

# التعرف على الخط العربي المكتوب يدوياً

أحمد سلام الرفاعي

## مسار المحاضرة

- مقدمة
- الدافع
- خصائص اللغة العربية
- التعرف على الخط العربي المكتوب يدوياً.
- التمثيل
- التقطيع
- الخصائص
- التعرف
- الاستخدامات
- خاتمة
- المراجع
- أسئلة

- نظم التعرف ممكن أن تكون متزامنة (on-line) أو غير متزامنة (off-line).
- المتزامن أسهل من غير المتزامن نظراً لأنه عندنا معلومات أكبر.
- التعرف غير المتزامن (off-line) على الكتابة اليدوية هو تحديد ماهي الحروف أو الكلمات الموجودة في صورة رقمية من الكلام المكتوب.
- لها فائدة عظيمة في التواصل بين الإنسان والآلة و تساعد في معالجة النصوص المكتوبة يدوياً.

## الدافع

- العربية يتحدثها ٢٣٤ مليون شخص وهي ثقافة مهمة لأعداد أكبر من الناس.
- الكلام العربي يختلف ولكن الكتابة العربية موحدة في مختلف أنحاء العالم العربي و تكون 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.

## خصائص اللغة العربية

Name	Isolated	Initial	Medial	Final
alif	ا	-		ا
baa	ب	ب	ب	ب
taa	ت	ت	ت	ت
thaa	ث	ث	ث	ث
jiim	ج	ج	ج	ج
Haa	ح	ح	ح	ح
khaa	خ	خ	خ	خ
daal	د	-		د
dhaal	ذ	-		ذ
raa	ر	-		ر
zaay	ز	-		ز
siin	س	س	س	س
shiin	ش	ش	ش	ش
Saad	ص	ص	ص	ص
Daad	ض	ض	ض	ض
Taa	ط	ط	ط	ط
Dhaa	ظ	ظ	ظ	ظ
ayn	ع	ع	ع	ع
ghayn	غ	غ	غ	غ
faa	ف	ف	ف	ف
qaaf	ق	ق	ق	ق
kaaf	ك	ك	ك	ك
laam	ل	ل	ل	ل
miim	م	م	م	م
nuun	ن	ن	ن	ن
haa	ه	ه	ه	ه
waaw	و	-		و
yaa	ي	ي	ي	ي

- اللغة العربية تتألف من ٢٨ حرف.
- 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 الذي يليه.

## خصائص اللغة العربية

أ ب ج د هـ و ز ح ط ي

العرش = مارث

دقار ميسر المزونة أ كودة

- الحركات (الفتحة، الضمة، الكسرة، السكون، التثوين، الشدة، المدة، الهمزة)
- كما أن بعض الحروف لها سوابق (ascenders) أو لواحق (descenders).
- تكتب اللغة العربية من اليمين إلى اليسار، والحروف عادة ما تكون متصلة حتى عند الطباعة.
- السطر الأساسي (baseline) هو السطر الذي عادة ما تتصل الحروف مع بعضها.

