**التعرف على الخط العربي المكتوب يدوياً**

# جامعة الملك فهد للبترول و المعادن

## كلية هندسة و علوم الحاسب الآلي

## قسم علوم الحاسب الآلي

**تعريب الحاسبات**

ع.ح ١\ ١

متطلب 9

**"مواد محاضرتي"**


عمل فريق:

**نحو القمم**

شعار الفريق:


أعضاء الفريق:

أحمد سلام الرفاعي

208602

بإشراف

الاستاذ حسني المحتسب

الثلاثاء، 19 كانون الأول، 2006

المطلوب

**متطلب رقم 9: مواد محاضرتي**

بعد تقديم محاضرتك مباشرة نتوقع منك تسليم ما يلي في هذا المتطلب:

- ملف المحاضرة المقدمة بصيغة بوربوينت
- تقرير المحاضرة بصيغة وورد متضمنا ما يلي:
  - ✓ ملخص المحاضرة
  - ✓ المراجع المستخدمة
  - ✓ المشاكل التي واجهتك
  - ✓ الفوائد التي استفدت منها خلال إعداد وتقديم المحاضرة
  - ✓ 3 أسئلة من نوع "اختيار من متعدد" أو "صح أم خطأ" حول أهم المفاهيم التي قدمت في محاضرتك

**ملخص المحاضرة**

## التعرف على الخط العربي المكتوب يدوياً

التعرف الآلي على النصوص في الصور الممسوحة يمكننا من البحث عن الكلمات في ملفات ضخمة. بالإضافة إلى إمكانية ترتيب الرسائل البريدية، وتنسق النصوص القديمة بطريقة أسهل. إن مجال التعرف على الكلام المكتوب يدوياً فيه الكثير من التحديات التي تُطرق إليها في السنوات القليلة الماضية بشكل أكبر.

### أنواع التعرف

نظم التعرف ممكن أن تكون متزامنة (on-line) أو غير متزامنة (off-line). على سبيل المثال تكون متزامنة عندما يقوم الانسان بكتابة السلسلة الحرفية بالقلم على الـ PDA (المساعد الشخصي الرقمي)، وأيضاً تكون غير متزامنة عندما تعرض على نصوص مكتوبة سابقا مثل صور ممسوحة بالماسح الضوئي. مما لا شك فيه، أن التعرف المتزامن أسهل من غير المتزامن نظراً لأنه عندنا معلومات أكبر. (يعني نحلل حرف حرف بشكل مباشر يكون التركيز أكبر على ما يكتب الآن والشوائب أقل).

التعرف غير المتزامن على الكتابة اليدوية تتضمن تحديد ماهي الحروف أو الكلمات الموجودة في صورة رقمية من الكلام المكتوب. إن لها فائدة عظيمة في التواصل بين الإنسان والآلة و تساعد في معالجة النصوص الكتوبة يدوياً.

### الدافع

العربية يتحدثها 234 مليون شخص وهي ثقافة مهمة لأعداد أكبر من الناس. حيث إن الكلام العربي يختلف ولكن الكتابة العربية موحدة في مختلف أنحاء العالم العربي و تسمى وفقاً للمعيار العربي الحديث Modern Standard Arabic. كما أن العديد من اللغات الأخرى تستخدم الحروف العربية مثل الفارسية والكردية والأردية. لذا فإن القدرة على تفسير الكلام العربي المكتوب آلياً له فوائد واسعة.

يمكننا أيضاً على التعرف على الكتابات العربية القديمة، بنفس طريقة التعرف على اللغة الحديثة يمكننا أيضاً التعرف على اللغة القديمة. المعالجة الآلية تمكننا من زيادة جعل هذه المصادر متوفرة.

### خصائص اللغة العربية

اللغة العربية تتألف من 28 حرفاً. كل حرف له شكلان أو أربعة أشكال، واختيار شكل الحرف يكون على حسب موقعة في المقطع. أربعة مواقع محتملة، بداية المقطع، وسط المقطع، نهاية المقطع أو معزول. الحروف التي لا يمكن أن تكون في بداية المقطع أو وسطه لايمكنها الاتصال مع الحرف الذ ي يليه. الحروف مبينة في الشكل التالي.

