

(Bounding_box) = [0.5000 0.5000 4.0000 4.0000]

Major axis = 4.6188

Minor axis = 4.6188

فأن قيمة eccentricity ستكون 0.

4. عملية التصنيف (classification)و خطاء التصنيف (classification error)

ان خطاء التصنيف يعتمد على تداخل (overlap) بين الاصناف. ان الابتعاد بين مراكز الاصناف مع تصغير التداخل بين الاصناف ينتج عنه امكانية تصنيف بجودة عالية.

ان التصنيف الجيد يعتمد بالاساس على المتوسط الحسابي(mean) للصنف و كذلك على معامل اخر وهو معمال التشتت (variance) حيث انه كلما كان معامل التشتت قليل قلت نسبة الخطاء في التصيف.

وبالاعتماد على العلاقة السابقة فأن وجود معامل تشتت محدد فان ابعاد المتوسطات الحسابية للأصناف سوف يقلل من التداخل ويجعل التصنيف اكثر دقة. اما في الحالات التي يكون فيها معامل التشتت عالي وأن كان المتوسط للاصناف بعيد فأن التداخل موجود و يكون هنا التصنيف قليل الدقة.

وبذلك يكون التصنيف الدقيق محسوب بالعتماد على مقياس مركب يعتمد على معامل التشتت والمتوسط الحسابي في نفس الوقت.

ان المقياس الذي يحتوي على علاقة بين كل من (mean) و (variance) هو معامل (Mahalanobis distance classifier) ويحسب هذا المعامل بالاعتماد على العلاقة الرياضية التالية:

$$MD = transpose\ (x-u) * \sum^{-1} * (x-u) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

حيث ان...

$\sum$هو المصفوفة التغاير (covariance matrix) لفئتين تصنيف.
$\sum^{-1}$ هو معكوس المصفوفة السابقة.
u هو متجه المتوسط الحسابي $u = [u_1 \quad u_2 \quad \cdots \quad u_d]'$, حيث ان d هو بعد المصفوفة.

استنادا إلى المعادلات الواردة أعلاه يمكن الاستفادة منها من ناحية زيادة المسافة بين المصنفات، وبذلك يمكن القول ان مصنف الأمثل يمكن أن يبنى بأن يكون قادرا على التمييز بين فئتين من الكائنات. أن الأمثل هنا لا يعني أن العملية ستكون خالية من الأخطاء، إلا أن المصنف قادر على تقليل الاخطاء الى الحد الادنى (9).

5. تصنيف النصوص

الهدف الاساسي من هذه البحث هو تصنيف النصوص في صور الوثائق متعددة اللغات الى عربي ورومانى (انكليزي). هناك خاصيتين اساسية يمكن تقود عملية التحليل لأجراء التصنيف و من بعده اختيار الصفات لغرض التقسيم اولا ان النصوص العربية تكتب من اليمين الى اليسار لذلك اختيار الصفات الخاصة بالتصنيف يجب ان تعكس هذه الحقيقة. ثانيا الكتابة باللغة العربية هي بشكل منحنيات وهذا يعني (Bounding_box) للمكونات المترابطة سيكون اكبر افقيا مقارنتاً باللغة الانكليزية التي ممكن ان تكون الكتابة بها بشكل منحنيات او من الممكن ان لا تكون ايضا.
على سبيل المثال (centroid) للمكونات المترابطة في اللغة العربية يبدي ميل الى يمين مركز (Bounding_box) . و (aspect ratio) للغة العربية يتكون اصغر من اللغة الانكليزية.
هذا بالاضافة الى ان الكائنات في اللغة العربية ستميل الى الاستطالة اكثر من اللغة الانكليزية. أن قيمة (Eccentricity) تبين مقدار استدارة الكائنات. عموماً فأن اللغة الانكليزية تميل للاستدارة اكثر من اللغة العربية بسبب الطبيعة الانحنائية لأشكال الغة العربية. وهذا السبب ايضا يجعل حجم الكائن (Extent) في اللغة العربية اصغر من الكائنات في اللغة الانكليزية.