## خصائص اللغة العربية

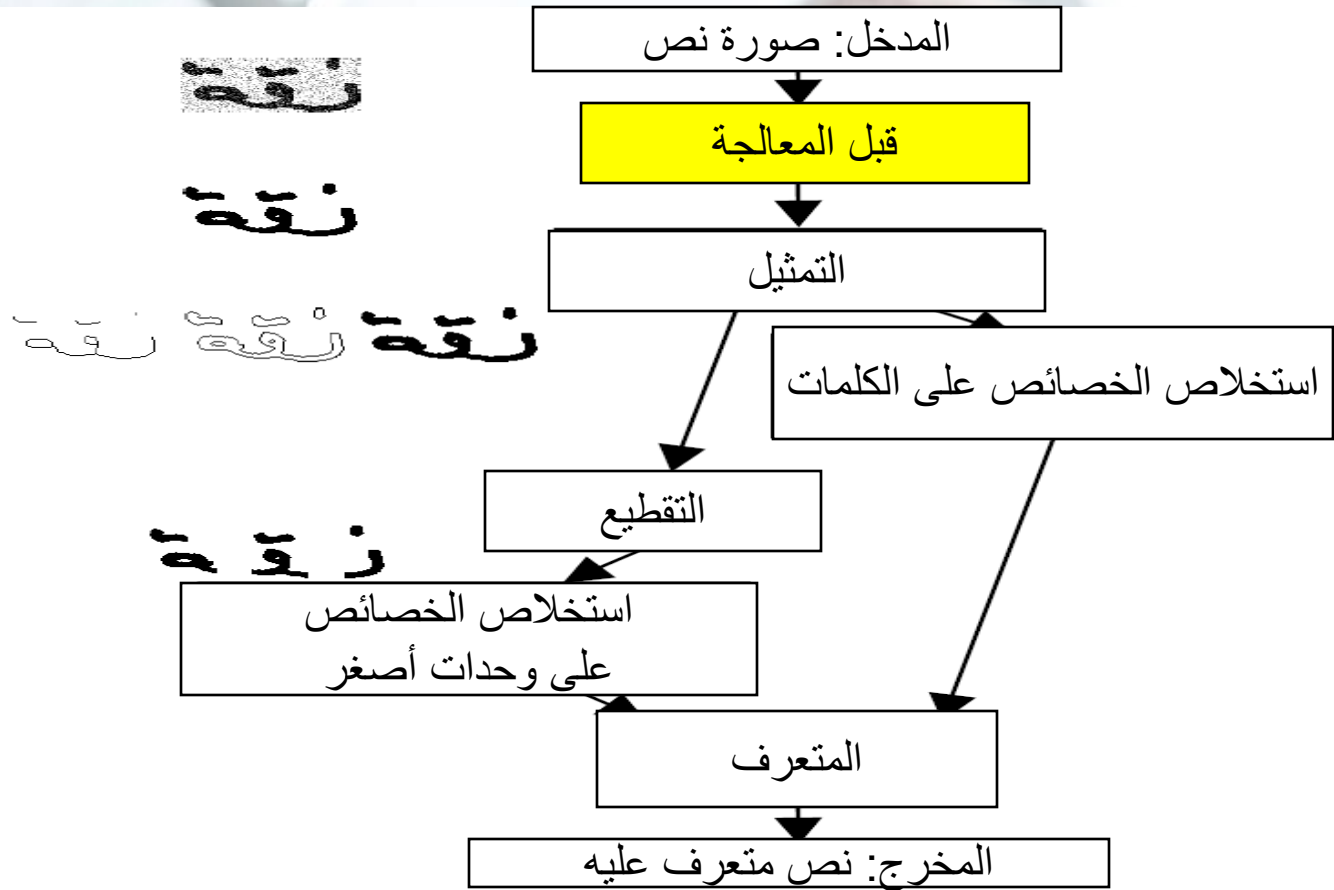
رَبِينَا رَبِينَا رَبِينَا رَبِينَا

رؤاد دعومس شاف هتلين

لا لا لا لا

- يعتمد اتصال الحرف في الكلمة على الحرف نفسه والحرف الذي يليه.
- 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).