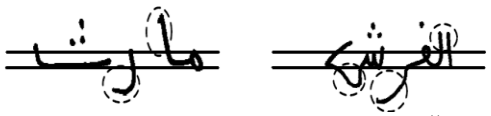
| Name | Isolated | Initial | Medial | Final |
|---|---|---|---|---|
| alif | ا | - | | ا |
| baa | ب | بـ | ـبـ | ـب |
| taa | ت | تـ | ـتـ | ـت |
| thaa | ث | ثـ | ـثـ | ـث |
| jiim | ج | جـ | ـجـ | ـج |
| Haa | ح | حـ | ـحـ | ـح |
| khaa | خ | خـ | ـخـ | ـخ |
| daal | د | - | | ـد |
| dhaal | ذ | - | | ـذ |
| raa | ر | - | | ـر |
| zaay | ز | - | | ـز |
| siin | س | سـ | ـسـ | ـس |
| shiin | ش | شـ | ـشـ | ـش |
| Saad | ص | صـ | ـصـ | ـص |
| Daad | ض | ضـ | ـضـ | ـض |
| Taa | ط | طـ | ـطـ | ـط |
| Dhaa | ظ | ظـ | ـظـ | ـظ |
| ayn | ع | عـ | ـعـ | ـع |
| ghayn | غ | غـ | ـغـ | ـغ |
| faa | ف | فـ | ـفـ | ـف |
| qaaf | ق | قـ | ـقـ | ـق |
| kaaf | ك | كـ | ـكـ | ـك |
| laam | ل | لـ | ـلـ | ـل |
| miim | م | مـ | ـمـ | ـم |
| nuun | ن | نـ | ـنـ | ـن |
| haa | ه | هـ | ـهـ | ـه |
| waaw | و | - | | ـو |
| yaa | ي | يـ | ـيـ | ـي |

**شكل 1: الحروف العربيّة و شكلها على حسب موقعها.**

كما تتميز اللغة العربية بوجود الحركات وهي عبارة عن مدود قصيرة أو تنوين أو تشديد الحرف. عادة هذه الحروف لا تكتب أثناء الكتابة.الحركات هي: الفتحة و الضمة والكسرة والتنوين والشدة والهمزة والمدة.بعض الحركات مبينة بالشكل التالي:

**شكل 2: تنوين الفتح، السكون، الكسرة، الضمة، الفتحة**

كما أن بعض الحروف لها سوابق أو لواحق ( descender) أو (ascenders)، كما هو مبين في الشكل التالي السوابق هي ما فوق أعلى سطر، واللواحق، هي ما تحت السطر الأسفل.



**شكل 3: السوابق واللواحق**

تكتب اللغة العربية من اليمين إلى اليسار، والحروف عادة ماتكون متصلة حتى عند الطباعة. يعتمد توصيل الحرف على الحرف على الحرف الذي يليه. السطر الأساسي (baseline) هو السطر الذي عادة ما تتصل الحروف عن بعضها. في الواقع قريب منه.

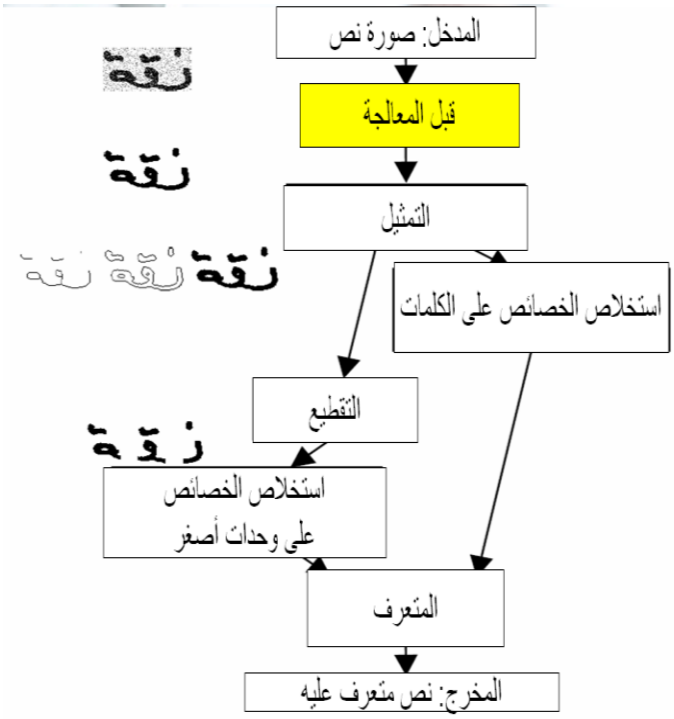
لا يوجد اتصال بين الكلمات، لذا يجب أن نضع فراغ. هناك ست حروف لا تتصل إلا من اتجاه واحد. لما يظهروا في كلمة ما تنقسم الكلمة إلى عدد من المقاطع.