أن المقارانات والتحليلات السابقة قد اعتمدت في البحث حث تم تطبيق تحليلات احصائية على عينات عشوائية من الوثائق من كلا اللغتين , وكما مبين في الجدول (1).

جدول رقم (1) المتسوط الحسابي لبعض الميزات

	Distance	Extent	Holes	eccentricity	Aspect
Arabic Script	**.3**	**.6**	**.2**	**.9**	**1.2**
English Script	**.2**	**.4**	**.7**	**.7**	**1.5**

بالاعتماد على التنتائج السابقة ،فأن مجموعة من الصفات قد حددت لغرض تصنيف اللغة داخل الوثائق وهذه الصفات هي:

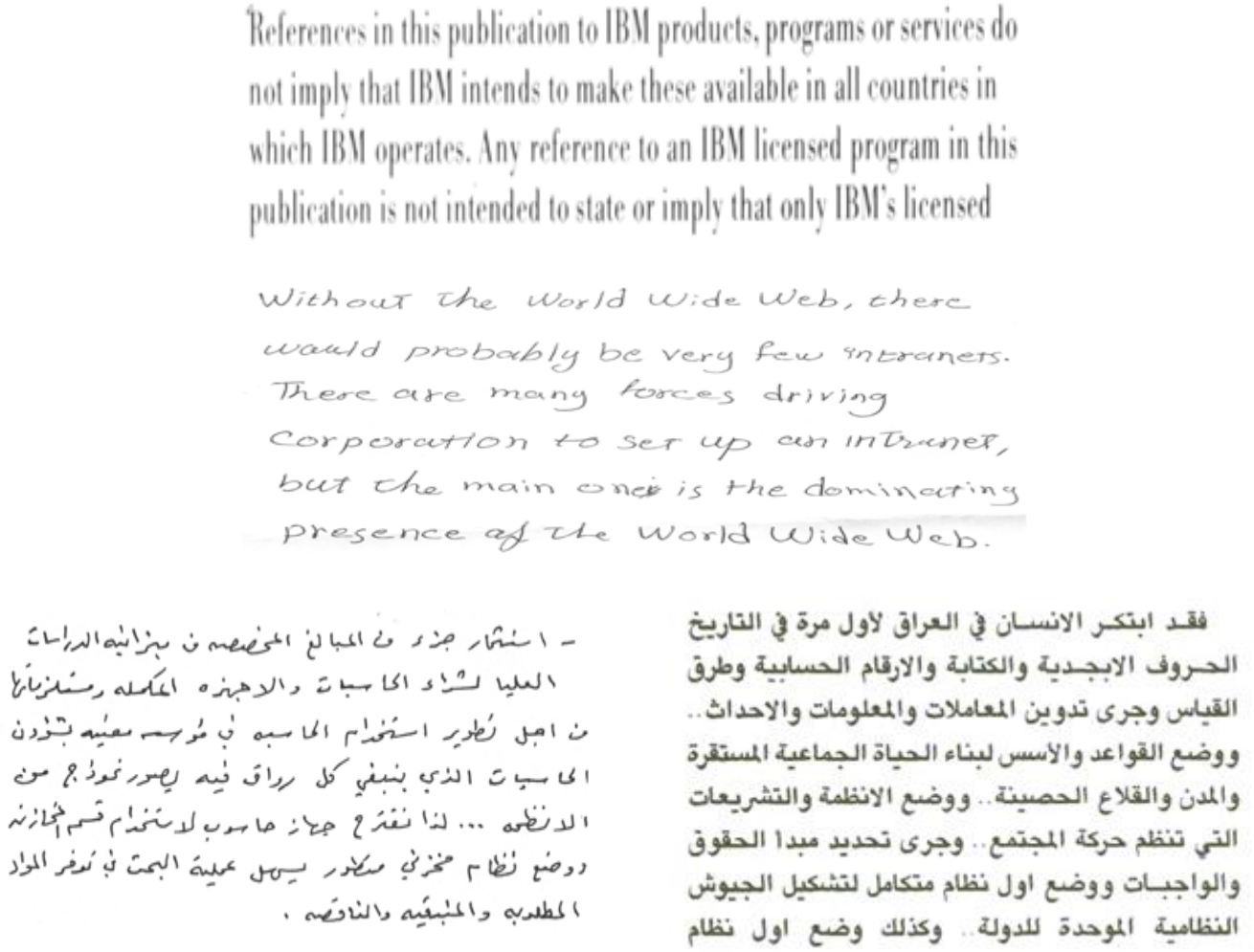
- centroid distance.
- angle of centroid distance vector.
- number of holes.
- eccentricity.
- aspect ration.
- extent.