# التعرف على الكتابة اليدوية - طريقة العمل

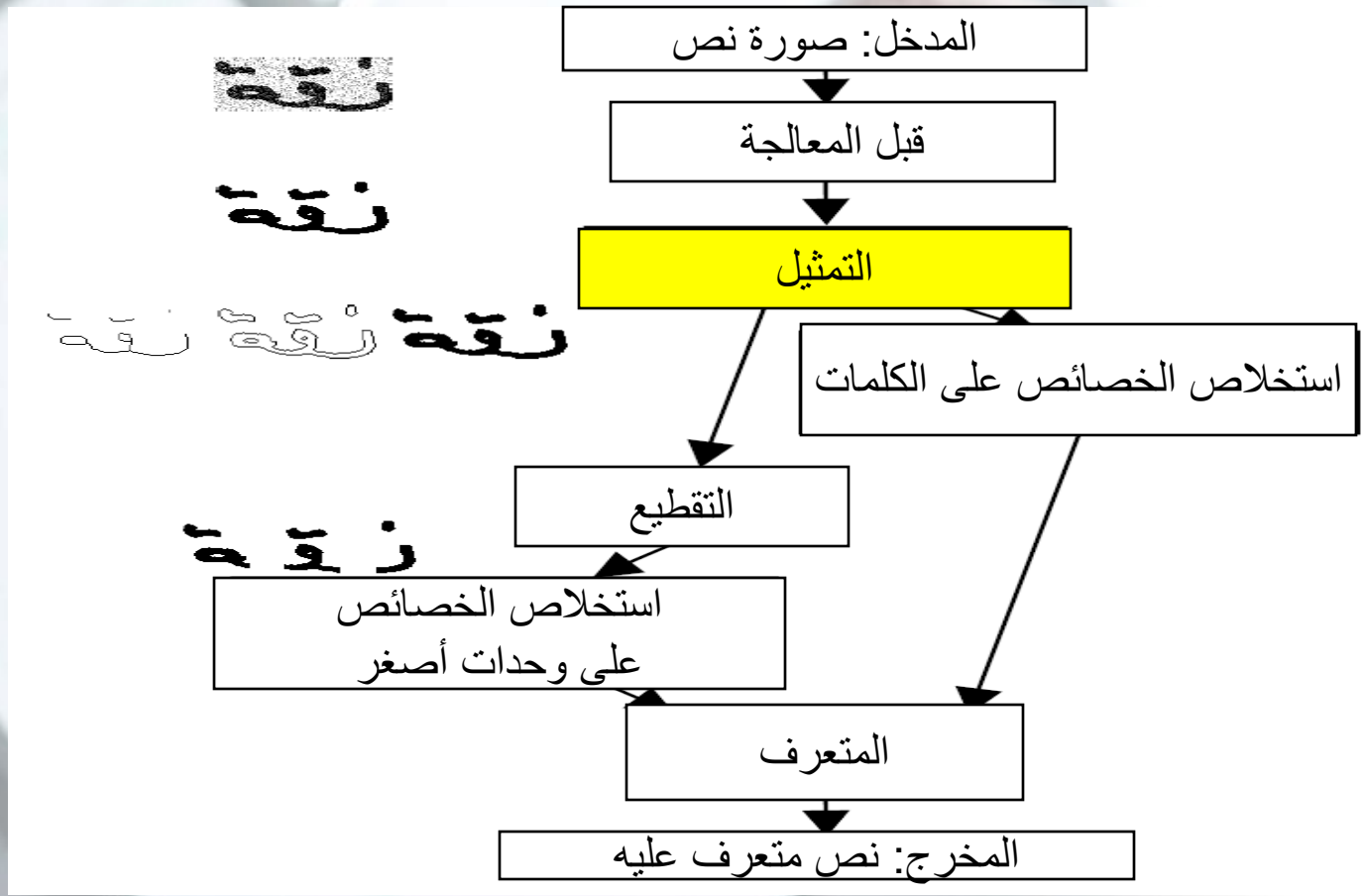




## مهمات ما قبل المعالجة

- التعرف على الخط الأساسي.
- عملية إزالة الشوائب، وتصحيح الميلان عن طريق معالجة الصور.
- مكان النص في الصورة، و فصل النص عن ما يحيطه