الرباط Ligature: هو تكوين الحرف باتصال حرف أو أكثر بشكل مقبول مثل اللام ألف (لا) أو يا ميم، لام ميم، وغيرها. الاشكال المختلفة لحرف لا مبينة بالشكل التالي.

حرف لا وطرق كتابته يدوياً.

**التعرف على الكتابة اليدوية – طريقة العمل**

للتعرف على الكتابة العربية يجب أن نمر بعدد من المراحل، أولاً مرحلة ما قبل المعالجة، ثم مرحلة التمثيل. إن هنالك طريقتان للتعرف الطريقة الكلية، بحيث نتعرف على الكلمة بشكل كامل، ولا يوجد حاجة إلى التقطيع، والطريقة الثانية هي طريقة هي التعرف على أجزاء من الكلمة كالأحرف وغيرها، وفي هذه الحالة نحتاج لمرحلة التقطيع. بعد ذلك نحتاج إلى مرحلة استخلاص الخصائص للكلمات أو الوحدات الأصغر، ثم مرحلة المتعرف، وسوف يتم شرح ما يجري بكل مرحلة فيما يلي. الشكل التالي يوضح الشكل العام للتعرف على الكلام المكتوب يدوياً.




**شكل 4: المخطط العام للتعرف على الكتابة اليدوية**

**قبل المعالجة**

تتم قبل المعالجة العمليات التالية: التعرف على الخط الأساسي، عملية إزالة الشوائب، وتصحيح الميلان عن طريق معالجة الصور، وتحديد مكان النص في الصورة، وفصل النص عن ما يحيطه

**التمثيل**

الصورة عادة ما تتحول إلى شكل موجز قبل التعرف، وهناك شكلان عادة ما يمثل الشكل بهما: الهيكل skeleton هو عبارة عن التعبير عن الكلمة بسمك نقرة ضوئية (بكسل) واحدة تظهر الخط الوسطي للنص. عملية الهيكلة skeletonization أو التنحيف thinning تسهل عملية تصنيف الصورة و أخذ خصائصها. الطريقة الثانية هي المحيط contour. مشاكل التنحيف هي التحديد بشكل خاطئ للخصائص، والالتباس الخاص بكل خوارزمية تنحيف أما طريقة المحيط تتجنب هذه المشاكل لأنها لا تخسر معلومات. الشكل التالي يبين كيف ستظهر الكلمة في كلا الشكلان.

شكل 5: طرق التمثيل: المحيط (يمين)، والهيكل (وسط)، والكلمة الاصلية (يسار).

**التقطيع**

التقطيع (segmentation) هي المهمة بتقسيم الكلمة إلى الحروف المكونة لها. الاتصال شيء أساسي باللغة العربية يجعل المهمة أكثر تعقيدا في أثناء التعرف على الحروف اللاتينية. والكتابة اليدوية عندها أصلاً اختلافات في الميلان slop و الامتداد stretch و الانحراف skew والحجم size وكيفية ظهور الحرف. كما أنه من الممكن الحرف يظهر فوق أو تحت الحرف السابق. و أيضاً في بعض الأحيان قد يظهر الحرف التالي قبل الحرف السابق. الشكل التالي يوضح بعض هذه الصعوبات.

شكل 6: ظهور الحرف التالي قبل السابق (يمين) أو التالي تحت السابق (يسار).