كل الصفات السابقة هي نسب لذلك لا حاجة لتقييسهم بحدود قيم معينة ،فقط (centroid distance) يجب ان تقيس بقيم بين [0,1] .
لكل وثيقة يرغب في تصنيفها يجب توليد ملخص احصائي لكل صفة من الصفات السابقة. القيم الاحصائية هي المعدل, الانحراف المعياري و الالتواء. حيث يولد متجه من 18 عنصر لغرض تصنيف كل وثقة للاستخدام في عملية التصنيف. كما واستخدم (Mahalanobis distance classifier) كعلاقة بين (mean) و(variance).

6. تحليل النظام ونتائج اختبار نظام التصنيف

لاغراض الاختبار تم الاعتماد على من الوثائق التي تم اختيارها عشواءيا .فصلت هذه الوثائق الى قسمين القسم الاول لتدريب النضام والقسم الثاني لاغراض فحص النضام شكل (1) يبين مجموعة من الوثائق المستخدمة لتدريب النظام. الوثائق التي استخدمت للتدريب تتكون من:

- 16 وثيقة مطبوعة باللغة الانكليزية
- 16 وثيقة مكتوبة باليد انكليزية
- 16 وثيقة مطبوعة باللغة العربية
- 16 وثيقة مكتوبة باليد باللغة العربية


References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM licensed program in this publication is not intended to state or imply that only IBM's licensed

Without the World Wide Web, there would probably be very few intranets. There are many forces driving corporation to set up an intranet, but the main one is the dominating presence of the World Wide Web.

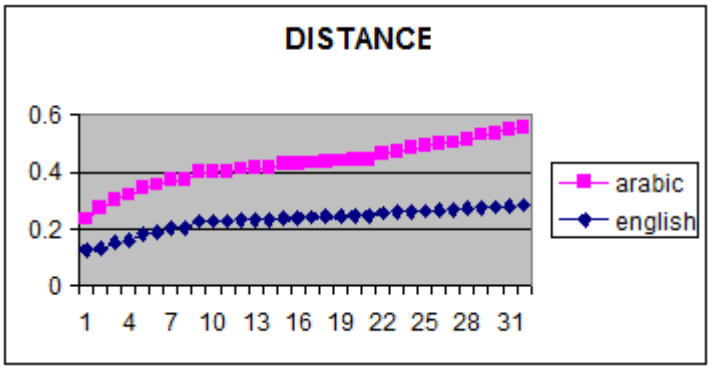
فقد ابتكر الانسان في العراق لأول مرة في التاريخ الحروف الابجدية والكتابة والارقام الحسابية وطرق القياس وجرى تدوين المعاملات والمعلومات والاحداث.. ووضع القواعد والأسس لبناء الحياة الجماعية المستقرة والمدن والقلاع الحصينة.. ووضع الانظمة والتشريعات التي تنظم حركة المجتمع.. وجرى تحديد مبدا الحقوق والواجبات ووضع اول نظام متكامل لتشكيل الجيوش النظامية الموحدة للدولة.. وكذلك وضع اول نظام