# التعرف على الكتابة اليدوية - طريقة العمل



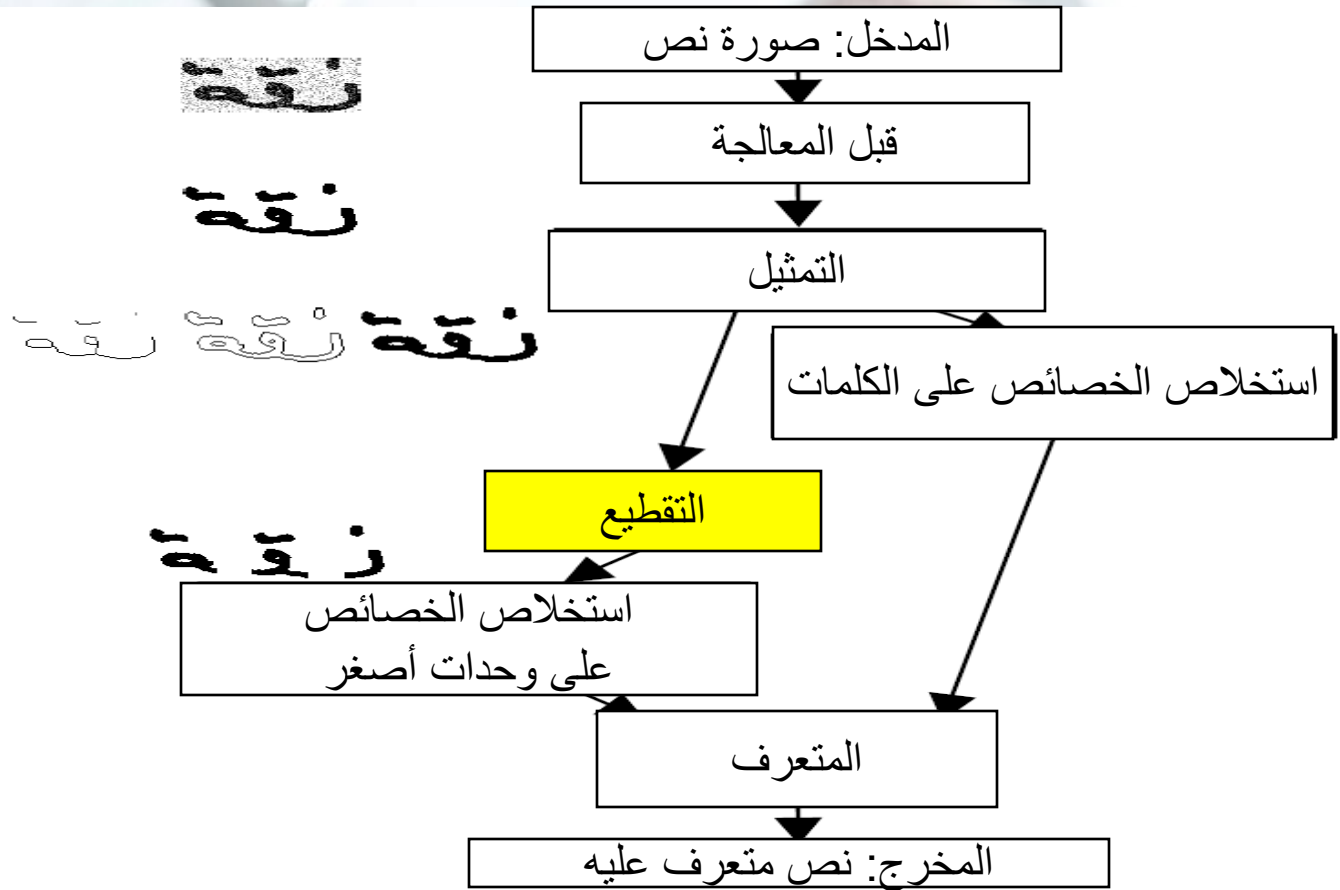
نقعة  
نقعة  
نقعة نقعة نقعة  
نقعة

## التمثيل

- الصورة عادة ما تتحول إلى شكل موجز قبل التعرف، الهيكلة skeleton هو عبارة عن التعبير عن الكلمة بسمك نقطة ضوئية (بكسل) واحدة تظهر الخط الوسطي للنص. الهيكلة skeletonization أو التحنيف thinning تسهل عملية تصنيف الصورة و أخذ خصائصها.
- الطريقة الثانية هي المحيط contour.
- مشاكل التحنيف هي التحديد بشكل خاطئ للخصائص، والالتباس الخاص بكل خوارزمية تنحيف أما طريقة المحيط تتجنب هذه المشاكل لأنها لا تخسر معلومات.