لهذه الاسباب قد يعتقد الكثيرون أن اللغة العربية أصعب للتعرف عليها من اللغة الانكليزية. ولكن هناك اعتبارات تجعل اللغة العربية أسهل مثل: عدم وجود حروف كبيرة وصغيرة، وخط أساسي قوي، وقصر طول الكلمة بالمعدل، نقاطه المميزة، تغير شكل الحرف على حسب موضعه بشكل نظامي.

**استخلاص الخصائص**

الخصائص هي عبارة عن قياسات عددية مأخوذة عن الصور أو عن مكان في الصور، هذه القياسات هي التي تمرر للمتعرف لكي يتعرف عليها

أمثلة على الخصائص:

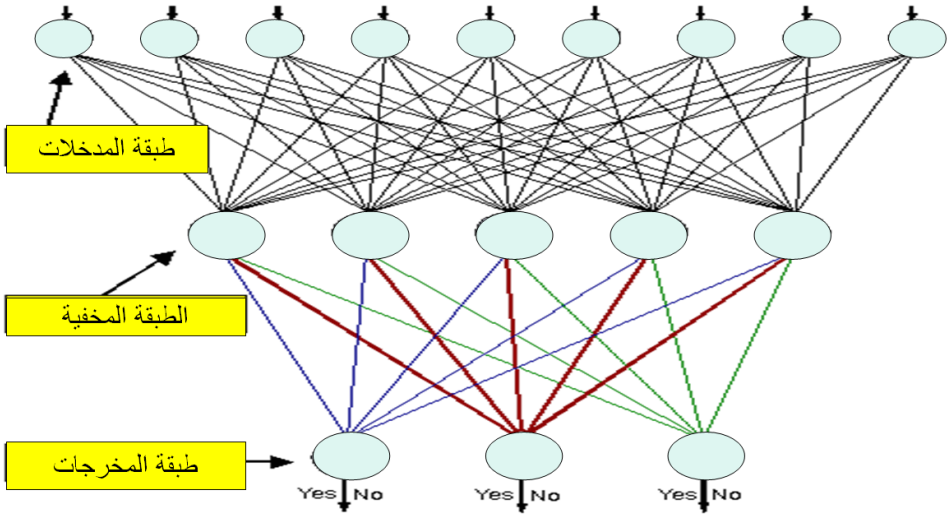
1. كثافة النقرة الضوئية أو البكسل.
2. تقعر التكوينات بالنسبة للخط الأساسي.
3. نسبة الطول إلى العرض.
4. السوابق واللواحق.
5. الخصائص البنيوية.
6. الدوران، ونقاط البداية والنهاية.
7. الطول والعرض.
8. النقاط على الحروف

**طرق التعرف**

التعرف من الممكن أن يكون مبني على القواعد أو مبني على الاحتمالات أو كليهما معاً. ونستطيع أن نبني هذا المعالج باستخدام بعض الطرق مثل الشبكات العصبونية، انموذج ماركوف المخفي والقواعد أو عن طريق هجين بين الطرق الاحصائية والقواعد.

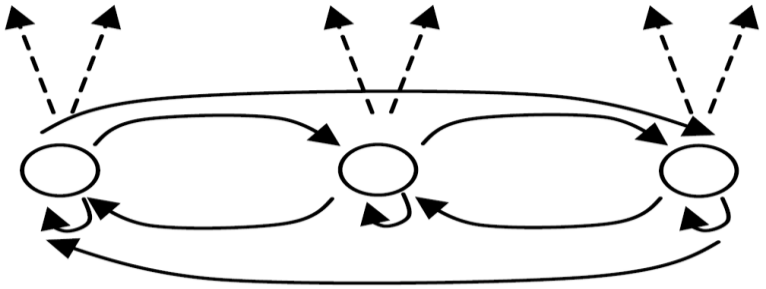
القواعد: إنشاء قواعد تعتمد على الخصائص البنيوية للحرف مثل المنحنيات المفتوحة بعدد من الاتجاهات.