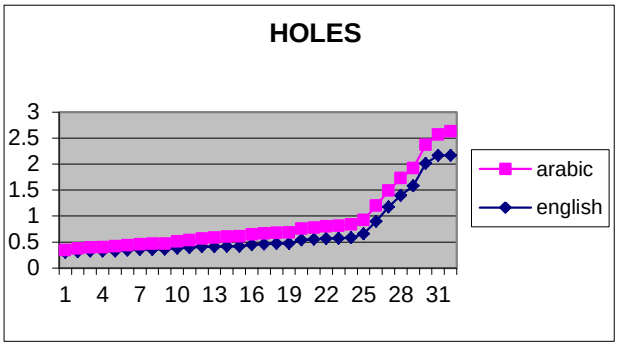
- استثمار جزء من المبالغ المخصصة من ميزانية الدراسات العليا لشراء الحاسبات والاجهزه المكمله وستلزمنا من اجل تطوير استخدام الحاسبه في موجهه بعضهم بتزودن الحاسبات الذي ينبغي كل رواق فيه يصور نموذج من الانظمه ... لذا نقترح جهاز حاسوب لاستخدام قسم لمجازنه وضع نظام مخزني مكثور يسهل عملية البحث في توفر المواد المطلوبه والمتبقيه والناقصه .

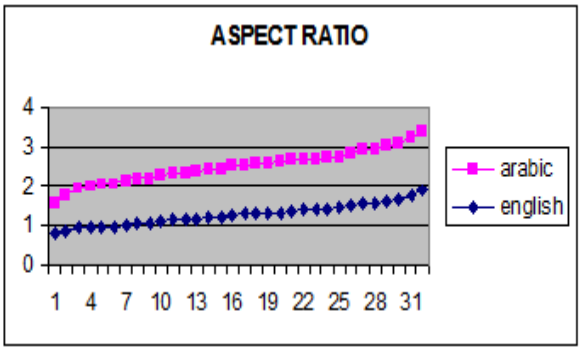
شكل (1) مجموعة نماذج من الوثائق المستخدمة لتدريب النظام

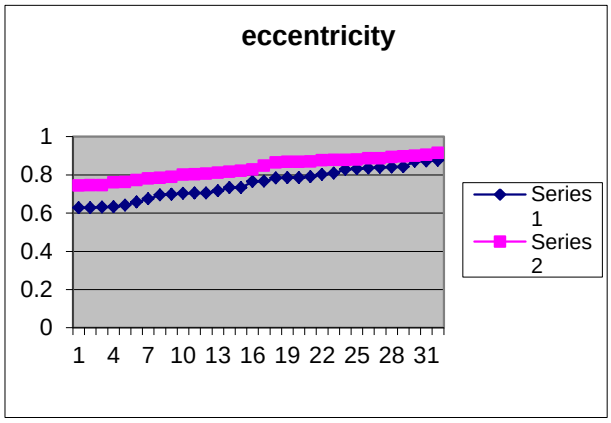
الوثائق المكتوبة باليد اخذت من اشخاص مختلفين بخلفيات مختلفة حتى لا يؤثر هذا على علمية التصنيف.متجه من 64 قيمة قد ولد لاغراض التصنيف وقد وجد ان هذا المتجه هو الحد الادنى المناسب طولاً للابتعاد عن نسبة الخطاء بالنظام.

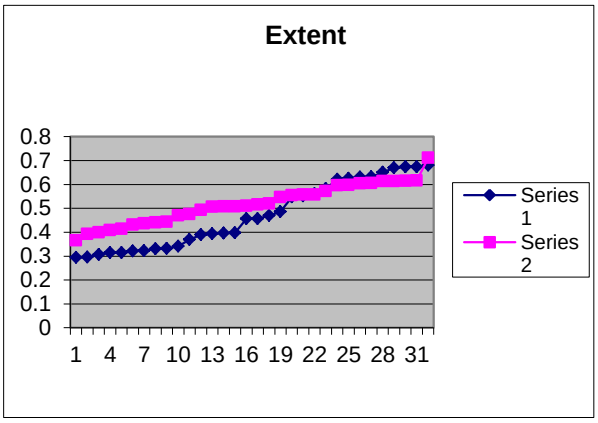
مخطط انحناء لصفات متجه التدريب مبينة بالشكل (2). الامر الاساسي الذي يجب ملاحظته بالمخطط هو ليس القيم او شكل منحنى المخطط وانما الانفصال بين المنحنيين لكل لغة .لكل صفة بالمنحنى نجد انفصالها عن الصفة المقابلة بالمنحنى للغة الثانية.