ط ب ل ب ن ه ط ب ل ب ن ه ط ب ل ب ن ه

# التعرف على الكتابة اليدوية - طريقة العمل



نقعة  
نقعة  
نقعة نقعة نقعة  
نقعة

# التقطيع

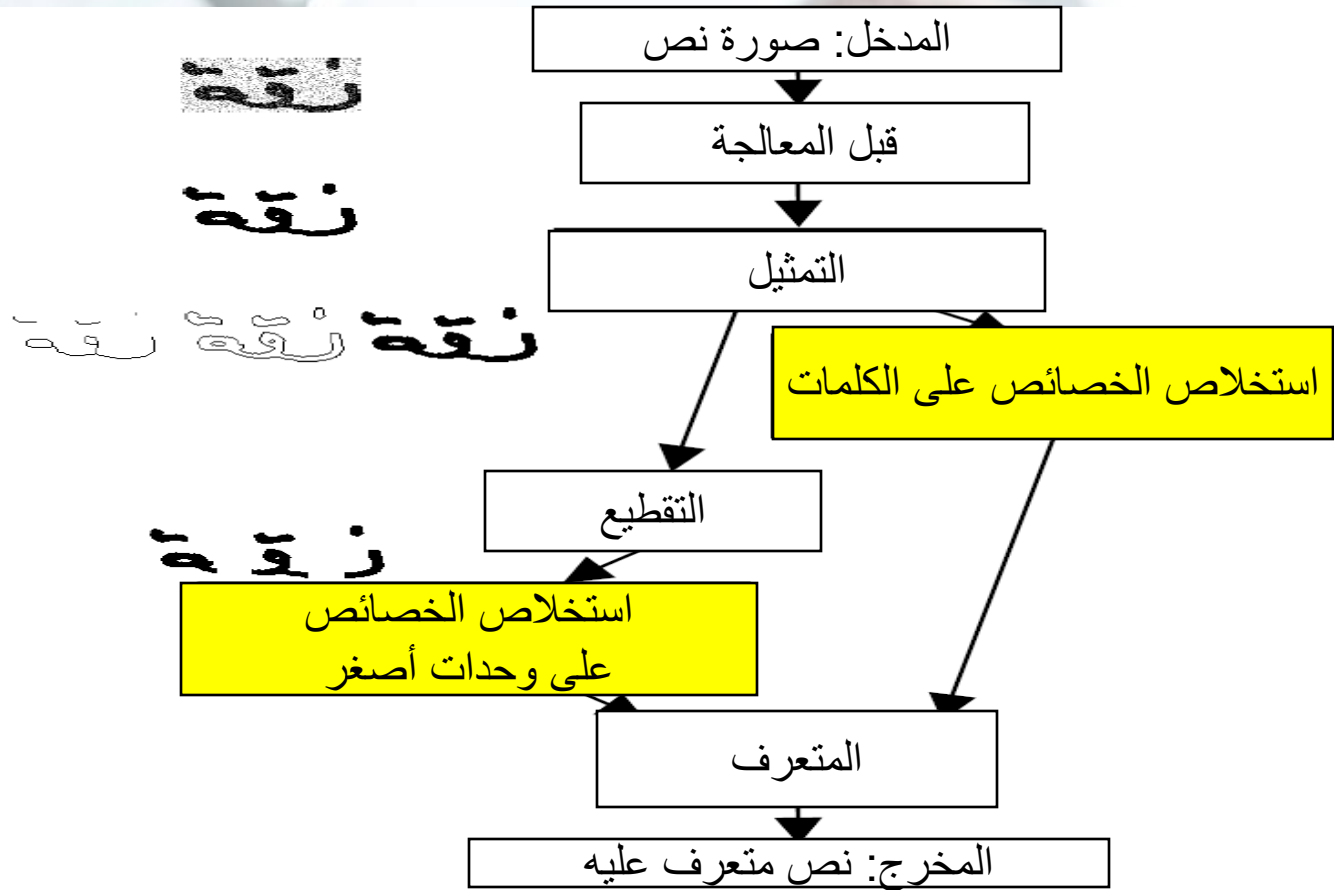
- التقطيع (segmentation) هي عملية تقسيم الكلمة إلى الحروف المكونة لها.
- الاتصال شيء أساسي باللغة العربية يجعل المهمة أكثر تعقيدا في أثناء التعرف.
- والكتابة اليدوية عندها أصلاً اختلافات في الميلان slop و الامتداد stretch و الانحراف skew والحجم size وكيفية ظهور الحرف. كما أنه من الممكن أن يظهر الحرف فوق أو تحت الحرف السابق. و أيضاً في بعض الأحيان قد يظهر الحرف التالي قبل الحرف السابق.
- 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، تغيير شكل الحرف على حسب موقعه بشكل نظامي.
- تقطيع الكلمات إلى حروف أو أجزاء من حروف أو أي وحدة أخرى.