الشبكات العصبونية: تتكون من عنصر معالجي بسيط و عدد كبير جداً من الترابط، الأوزان في العناصر تدرب من خلال بيانات تدريبية. وهي مقسمة إلى طبقة مدخلات، و طبقات متوسطة "مخفية" و طبقة مخرجات نهائية.
المعلومات تذهب من خلال البداية إلى النهاية التي تعطي الحرف المطلوب. الشكل التالي يري الشكل العام للشبكات العصبونية.




شكل 7: الشبكات العصبونية

انموذج ماركوف المخفي(HMM) تعتبر مناسبة لتعلم الخصائص التي من الصعب وصفها بشكل بديهي.
متسلسلة باتجاه واحد، فيها states حالات و احتماليات propablitites للانتقال بين هذه الاحتماليات على حسب النظر إلى متسلسلة من البيانات أو الملاحظات observations . إن كان عندنا س حالة، و ص ملاحظة محتملة، يعتمد الاختيار على الاحتمالية المترافقة مع كل حالة. الهدف هو لإعادة بناء مسار الحالات او "path" من الملاحظات، لكي يتعلم معاني البيانات. في التعرف على النصوص، الملاحظات من الممكن أن يكون عبارة عن مجموعة من البكسلات، و الحالات هي عبارة أجزاء من الحروف. لإيجاد المسار هي أخذ أعلى احتمالية من عدد من النماذج. الشكل التالي يري الشكل العام لنموذج ماركوف المخفي.



شكل 8: نموذج ماركوف المخفي.

**الاستخدامات**

أثبت التعرف على الخط العربي المكتوب يدوياً جدواه، في التطبيقات المحصورة مثل التعرف على الارقام المكتوبة على الشيكات يدوياً، لقاعدة بيانات AHDB التي تحتوي على أرقام و أكثر الكلمات استخداماً، وكلمات مكتوبة يدوياً، والتعرف على العناوين البريدية، لقاعدة بيانات تحوي أسماء مدن وأرقام وكلمات تظهر في العناوين.

## المراجع المستخدمة

Offline Arabic Handwriting Recognition: A Survey. By Liana M. Lorigo, Venu Gvindaraju.
بالإضافة مراجع أخرى من الانترنت و محاضرة المادة عن التعرف الضوئي عن الكتابة العربية تمت قراءتها والاستفادة منها بشكل غير مباشر.

## المشاكل التي واجهتني

الحصول على الترجمة الدقيقة.
فهم المادة العلمية
محاولة الحصول على عرض ليس عام جداً يصلح لأن يعرض لأناس غير اختصاصين، وليس متخصص جداً بحيث لا يمكن أن يفهمه إلا من له باع كبير في مجال التعرف.
استخلاص المعلومات المهمة من الأوراق البحثية.
الحصول على مراجع سهلة الفهم، من الممكن عرضها على الطلاب يستدعي وقتاً جيداً.

## الفوائد التي استفدتها من المحاضرة

فتحت لي الآفاق في مجالات بحثية جديدة.
تقدير المحاضرة الناجحة، لما ورائها من العمل الجاد.
التواصل مع الآخرين باستخدام اللغة العربية، كيف يتم ذلك؟ لأن المراجع باللغة الانكليزية، وخلفيات المستمعين قد تكون بالانكليزية، فتحتاج إلى إظهار المصطلح العربي الدقيق إلى جانب المصطلح الانكليزي.
الثقة بالنفس والقدرة على الانجاز.

## ثلاث أسئلة

مشاكل التمثيل بالمحيط هي التحديد بشكل خاطئ للخصائص، والالتباس الخاص بكل خوارزمية تنحيف أما طريقة التمثيل بالتنحيف (الهيكلة) تتجنب هذه المشاكل لأنها لا تخسر معلومات. الإجابة: ☒

طريقة الشبكات العصبونية في التعرف على الخط العربي المكتوب يدوياً بحاجة إلى تدريب من خلال بيانات تدريبية حتى تعطي نتيجة صحيحة.الإجابة: ☑

من الأمثلة الناجحة على استخدامات التعرف على الخط العربي المكتوب يدوياً هي التعرف على الأرقام و الكتابات المكتوبة على الشيكات.الإجابة: ☑