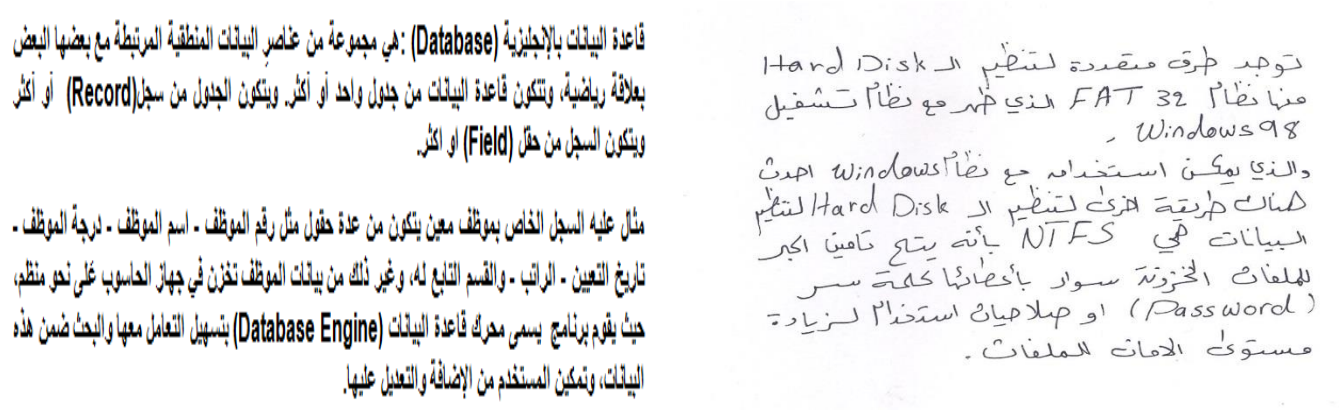
شكل (2) منحنيات خواص النصوص.

أن بعض الصفات تبين انفصال اكبر من الصفات الاخرى.الجدول (2) يبين نتائج فحص النظام.

جدول رقم (2) نتائج تصنيف الوثائق

نوع الوثيقة	عدد الوثائق	عدد الكتل (blocks)	النسبة المئوية (%) لمقدار صحة التصنيف
انكليزي– عربي طباعة	40	6210	100
انكليزي –عربي خط يد	40	7349	100
انكليزي عربي خط يد و طباعة	80	13559	100
انكليزي–عربي خط يد و طباعة تدوير 90 درجة	5	1053	100
انكليزي–عربي خط يد و طباعة تدوير 180 درجة	10	1752	100

شكل (3) تبين نماذج من الوثائق التي اخلت الى النظام وتم تصنيفها.



شكل (3) نماذج من الوثائق التي استخدمت للتصنيف بواسطة النظام

7. الخوارزمية المقترحة للتصنيف

الخوارزمية في حالة عامة

- المدخلات: صورة ثنائية (binary image) للوثيقة المراد تصنيفها.
- المخرجات : تصنيف مجموعة احرف الموجودة في الوثيقة المدخلة الى نوع اللغة التي تنتمي اليها.

- في المرحلة الأولى: تحديد العناصر المترابطة في الصورة الثنائية (binary image) للوثيقة.
- في المرحلة الثانية : تنفيذ بالتتابع ما يلي لجميع العناصر المترابطة:
  - المرحلة الثانية -1: حساب الميزات (Centroid, Bounding_box, Eccentricity, Euler_number, Extent) لكل كائن مترابط.
  - المرحلة الثانية -2: استخلاص القيم التالية لكل كائن:

holes=1-Eulleer number
h=height of Bounding_box
W=width of Bounding_box
Aspect ratio=h/w
cx=x coordinate of centroid
bx=x coordinate of the center of the Bounding_box
by=y coordinate of the center of the Bounding_box
disx=x distance from the center of box to the centroid
disy=y distance from the center of box to the centroid

$$vdistx = bx + \frac{w}{2} - cx$$

$$vdisty = by + \frac{h}{2} - cy$$

$$tdto = \sqrt{(vdistx^2 + vdisty^2)}$$

$$dtot(normal) = dtot - \min(dtot))/(\max(dtot) - \min(dtot))$$

$$theta$$
$$= angle\ distance\ (dtot) vector\ makes\ with\ horizantal\ direction$$
$$= arctangent(vdisty/vdistx)$$

- في المرحلة الثالثة: احتساب المتوسط الحسابي , معامل التشتت و الألتواء للقيم المحسوبة في المرحلة الثانية.
- في المرحلة الرابعة :ينظم متجه الميزات(feature vector).
- في المرحلة الخامسة :يرسل متجه الميزات الى معادلة المصنف(classifier).