الجليج

خط الجليج

Printed, Handwritten					Remarks
ب	ا	ب	د	د	Vert. line may be missing
ث	ث	ث	ث	ث	Dot pattern varies
ش	ش	ش	ش	ش	Dots, curve shape vary
ي	ي	ي	ي	ي	Curves' angles, sizes vary

# التعرف على الكتابة اليدوية - طريقة العمل



نقعة

نقعة

نقعة نقعة نقعة

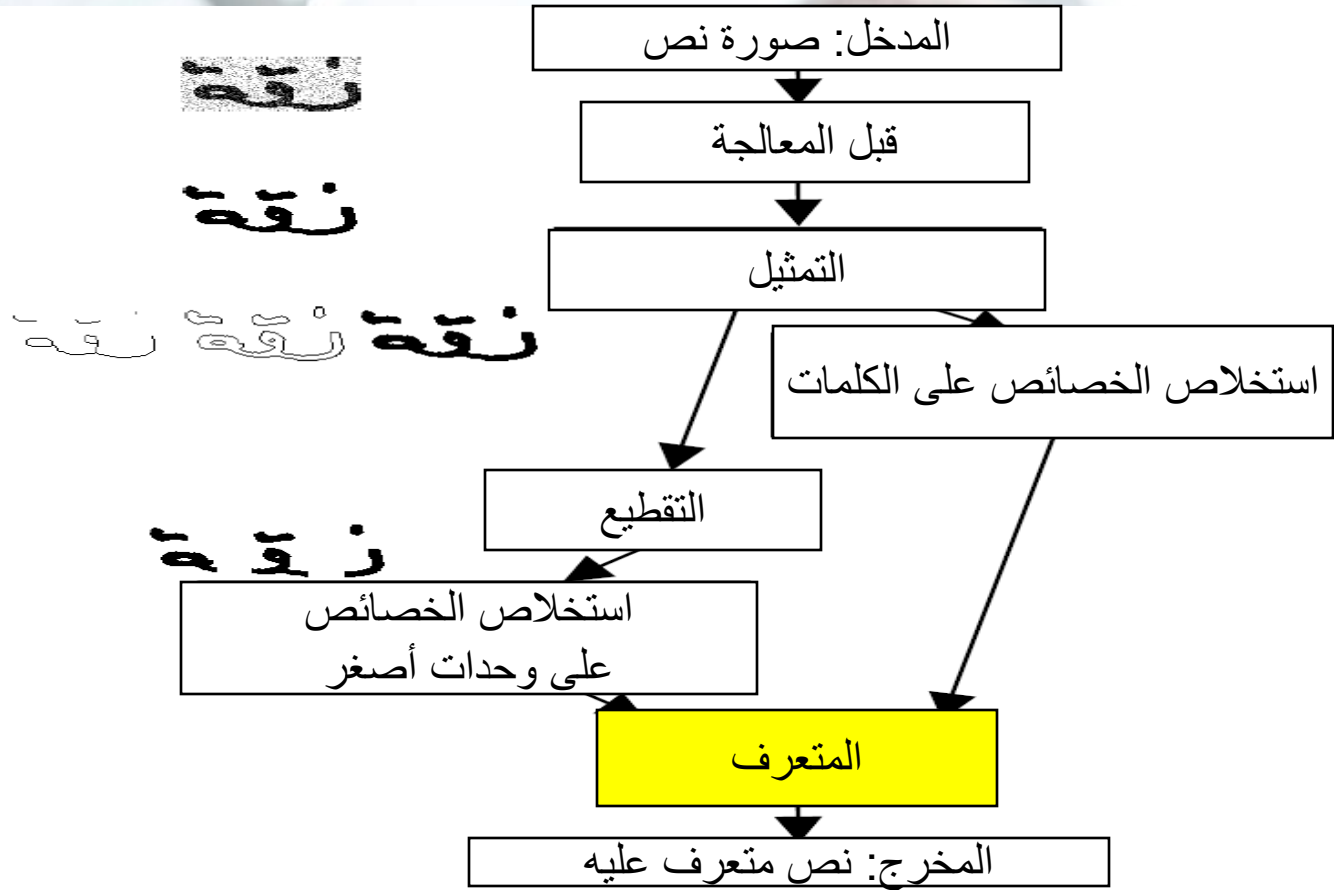
نقعة

## الخصائص

- هي عبارة عن قياسات عددية مأخوذة عن الصور أو عن مكان في الصور.
- الخصائص هي المعلومات التي تمرر للمتعارف.
- من الممكن استخدامها للتقطيع.
- أمثلة على الخصائص:
  - كثافة النقرة الضوئية أو البكسل.
  - تقعر التكوينات بالنسبة للخط الأساسي.
  - نسبة الطول إلى العرض.
  - السوابق واللواحق.
  - الخصائص البنوية.
  - الدوران، ونقاط البداية والنهاية.
  - الطول والعرض.
  - النقاط على الحروف. وغيرها.



# التعرف على الكتابة اليدوية - طريقة العمل



نقطة

نقطة

نقطة نقطة نقطة

نقطة



## معالج التعرف

- التعرف من الممكن أن يكون
  - مبني على القواعد.
  - مبني على الاحتمالات.
  - أو كليهما معاً.
- باستخدام بعض الطرق مثل
  - الشبكات العصبونية.
  - انموذج ماركوف المخفي.
  - القواعد.
  - هجين بين الطرق الاحصائية والقواعد.

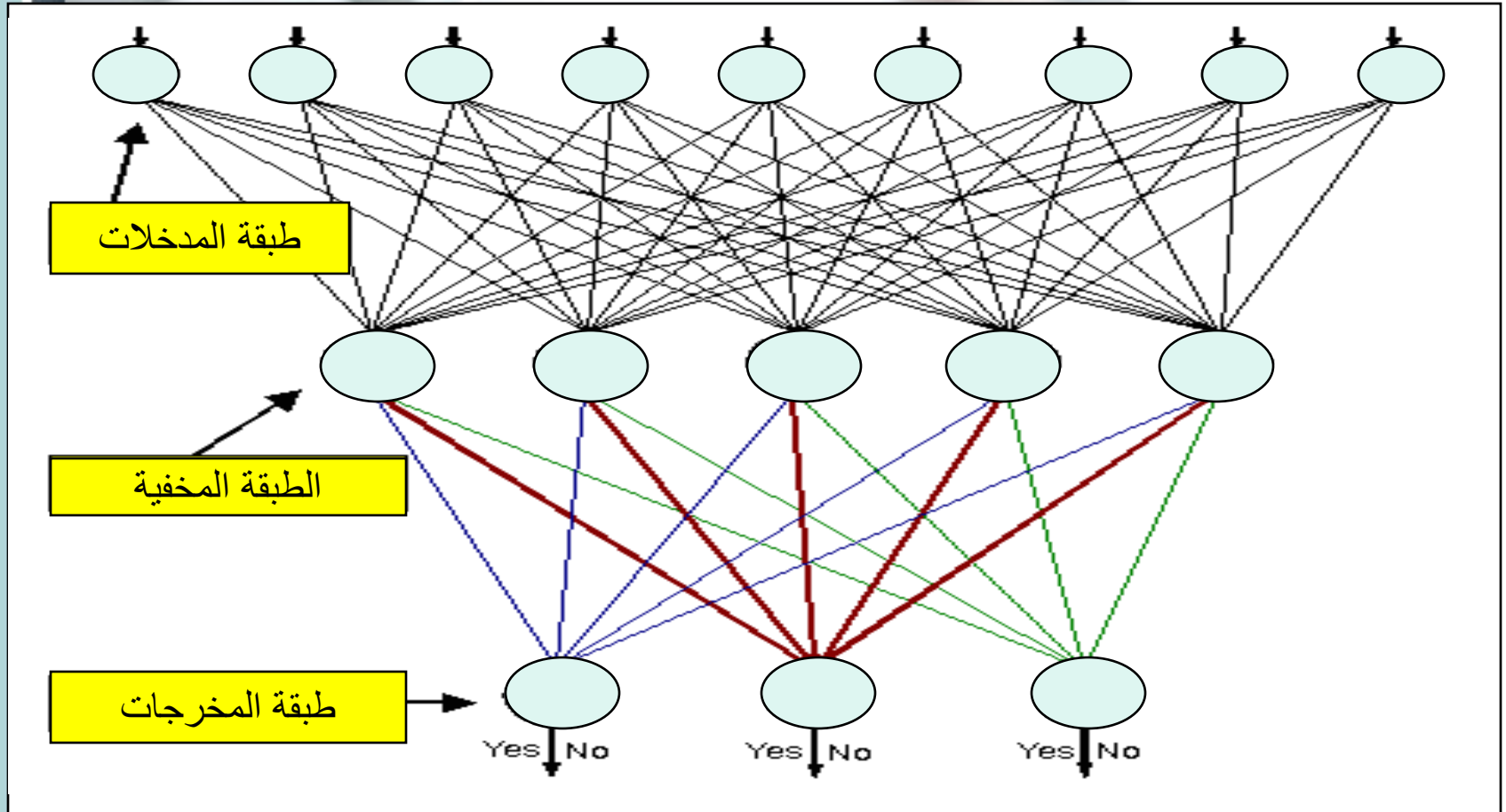
## القواعد

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

## الشبكات العصبونية

- Neural Networks
- تتكون من عنصر معالجي بسيط و عدد كبير جداً من الترابط،  
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

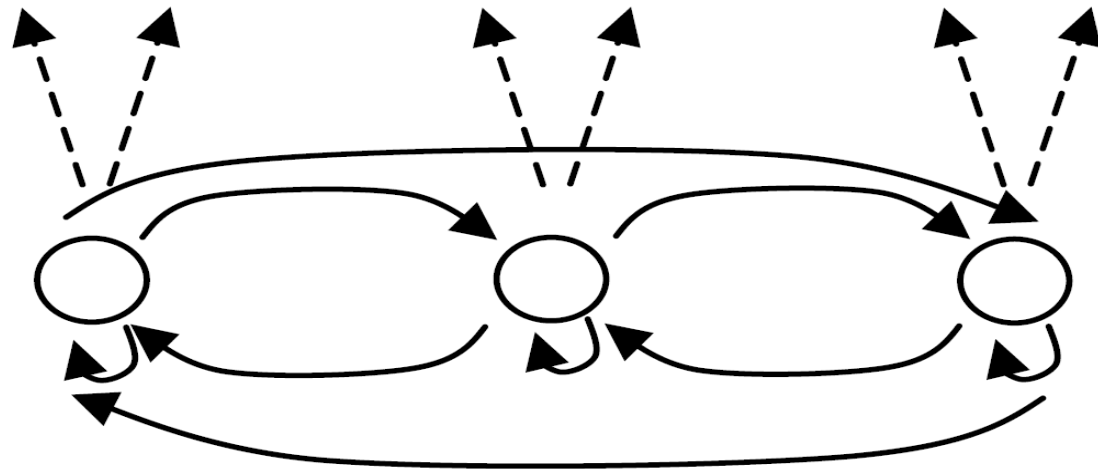
# الشبكات العصبونية



## انموذج ماركوف المخفي

- Hidden Markov Model (HMM)
- تعتبر مناسبة لتعلم الخصائص التي من الصعب وصفها بشكل بديهي.
- وهي عبارة عن متسلسلة باتجاه واحد فيها حالات واحتماليات للانتقال بين هذه الحالات على حسب الملاحظات.
- 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.

# انموذج ماركوف المخفي



- التعرف على الأرقام المكتوبة على الشيكات يدوياً.  
– مثل قاعدة بيانات AHDB التي تحتوي على أرقام و أكثر الكلمات استخداماً، وكلمات مكتوبة يدوياً.
- التعرف على العناوين البريدية، مثل قاعدة بيانات تحوي أسماء مدن وأرقام وكلمات تظهر في العناوين.
- التطبيقات المحصورة في التعرف الآلي على الكلام المكتوب يدوياً، أثبتت جدواها.

## خاتمة

- أهمية التعرف على الخط المكتوب يدوياً.
- الطريقة العامة للتعرف.
- بعض خوارزميات التعرف.
- بعض الاستخدامات.



## المراجع

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

وأخيراً

- شكراً لكم على حسن استماعكم
- أيّ