في المرحلة السادسة: بالاعتماد على نتائج المصنف(classifier) نحصل على التصنيف النهائي لاحرف الوثيقة المدخلة للنظام.

8. الاستنتاجات

هذا البحث يعرض طريقة لتصنيف النصوص الى (العربية و الرومانية) دون أي محاولة لاكتشاف الحروف اوتشكيل الكلمات. الفكرة الأساسية هي لاستخراج معلومات شكل الكلمة على مستوى الصفحة ، للحد من حساسية للضوضاء على المستندات الممسوحة ضوئيا . حيث يستند أساسا على معالجة تحليل المكونة الموصولة ، و المكونات المرتبطة تشكيلياً. وقد تم استخدام هذه الميزات الأساسية و استغلالها في مجموعة متنوعة من الطرق. هذه الطريقة سريعة وغير مكلفة حسابيا ، ليست حساسة لاختلاف الخطوط أو إلى الطابع الخط . تصنف هذه خوارزمية المتقدمة النص من ناحية اللغة وكذلك طريقة كتابة النص ان كان يدوية او طباعة.

تم التصنيف النصي بواسطة خوارزمية تصنيف الى نوعين من النصوص ، وهذه النصوص هي الرومانية والعربية. تم الحصول على دقة التصنيف من 100 ٪ .

بعد هذه النتائج، تم التأكد بأن المصنف يمكن استخدامه لتصنيفالوثائق متعددة اللغة. تم استخراج ست ناقلات الميزات من كتلة النص و يتم تغذية هذه إلى مسافة المصنف للتمييز بين النصوص.

المراجع

[1] Dr. Firoj Parwej, ***" The State of the Art Recognize in Arabic Script through Combination of Online and Offline***", International Journal of Computer Science and Telecommunications ,Volume 4, Issue 3, March 2013.

[2] R.M. Haralick, "***Statistical and structural approach to texture***". Proceeding of IEEE, Vol 67, 1997 pp. 786-804.

[3] J. Ding, L. Lam and C.Y. Suen. "***Classification of Oriental and European scripts by using characteristic features***". Fourth International Conference on Document Analysis and Recognition, ICDAR97, Ulm, Germany, 1997, Vol.2, pp. 1023-1027.

[4] B Waked, S. Bergler, C.Y. Suen and S. Khoury. Skew Detection, "***Page Segmentation, and Script Classification of Printed Document Images***". In Proceeding of the 1998 IEEE International Conference on System, Man, and Cybernetics, San Diego, CA, October 1998.

[5] Ali Selamat and Ng Choon Ching, "***Arabic Script Documents Language Identifications Using Fuzzy ART***", Second Asia International Conference on Modelling & Simulation, 2008 IEEE.

[6] S. Ben Moussa, A. Zahour, A. Benabdelhafid, A.M. Alimi, "***Fractal-Based System for Arabic/Latin, Printed/Handwritten Script Identification***", IEEE Pattern Recognition, 2008. ICPR 2008. 19th International Conference on 8-11 Dec. 2008.

[7] Alireza Behrad, Malike Khoddami and Mehdi Salehpour, "***A Novel Framework for Farsi and Latin Script Identification and Farsi Handwritten Digit Recognition***", JOURNAL OF AUTOMATIC CONTROL, UNIVERSITY OF BELGRADE, VOL. 20:17-25, 2010.

[8] Haralick and Shapiro,"***Computer and Robot Vision***", vol I, Addison-Wesley 1992, Appendix A.

[9] C. Jose & R. Neil & W. Curt (2000) "***Numerical and Adaptive Systems: Fundamentals Through Simulations***". John Wiely & Sons, Inc. 2000.